EXPLORING EXPLAINABLE OPTIMIZATION IN MEDICAL SEGMENTATION NETWORK FOR MULTI-SCALE GENERALIZATION WITH ANATOMICAL ATLAS

By

Ho Hin Lee

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

August 11st, 2023

Nashville, Tennessee

Approved:

Bennett A. Landman, Ph.D.

Yuankai Huo, Ph.D.

Ipek Oguz, Ph.D.

John Jeffrey Carr, M.D., M.Sc.

Zhoubing Xu, Ph.D.

The dedication page is optional. The Copyright page is also optional. If you do not use them, please delete them. If you have a dedication, then center it in the middle of this page. If no Copyright page, begin printing page numbers here, using lower case Roman numerals and continue consecutive Roman numeral numbering throughout the preliminary pages.

ACKNOWLEDGMENTS

It's truly astonishing to realize that six years have already passed since I first arrived in the US to pursue my master's degree in 2017. Without a doubt, residing in an unfamiliar country for such a prolonged period of time has been a transformative experience. I can vividly recall that my primary objective upon arrival was to seek out the opportunities that aligned with my aspirations. Looking back at the journey of the past six years, I have encountered a multitude of emotions, ranging from immense joy to nerve-wracking moments, from frustrating setbacks to heartbreak. These diverse experiences have played an instrumental role in shaping me into the person I am today, influencing not only my academic pursuits but also my lifelong voyage.

When I reflect on the steps I have taken, I am overwhelmed with gratitude towards the multitude of individuals who have accompanied me during my Ph.D. endeavors. At the forefront, I want to express my deepest appreciation to my father, Siu Hung (Patrick) Lee, and my mother, Wun (Ivy) Hui. Despite the fluctuating economic circumstances our family has faced over the past few years, their unwavering support has remained constant. They have consistently stood by my side, offering their encouragement and guidance, propelling me forward to conquer every frustration and obstacle I encountered. Their unwavering belief in me has been the driving force behind my journey. I am profoundly proud to be their son, and now, I recognize that it is my responsibility to leverage my abilities to support them in return.

Before embarking on my Ph.D. journey, I am immensely grateful to have had the opportunity to work alongside and learn from exceptional mentors during my internships at Siemens Healthineers. I extend my heartfelt appreciation to Rui Liao, Yue Zhang, Xiao Chen, Zhoubing Xu, and Siqi Liu for their invaluable advice and guidance in the field of medical AI. Their mentorship ignited a deeper passion within me for this domain. It was through their recognition and recommendation that I had the privilege to meet my Ph.D. supervisor, Professor Bennett Landman. Professor Landman, I am profoundly grateful for your patience and unwavering guidance throughout my Ph.D. journey. Your mentorship has enabled me to confidently present my research ideas to diverse audiences. I am grateful for the numerous opportunities you have entrusted me with, and for treating me as both a friend and a student. Your feedback and insights have continuously shaped me into a better researcher. I still remember your request for an immediate call when I received the return offer from Microsoft. Your warm approach fills me with gratitude for having you as my mentor. I am thankful to have had your support during every step of my Ph.D. Another individual I deeply appreciate is Professor Yuankai Huo. If Professor Landman can be described as the guiding star that illuminates my path, Professor Huo can be likened to the one who paves the way for me to follow that direction. Professor Huo possesses an exceptional talent for rescuing me from research dead ends. His invaluable feedback has consistently reminded me of the importance of simplicity in writing. I am grateful for his patience in forgiving my occasional careless mistakes in writing over the past few years, while continually pushing me to reach new heights.

In addition to my previous supervisors and my family, there are a few more people I would like to express my gratitude to. Firstly, I want to extend a heartfelt thank you to my partner, Yu Zhao. Throughout my Ph.D. journey, her unwavering support, happiness, and kindness have been a source of strength for me. Her constant presence have helped me navigate through moments of frustration and kept me motivated. I am truly grateful to have you by my side as a partner, and I never felt alone while walking this path. Furthermore, I would like to express my appreciation to all my colleagues from the MASI lab. Each member of the lab is incredibly talented and brilliant. The collaborative discussions and brainstorming sessions we had as a team have been instrumental in shaping and generating numerous research ideas. The generosity and camaraderie displayed by everyone in the lab have made me feel a strong sense of belonging to a larger family. I truly enjoy working with all of you as a team, and I am grateful for the opportunity to collaborate and learn from such exceptional individuals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Overview

Medical imaging is an essential clinical tool in diagnostic investigation and provides efficient visualization of internal organs, bones, soft tissue and blood (50). With different imaging protocols and vendors, we are allowed to visualize different contextual information from organs with different appearance in tissues/organs, thus becoming more easily to classify patients' conditions (18; 213). However, only qualitative visualization can be provided from the images and perform analysis with our vision. It is limited to investigate the fine-grained details in each specific organ region and adapt quantitative measurements without addition guidance. To extract meainingful information from the imaging, multiple ways have been proposed with respect to the corresponding anatomical regions, such as bounding box to search for potential lesion (167), pixel/voxel-wise localization of organ regions (200; 125) and streamline analysis for white matter flows in brain region (20). To adapt the fine-grain measurement of different organs, image segmentation is an essential medical task to localize and capture the volumetric context of each organ by classifying each pixel/voxel into a corresponding semantic value (e.g., 1,2,3) (125; 200; 257; 182; 121). However, the most naive way to obtain such semantic segmentation mask is to annotate manually with our bared hands. Although the manual annotations are regarded as golden standard, such process requires significant amounts of time efforts and efforts for clinicians. To reduce the manual efforts, different automated segmentation approaches including atlas-based methods (216; 19) and machine learning approaches (125; 222), have been proposed to demonstrate the opportunity of automatic segmentation in imaging domain.

Deep learning models have been widely explored for both computer vision and medical image analysis , especially for medical image segmentation. Medical image segmentation can be divided into two perspectives: 1) binary organ segmentation and 2) multi-organ segmentation. Binary organ segmentation is to assign one-hot semantic value meaning to the image pixels/voxels, while multiple values are assigned to different organs in the image. Meanwhile, it is vitally important to localize multiple organ of interests in pixel/voxel-wise setting and further investigate the qualitative/quantitative measures within the organ regions. In general, 2D segmentation approaches take separated slices for learning slice-wise representation (32; 31; 257; 29), while 3D segmentation approaches have also taken places to extract representations from a complete volumetric scan (125; 141). Approach in different dimensionality has its corresponding advantages, such as faster and low-cost memory in training for 2D approaches, extract spatial information between slices for 3D

approaches. Furthermore, with the introduction of conceptual theory in vision transformers (ViTs) (52), multiple transformer networks achieve state-of-the-art performance on several 3D medical imaging benchmarks for semantic segmentation (253; 74; 73). The key contribution of vision transformer is to extract features with large receptive field with limited inductive bias, while the computation of attention mechanism is quadratic complexity with respect to the input resolution, especially for volumetric medical images. Therefore, hierarchical transformers such as swin transformer, further address the limitation in vanilla transformer design and extend the encoding design with the prior knowledge from convolution neural network (CNN) to tackle the high-resolution prediction task. However, the sliding window strategies for extracting attentional features are still highly dependent on the defined resolution of the input patches. Meanwhile, the conceptual and engineering idea of sliding window behaves similarly with the kernels in convolution. Therefore, an alternative of revisiting convolution modules can be explored to simulate the capability of hierarchical transformers and enhances the segmentation performance with efficiency.

Apart from the generic design of the neural network backbone, another important aspect in deep learning model is to interpret and evaluate the effectiveness of the meaning of the learned feature in a trained model. Current deep learning models are often denoted as "black boxes", which demonstrates the insufficient context bringing out from the learned model in terms of mathematical derivation and the learned features. Although a lot of studies have been proposed to enhance the transparency of the model, it is still challenging to enhance the interpretability of the learned features, especially for multi-object semantic segmentation. It is easy to explain the features learned for single organ segmentation, while the single learned features cannot separably explain each organ in multi-organ segmentation tasks. With the advance of contrastive learning (33), a variant of self-supervised learning that learns the meaningful representations by pulling similar representations together and pushing dissimilar representations away, it demonstrates feasibility to define feature embeddings with semantic meanings introduced by ourselves. Another explorative direction can be focused on defining an explainable feature space that benefits the robustness of multi-organ segmentation task by introducing different meaningful context to classify embeddings.

With the enhanced quality of segmentation predictions from deep learning models, we then generate multi-organ segmentation for large population imaging samples from research cohort. We observe that substantial variation in morphology of different organs are demonstrated across populations due to the variable demographics (e.g., ages, sex, blood type). Although we know that the patients' condition is healthy from demographics, there is no standardized biomarkers in imaging level to define the healthy condition in organs. Therefore, a reference template is needed to generalize the anatomical information across population with preserved contrast characteristics (due to the variability of imaging protocols).

In this thesis, We aim to innovate explainable learning strategies to optimize deep neural network for

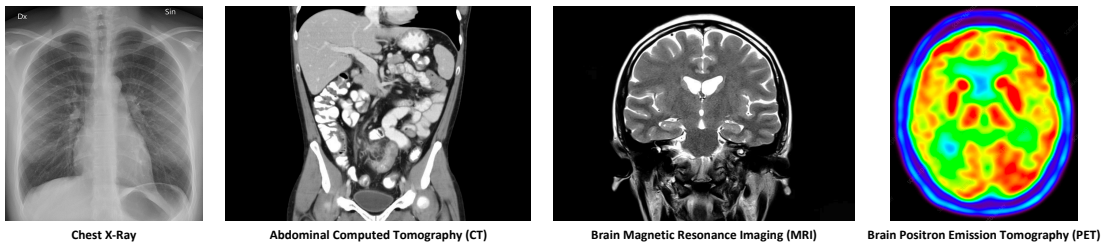


Figure 1.1: To visualize different anatomical regions with meaningful context, several imaging modalities have been widely leverage in current clinical scenario: 1) X-ray, 2) computed tomography (CT), 3) magnetic resonance imaging (MRI) and 4) positron emission tomography (PET).

medical image segmentation and generate organ-specific atlases to enhance the corresponding anatomical understanding. Novel approaches are used to increase the explainability of representations extracted and the computed gradient in deep learning models in theory. With the enhanced robustness of segmentation model across multi-modality imaging, we further leverage the predicted segmentation labels to generate anatomical reference template to provide a clearer organ-specific understanding across large-scale populations.

1.2 Medical Imaging

To visualize the interior of our human body, medical imaging is an essential platform that provides the fine-grained details of each organ/tissue structure. Common imaging modalities in current clinical scenario can be divided into multiple perspectives, such as 1) X-ray (207), 2) computed tomography (CT) (18), 3) magnetic resonance imaging (MRI) (213) and 4) positron emission tomography (159) (as shown in Figure 1.1). Within the imaging modalities, substantial differences are also demonstrated across organs/tissues within the perspectives of MRI by leveraging different protocols, as shown in Figure 1.2. Moreover, contrast enhancement is performed in the imaging procedures by injecting contrast agent (e.g., barium sulfate, iodine) into patient's body before imaging. Its goal is to enhance the anatomical details between neighboring organs with respect to the retaining time of contrast agent in blood flow (80). Such great variety of contextual information further allow clinicians to enhance the confidence level of classifying patient's conditions with imaging biomarkers. Five different contrast phases are typically defined in the imaging cycle: 1) non-contrast phase, 2) early arterial phase, 3) late arterial phase, 4) portal venous phase and 5) delayed phase (as shown in Figure 1.3) (46). The variation of the contrast intensity benefits to capture and specify the fine-grained details in each organ. For instance, both early and late arterial phases abdominal CT can demonstrate a higher intensity contrast for aorta and the internal substructures of kidneys (201), while portal venous phase CT demonstrates additional contrast respect to hepatic vein and liver (69). By contrast, non-contrast abdominal CT is also

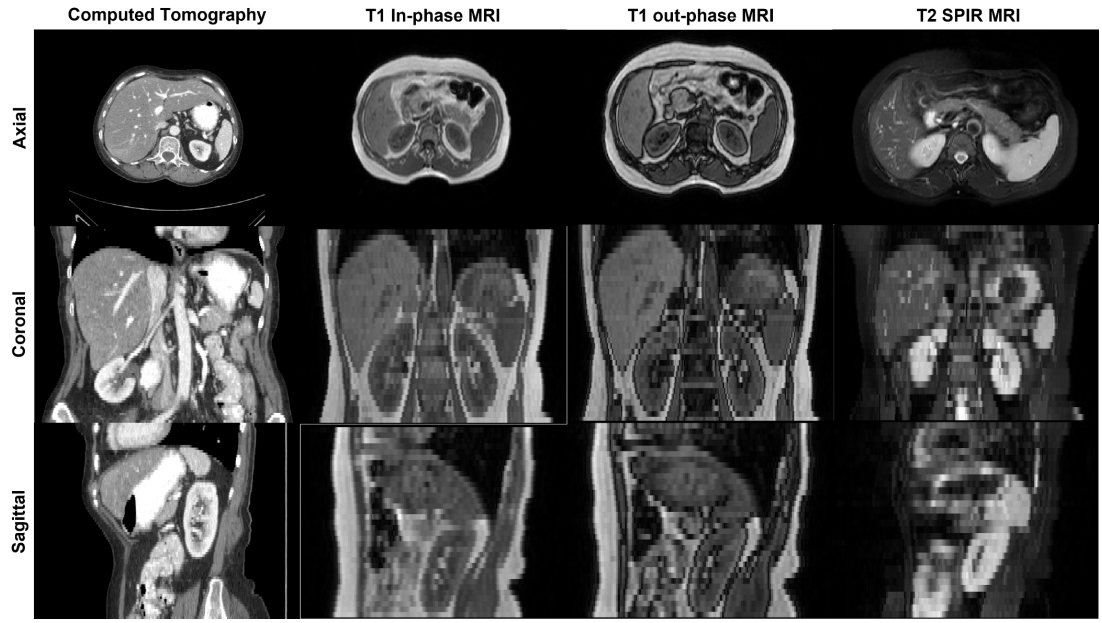


Figure 1.2: With different imaging protocols, substantial intensity differences are demonstrated across organs/tissues, even within the perspective of MRI modality.

widely and conveniently available for broadly usage in clinical scenarios (162; 191). It is routinely acquired as a diagnostic modality for detecting renal stone or intramural hematoma in aorta organ (59; 193).

1.3 Medical Image Segmentation

With the enhancement of contrast intensity towards different organs, we can localize each organ of interest with the corresponding anatomical and structural details. However, it is challenging to automatically extract the contextual information of organs such as volumetric and morphological measures, because there is no additional pixel/voxel-wise guidance to extract the organ regions specifically. To generate accurate pixel/voxel-wise localization for each organ, image segmentation is performed to generate a one-hot mapping that bound the region of interest for localization (169), as shown in Figure 1.4. It is vitally important to have an accurate organ segmentation mapping to extract meaningful context for further investigation, such as verifying patients' condition in liver after performing liver transplant. However, the most naive way to generate segmentation is to manually annotate the region of interests, which consumes significant amount of time and efforts for clinicians. To reduce time of efforts, mathematical algorithms and machine learning technique are proposed to enhance the automatic segmentation with robustness (39; 182; 45; 150). Popular medical image segmentation tasks have been performed across scale, such as multi-organ segmentation in abdominal region (125; 198), liver and tumor segmentation (139; 15), brain and tumor segmentation (172; 75; 103; 230; 218; 73)

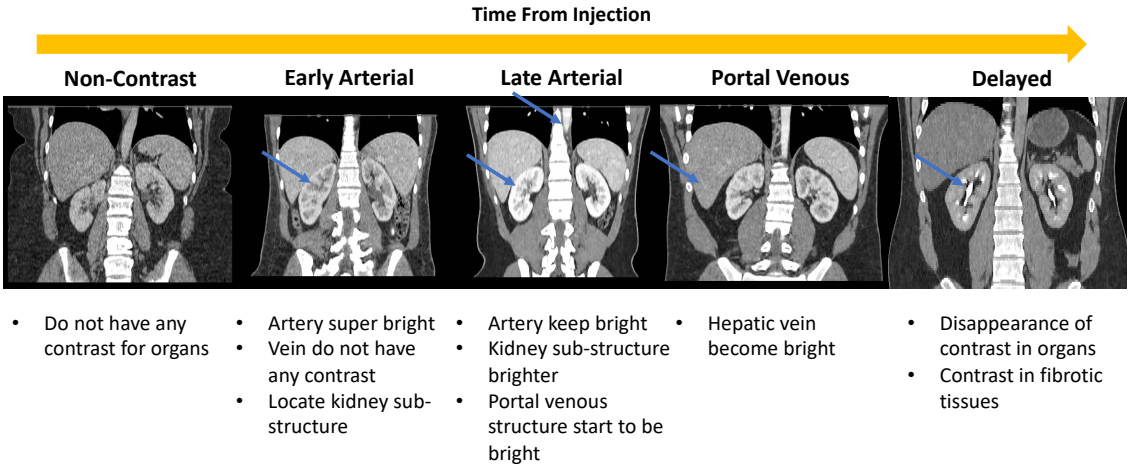


Figure 1.3: By the time of injecting contrast agent for contrast enhancement, the contrast intensity of particular organs are enhanced, which provides feasibility to localize clear anatomical context between neighboring organs.

and cell segmentation (149; 13). Early approaches in medical image segmentation depends on the edge detection (239), template matching techniques (116), statistical shape model (166) and active contours (194). To enhance the robustness and computation efficiency for segmentation, machine learning techniques are following with the metric-based approaches and introduce to extract hierarchical feature representations of images and improve the segmentation performance. With the introduction of machine learning concepts, the perspectives of automatic segmentation are further divided into two categories: 1) semantic segmentation (30; 31; 133), and instance segmentation (28; 231). Semantic segmentation is a pixel/voxel-level classification that assigns a corresponding category to each pixel / voxel in an image, while instance segmentation further distinguishes multiple instances on the basis of specific categories.

1.4 Organ Segmentation with Machine Learning

In each region of interest in the human body, complicated spatial relationship exists between neighboring organs and tissue structures. It is challenging to explore the feasibility of organ localization without addition guidance. Therefore, an accurate segmentation algorithm is demanded to extract contextual information across scales and further perform population-wise clinical measurements with the segmentation guidance. Deep learning, one of the evolutional machine learning techniques, have contributed significant efforts in generating automatic segmentation throughout 2020s, especially for medical imaging (120). The simple insights behind deep learning is to generate a network to model human neural system and the basic unit of transferring the information throughout the network is a neuron. In biology, each neuron receives signals with

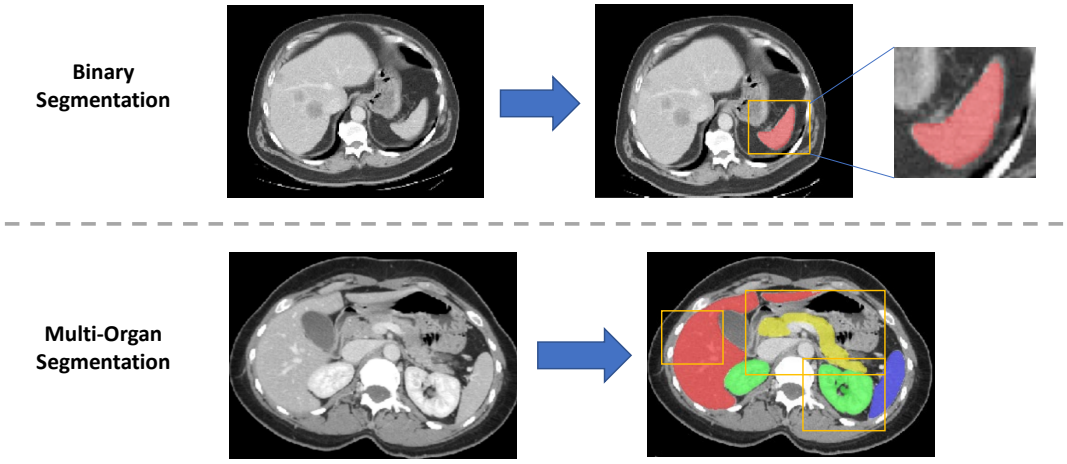


Figure 1.4: Medical image segmentation consists of two main perspectives: 1) binary segmentation for single organs and 2) multi-organ segmentation with multiple semantic values (e.g. 0, 1, 2, 3, 4).

multiple inputs from dendrites and computes an output signal through axon. The output signals are eventually branched out and connects to the dendrites in other neurons through synapse. For computational model, we model the input signal as x_0, which interact multiplicatively with other dendrites throughout a synapse w_0. The synapse portion is assumed to be learnable and influence the learning process with iterative optimization. In practical perspective, deep learning is to learn distinctive features from a population of samples (regard as training data) and perform predictions to unseen samples (regard as testing data). In summary, we further divide the directions of current deep learning approaches into four perspectives:

1. Supervised Learning: leverages imaging samples with good quality labels for training
2. Unsupervised Learning: leverages imaging samples only without using labels for training
3. Semi-Supervised Learning: leverages both labeled and unlabeled imaging samples for training
4. Self-Supervised Learning: learns meaningful features from unlabeled imaging samples and further refine features for specific downstream task

Briefly, supervised learning is a direct learning strategy that depends on the nature of the label given. For instance, one-hot labels (e.g. 0, 1, 2) are given to learn for classification task, while one-hot pixel/voxel mappings are used to provide learning guidance for segmentation task. For unsupervised learning, it mainly focused on feature clustering to perform classification and training generative model to synthesize images in different domains. Both supervised and unsupervised strategies are demonstrated to be task-oriented, while both semi-supervised and self-supervised learning are the ways that tend to enhance the robustness and

generalizability of the model. Before we deep more further into the current state-of-the-art training strategies, we first look into the backbone of deep learning and the way to adapt volumetric medical images for different downstream tasks.

1.4.1 Convolutional Neural Network (CNN) Approaches

To train a task-specific model, we need to extract meaningful features from images. However, the most naive approach is to transform the complete image into a 1-D feature vector (225), which contains a lot of contextual information that may not relate to the task and needs high computational power for training. Therefore, convolutional neural network is proposed to leverage convolution kernels that slide along the complete image and extract a static value for each kernel-bounded region as summarized features with scalable efficiency (120). A simple convolutional network consists of a sequence of layers such as convolutional layer, pooling layer, normalization layer, activation layer and fully-connected layer. Different combinations of network design have demonstrated to be beneficial to the downstream tasks. A lot of well-known network designs have been proposed for specific tasks (e.g. ReseNet (79), DenseNet for classification (94), U-Net for segmentation (177)). However, for medical imaging, it is reconstructed in a volumetric setting, which contains rich spatial context between organs. An extension of 3D convolution is leveraged to perform in medical task, especially for semantic segmentation (39). The volumetric images are downsampled to low resolution and directly input into a 3D network for training. However, such downsampling approach may introduce fuzzy interpolation operations, which contribute to the loss of anatomical details across organs. To preserve the high-resolution context in images, patch-wise and coarse-to-fine approaches are followed to propose and adapt features across scales (e.g., local to global) (182; 259; 200; 125). Throughout the advancement of CNNs, we observe that small convolution kernel sizes have been continued to leverage throughout different well-known block designs along 2020s. Such small kernel sizes only contribute the learning convergence in locality and have limited Effective Receptive Field (ERF) to extract spatial features from volumes with different Field Of Views (FOVs) (79; 112; 94). However, if we increase the kernel sizes in convolutional layers, both the model parameters and computational efficiency significantly enhance, which is challenging to be practical usage (196). In order to tackle this limitation alternatively, the concepts of transformer in NLP domain are introduced into the visual cognitive task and leverage pure attention schemes to enhance the spatial feature correspondence across imaging samples.

1.4.2 Vision Transformer Approaches

As transformer networks provide different computational advantages in NLP tasks, Dosovitskiy et al. proposed vision transformer (ViT), an extenion of the NLP transformer, which transform sequences of image

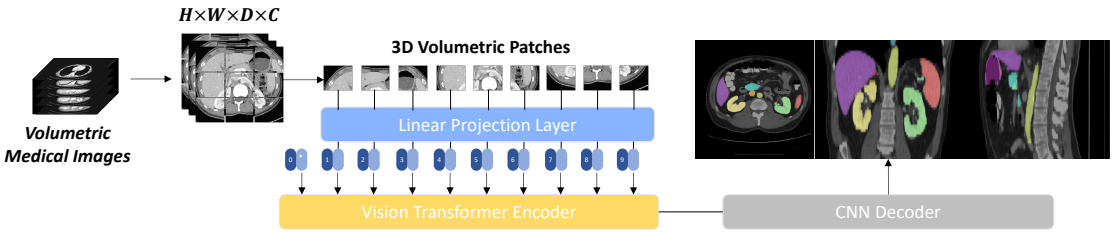


Figure 1.5: With the introduction of vision transformer, volumetric images are divided into volumetric patches in specific dimensions (e.g., $16 \times 16 \times 16$) and are projected as a one-dimensional feature vector with a linear layer for input. The vision transformer modules are then leveraged to be the encoder backbone following with a CNN decoder via skip connections for volumetric segmentation.

patches into sequences of linear embeddings and perform attentional mechanism to extract correspondence between embeddings, instead of the convolutional operations from CNNs (52). More specifically, as shown in Figure 1.5, an input images is split into a set of image patches with a dimension of 16×16, and also termed as visual tokens. The visual tokens are then projected into a set of encoded vectors with fixed dimensions using a Multi-Layer Perceptron (MLP). Meanwhile, a position encoding vector is concatenated with the encoded vectors to preserve the positional context during the attention mechanism. The attention map are computed via a multi-head attention network and further generate an output prediction with a MLP two-layer classification network. The vision transformer claimed to effectively model the long-range relationships and dependencies between visual elements with limited inductive biases in the visual domain. In medical domain, the growth of interest in adapting transformer network is also demonstrated (188; 78; 138). The lack of image-specific inductive biases and the scaling behaviour in ViTs claim to be enhanced by pre-training and bring significant improvement with downstream classification tasks. However, a vanilla ViT faces challenges when it is applied to other visual cognitive tasks such as object detection and semantic segmentation. The biggest challenge in ViT is the design of computing global attention, which has a quadratic complexity with respect to the input image sizes, especially for high-resolution medical images with dense features across scales. Therefore, hierarchical transformers are proposed to bridge these gaps by introducing the sliding window strategy (e.g. attention within local windows) into ViTs termed Swin Transformer, allowing to behave similarly with the convolution operation in CNNs (145; 21; 104; 73; 253), as shown in Figure 1.6. The integration of Swin Transformer into medical domain with different designs (e.g., SwinUNETR, nnFormer) achieves the state-of-the-art performances on multiple 3D volumetric datasets benchmark in semantic segmentation. The key contribution of such significant performance enhancement is divided into two perspectives: 1) scaling behavior with large model parameters and 2) global self-attention with large receptive fields.

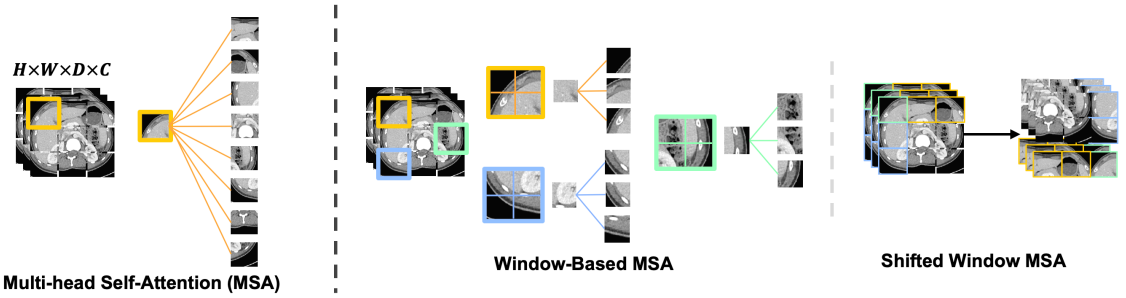


Figure 1.6: To tackle the limitation of vanilla vision transformer block in high-resolution task, hierarchical transformer is proposed to adapt window-based Multi-head Self-Attention (MSA) and shifted-window Multi-head Self-Attention to extract fine-grained representations, introducing the prior knowledge from the convolution kernels into transformer backbone.

1.4.3 Challenges & Exploration

From the milestone from CNNs to ViTs, we observe that the intrinsic structure of transformer still needs the integration of prior knowledge from CNNs to enhance both the robustness and generalizability in high-resolution prediction. Also, the computation of attention with shifted window strategies are unscalable and pretraining may need to perform for enhancing model stability due to limited inductive biases. With the summarized key contribution in ViTs, the behavior of self-attention scheme with large receptive fields can be demonstrated with convolution operation using large kernels, while multiple challenges existed to enhance the feasibility of adapting large kernel convolution efficiently. An efficient way to perform large kernel convolution is highly demanded to compute scalable operations with inductive biases. With the revisit of depthwise convolution, it enables to reduce the model parameters significantly by computing convolution along only each channels independently, instead of traditional convolution across all channels to mix features. Meanwhile, it allows to extend current convolution block designs to adapt large kernel size (e.g., 31×31 (48), 51×51 (143)) as a generic backbone in natural imaging domain. However, limited findings have been proposed to explore the feasibility of leveraging large kernel convolution in volumetric downstream tasks in medical domain.

Another perspective that has not been explored is the maximum kernel size for volumetric convolutions without trading off computational efficiency. From previous works in natural imaging domain, the kernel size of the convolution have been expanded to a size, which is approximately equal to an image patch. Meanwhile, the model performance becomes saturated and even degraded until the kernel is scaling up to a particular size. Such phenomenon is due to the limited learning convergence in locality and the concept of structural re-parameterization is introduced into CNNs (91; 49). Combining the weights from different branches with small kernels, can enhance the convergence of locality learning and improve the downstream

task performance with a more widely distributed effective receptive field. However, the parallel branches may hinder the training efficiency and it is challenging to optimize extreme large kernel convolution without addition learning guidance.

1.5 Explainable Machine Learning in Medical Domain

Apart from looking into the generic design of the neural network backbone to adapt robust organ segmentation, another important aspect in the medical domain is to explain the ongoing process in the neural network to the clinical teams. Modern machine learning models are considered as "black boxes", referring to the high difficulty in understanding how they arrive at their predictions. The "black box" characteristics may lead to a problematic consequences of the significant errors or bias of the model. Instead of only improving the model performance across different imaging cohorts, we need to additionally justify the question of "Why the model works", especially in the medical domain. Explainable machine learning, also known as interpretable machine learning, is proposed to investigate the ability of the machine learning models to provide justifications for its predictions. Significant efforts have been put to widely adopt and standardize approaches for model interpretability. In summary, two major perspectives can be divided: 1) perceptive interpretability and 2) theoretical interpretability (204). The perceptive interpretability is to consider the generated predictions that can be immediately interpretable (e.g., Class Activation Maps (CAM) (252), Local Interpretable Model-agnostic Explanations (LIME) (173; 153) and Layer-wise Relevance Propagation (LRP) (118)), while the theoretical interpretability is to consider the generated intermediates with more than one layer of cognitive processing and provide the corresponding meanings with mathematical guidance.

In medical image segmentation perspectives, efforts have been shown to investigate the direction of adapting saliency or uncertainty mapping into the segmentation network and explain the uncertain boundary prediction from the model. However, limited works have been proposed to enhance the interpretability of the learned representations during the model training process, especially for multi-organ segmentation. When we train a model for single organ segmentation, we can direct hypothesize that the learned feature is corresponding to the segmented targets, as shown in Figure 1,7. However, to train a multi-organ segmentation model, still only one feature mapping is generated for each input and concretely summarize the representations for multiple organs. Such single high-level feature is challenging to directly correspond the learned feature to each specific organ and limits to explain each targeted organ semantic meaning in the feature level.

1.5.1 Defining Latent Space with Contrastive Learning

In natural image domain, contrastive learning, a variant of self-supervised learning strategy, is proposed to pre-identify the feature into different clusters by evaluating the similarities and difference between data

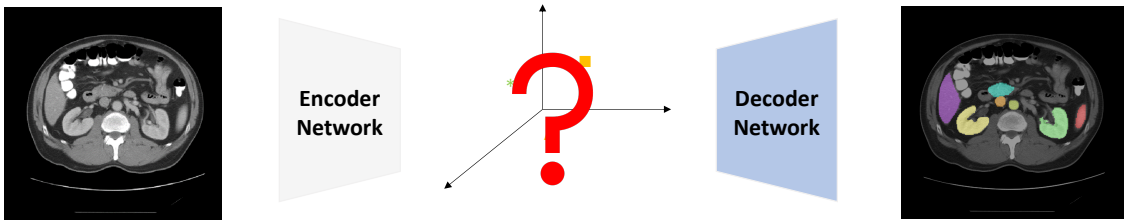


Figure 1.7: For multi-organ segmentation task, we can easily evaluate the robustness of the model with different evaluation metrics. However, it is challenging to interpret the fine-grain details of the semantic meanings in feature level. As each input image are then encoded as a high dimensional feature map only, it is difficult to explain the correspondence of the feature map and the organ semantics in high dimensional latent space.

samples (33). The goal of contrastive learning is to learn the distinctive representations that capture the underlying structure between samples and use the same set of samples for downstream tasks (e.g., image classification, semantic segmentation) (237; 219). Meanwhile, we further observe that the semantic meanings can be further introduced in the contrastive learning process and classify the cluster with different semantics, thus to enhance the feasibility of tackling the limitation in feature interpretability of multi-organ segmentation network (23). The key theoretical concepts consists of two aspects: 1) pulling a target image feature and a similar sample feature together as a ”positive pair”, and 2) pushing the target image feature apart from numerous dissimilar samples’ features as ”negative pairs”. Each feature output from the encoder network are projected into L1-normalized space and evaluate pairwise cosine similarity within each minibatch. One common approach to contrastive learning is to adapt Siamese network architecture, where two copies of the same network are leveraged as the pairwise features of each sample. The pairwise features are then evaluated with different combination of similarity metrics (e.g., cosine similarity, mutual information) to determine whether the pair is positive or negative (70; 35). Another well-known genric backbone for contrastive learning is to adapt data augmentation techniques to generate multiple views from the same sample. The trained model aims to maximize the similarity between views of the same sample, while minimizing the similarity between views of different samples (33; 108). In summary, with the key advantages of contrastive learning, it provides feasibility to create a latent space with multiple semantic meanings to enhance the interpretability of the learned feature without any label guidance.

1.5.2 Contrastive Learning for Medical Image Segmentation

The initial work efforts have been put to leverage contrastive learning to enhance "image-wise" downstream task, such as image classification (33; 108; 206). Several works start to extend the ability of contrastive learning into pixel-wise setting for enhancing high-resolution prediction task. After we generate the feature representations from the encoder network, the concepts of "image-wise" contrastive learning is to perform average pooling and flatten the feature representation into 1-D vector, while the idea of "pixel-wise" contrastive learning is to compare each feature pixel with dense projection across sample pairs. Furthermore, to enhance the correspondence between features in each semantic class, ground-truth labels are leveraged to sample the feature pixel in the latent space. The feature pixels in the same class are classified as "positive pairs", while the feature pixels in different classes are defined as "negative pairs" (219; 90; 248). However, such feature sampling strategies need additional guidance from the ground-truth label. Also, each feature pixel may consist of the semantic meanings with multiple organs, as the convolution summarize the region context as a static value according to the convolutional kernel sizes. With the limited availability of ground-truth pixel-wise labels, it is challenging to adapt accurate correspondence between pixel-wise features and organ semantic meanings and enhance the feature interpretability for further clinical analysis.

1.5.3 Challenges

With the introduction of the contrastive learning concept, it brings out multiple advantages to enhance the stability and interpretability of deep learning models, such as stable performance with limited samples training, classifying representations into semantic interpretable clusters. However, the proposed strategies are usually focused on image-level representations and classify the image representations into initial semantic clusters. Furthermore, the input image for contrastive learning is in object-centric setting, which demonstrates that only one object exists within the image (e.g., airplane, car, dog, cat). Therefore, the models can easily extract the object-semantic representation and classify into corresponding clusters. We hypothesize that the conceptual idea of contrastive learning can further extend the capability of sub-image-level feature encoding, to advance pixel-level segmentation tasks. However, some gaps need to be filled to achieve the latter goal, especially for multi-organ segmentation tasks in medical imaging. Volumetric imaging (e.g., CT, MRI) usually exists multiple objects in different anatomical regions. For instance, spleen, liver, kidney and stomach may exist in the same axial slice of an abdominal CT scan. Multiple semantic meanings (e.g., different organs) are difficult to align with the single feature extracted from a complete image.

Apart from adapting multiple semantic meanings, when we further observe the fine-grained details across different organs in medical imaging, we found that the contrast intensity are significantly varied in large range of level due to the contrast enhancement procedures with different protocols. For instance, the contrast

appearance of liver is significantly different between portal venous and non-contrast phases, while the contrast appearance of gall bladder is similar even thought they are in different phases. With the inspiration of previous work in contrastive learning, image-wise labels (e.g., 1: spleen, 2: right kidney) have been introduced to assign the organ semantic meanings with the learned feature clusters. However, such fix labels are limited to represent the dynamic changes of the contrast intensity within organ ROIs. It is challenging to extract such dynamic knowledge to control the contrastive learning process without additional guidance.

1.6 Generalizing Organ Context across Population

After we robustly trained the segmentation model with interpretability, we can then perform inference to the unseen domain samples and generate segmentation mapping to localize each organ ROIs. However, we observe that the quantitative measures of each organ segmentation varies significantly across population. With the substantial variation in demographics, such as ages, height, weight...etc, the volumetric morphology of each organ significant differ across patients. Such large range of difference is limited to provide a clear understanding of defining biomarkers in conditions. For instance, we may raise a question: What is the biomarker of kidney organs in healthy condition? It is challenging to standardize the population context of specific organs across multiple volumetric scans with different imaging protocols.

1.6.1 Anatomical Atlas Reference

To standardize the population context of organ ROIs, we usually define an anatomical atlas template to align and generalize the anatomical details of each subject scans. An anatomical atlas is a standardize representation of specific anatomical structure and leveraged as a common spatial framework for comparing and analyzing different imaging data, such as CT and MRI. Such standardized template is typically created by combining the images to common coordinate system using mathematical transformations. After the forward transformations of each subject scans to the atlas template, we compute an average representation of the anatomy and demonstrate the distinctive contrast appearance across all scans. Meanwhile, the anatomical atlas typically consists of the voxel-wise label of the corresponding anatomy, such as brain tissues (60; 189), heart and major blood vessels (152). Using the atlas template can enable researchers and clinicians to analyze different imaging data in a standardized and consistent manner, without having fluctuated performance to identify the specific organ ROIs like deep learning models.

1.7 Contributed Work

Significant efforts have been directed toward integrating machine learning into visual recognition tasks in medical domain, especially for medical image segmentation. Image segmentation provides pixel/voxel-wise

localization of each organ/tissue target and allows researchers and clinicians to perform quantitative measures for investigating biomarkers. However, annotating volumetric labels is time-consuming. There is limited interpretability of features across learned current models. Furthermore, the volumetric morphology of organ of interests significantly varies with the substantial variability of demographics across population. Such variability limits the feasibility of generalizing population-wise features on specific organs to further investigate the corresponding biomarkers in conditions. In this dissertation, we first investigate training strategies to enhance the robustness in deep learning models across multi-modality imaging. We were inspired by the current progress of vision transformers and revisited the capabilities of large kernel convolution. We model the spatial frequency of the human visual behavior and derive a theoretical re-parameterization strategy to enhance both the learning capability and explainability of large kernel convolution for volumetric segmentation **(Contribution 1)**. With the basis of current network designs, we investigate self-supervised learning strategy to dynamically adapt multi-contrast imaging and integrate multiple semantic meanings to enhance the feature explainabilty, thus enhancing model generalizability with interpretable features **(Contribution 2)**. With the generated segmentation across large population cohort, we create multiple atlas templates to generalize the population characteristics across organs, and thus to facilitate the progress of revealing distinctive organ-specific biomarkers and align multi-scale findings from cellular level to organ level **(Contribution 3)**.

1.7.1 Contribution 1: Explore Deep Learning Optimization for Medical Image Segmentation

1. We proposed a coarse-to-fine hierarchical network to integrate the coarse segmentation output from low-resolution model as prior knowledge and refine the segmentation prediction within the coarse prior-bounded region.

2. We further extended the coarse-to-fine hierarchical network with vision transformer as the generic backbone and employ teacher-student structure to adapt cross-domain samples with unsupervised training.

3. We proposed a semi-supervised learning strategy to adapt large population of unlabeled samples and leverage a segmentation quality score prediction as an unsupervised loss function to backpropagate for unlabeled samples.

4. We explore the key contribution from vision transformer and proposed a large kernel convolution neural network to modernize the capabilities of hierarchical transformer for volumetric segmentation efficiently.

5. We further explore the feasibility of adapting convolution with large kernel sizes (e.g., $21 \times 21 \times 21$) for medical image segmentation. We simulate the spatial frequency in the human visual behavior as

a Bayesian prior knowledge to enhance the learning convergence for large kernels, thus to efficiently enhance the robustness with large receptive field for volumetric segmentation.

1.7.2 Contribution 2: Enhance Feature Interpretability for Medical Segmentation Network

1. We proposed a semantic-aware contrastive learning framework that integrates semantic meanings and classifies the learned feature into clusters with multiple semantic meanings, thus to enhance the feature interpretability in segmentation network.

2. Instead of using multiple fixed labels (e.g., 0,1,2) to assign the semantic meanings into the learned features, we control the cosine distance separation between pairwise features with a correlation weighting computed by the contrast level differences across organs. Such dynamic characteristic further leverages as a prior weighting to enhance the model generalizability across multi-contrast CT.

1.7.3 Contribution 3: Generalize Population-wise Biomarkers with Organ-specific Atlas

1. We constructed multi-contrast atlas template to generalize both the contrast characteristic and the variable volumetric morphology of kidney organs in healthy conditions across population.

2. We further extended to generate atlas template with arterial phase CT and generalized the fine-grained details of the kidney substructure organs.

3. We constructed unbiased eye atlas template with both metric-based and deep-learning based hierarchical registration to adapt the variability of eye organs across population.

CHAPTER 2

RAP-Net: Coarse-To-Fine Multi-Organ Segmentation With Single Random Anatomical Prior

2.1 Overview

[1] Performing coarse-to-fine abdominal multi-organ segmentation facilitates extraction of high-resolution segmentation minimizing the loss of spatial contextual information. However, current coarse-to-refine approaches require a significant number of models to perform single organ segmentation. We propose a coarse-to-fine pipeline RAP-Net, which starts from the extraction of the global prior context of multiple organs from 3D volumes using a low-resolution coarse network, followed by a fine phase that uses a single refined model to segment all abdominal organs instead of multiple organ corresponding models. We combine the anatomical prior with corresponding extracted patches to preserve the anatomical locations and boundary information for performing high-resolution segmentation across all organs in a single model. To train and evaluate our method, a clinical research cohort consisting of 100 patient volumes with 13 organs well-annotated is used. We tested our algorithms with 4-fold cross-validation and computed the Dice score for evaluating the segmentation performance of the 13 organs. Our proposed method using single auto-context outperforms the state-of-the-art on 13 models with an average Dice score 84.58% versus 81.69% ($p < 0.0001$).

2.2 Introduction

Creating a robust and accurate pipeline for volumetric abdominal organ segmentation in computed tomography (CT) domain is challenging. Deep neural networks have been used for performing sematic segmentation in medical imaging perspective with 2D images or 3D patches, such as DeepLab (30), UNet (177) and VoxResNet (27). Current major challenges of abdominal organ segmentation are: 1) weak intensity boundaries between abdominal organs, 2) large morphological variation of different organs, and 3) high resolution of 3D volumes.

To overcome the limitation of using 2D CNNs, 3D volumes are sliced in axial, coronal and sagittal directions and used for coarse organ detection and perform segmentation with the detection output (256). Both 2D and 3D-patch based learning methods target the single organ approaches. Roth et al. proposed coarse-to-fine pipeline with scaling input volumes at different levels using multi-scale pyramid networks and computed refined prediction at the last selective level for abdominal multi-organ segmentation (182). Zhu et. al. added expanded bounding box into a coarse-to-fine pipeline to abstract the ROI of small targets and

[1] Published at: Lee, Ho Hin, et al. "Rap-net: Coarse-to-fine multi-organ segmentation with single random anatomical prior." 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021. (125)

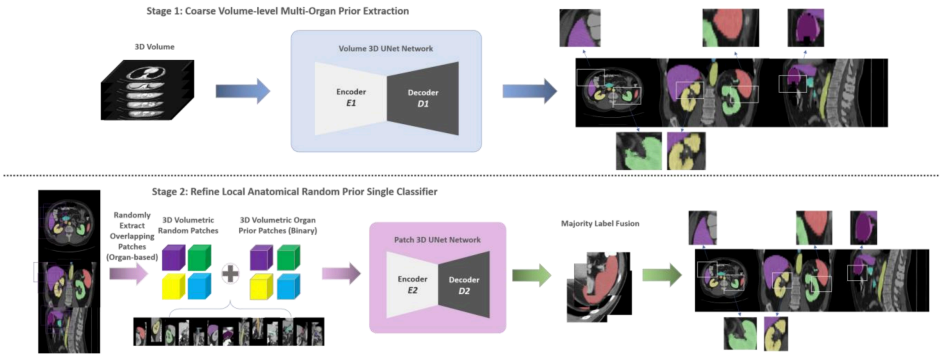


Figure 2.1: Overview of the pipeline combining the global and local level representations. The proposed hierarchical coarse-to-fine multi-organ segmentation framework consists of two main stages: 1) coarse global anatomical multi-organ prior extraction and 2) a refined single classifier with local anatomical random prior. The coarse- and refined-model are end-to-end optimized separately. The predicted segmentation patches of each single organ are finally fused as multi-channel volume and converted to multi-organ segmentation mask with majority voting.

improve robustness for refining single organ segmentation (259). However, numerous organ-corresponding models are needed to be trained to obtain refine segmentation performance. Significant effort is allocated for tuning hyperparameters, and there is a lack of flexibility to combine contextual information across scales due to high complexity of training strategies. Backpropagating multiple loss functions for all organs to obtain a single model is not possible due to the missing organ regions in the extracted patches. Therefore, a single high-resolution refine model framework integrating global and local contextual information to refine segmentation is needed.

To address the challenges, we introduce a 3D hierarchical coarse-to-fine framework RAP-Net, performing multi-organ segmentation at refined level with single model. Briefly, we initially down-sampled the volumetric images and used a traditional approach to generate a coarse segmentation for each organ. The coarse segmentation then acts as an anatomical prior for each organ and extracted corresponding organ patches using the prior information. We transformed the corresponding organ coarse segmentation to binary anatomical prior and integrated as the second channel input with the image patches. The refined model is then trained with all binary labeled organ patches end-to-end. Such training strategies help the refined model encode the variability of shape and local intensity across all organs and limit the segmentation region with the anatomical prior, generating a robust and accurate performance for multi-organ segmentation.

2.3 Materials and Methods

Our work aims to create a single model adapting high-resolution multi-organ context across global and local levels. As shown in Figure 2.1, the backbone of the hierarchical approach is based on the state-

of-the-art (200), which consists of 13 refined organ models separately capturing the local representation of each corresponding labeled organs. We propose a refined organ-voxel classifier that maintain the ability of classifying the morphological variations and abstracting the local contrast characteristics across all organs. By utilizing the anatomical prior information extracted from all organs, the extracted local representation from the single patch-wise multi-organ segmentation model integrates all organs' local intensity feature and the global morphological information, to generalize the structural variability and eliminate the possibility of over-segmenting towards neighboring organs. The proposed code is available at https://github.com/MASILab/coarse_to_fine_prior_seg.

2.3.1 Data

A clinical research cohort with 100 patient volumes in portal venous phase was retrieved in de-identified form under the approval of the local IRB (institutional review board). The range of slice numbers across all volumes in the cohort is between 42 and 149 with a dimension of 512×512. The resolution for x, y and z-axis are in the range of $0.5 - 0.9$ mm, $0.5 - 0.9$ mm and $2.5 - 7.0$ mm, respectively. We first randomly split the dataset into 80 volumes as the training and validation set, and the other 20 volumes as the testing dataset. Each volumetric scan is manually annotated with 13 classes of multiple abdominal organs including spleen, right and left kidneys, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava (IVC), portal splenic vein (PSV), pancreas, right and left adrenal glands.

We initially down-sampled each volume to a resolution of $2 \times 2 \times 6$ mm and pad/crop to a constant dimension of $168 \times 168 \times 64$ as the input for the coarse stage segmentation model. The predicted multi-organ segmentation mask of each volume is then converted back to its original corresponding resolution and extracts volumetric patches with the coarse segmented mask in dimension of $128 \times 128 \times 48$ as the input for the refined stage segmentation model. The single refined model generates a 3D binary mask, corresponding to the organs extracted from the anatomical prior.

2.3.2 Global Anatomical Multi-Organ Prior Extraction

We adapt a volume-based 3D UNet architecture as the initial stage of RAP-Net and provide supervisory end-to-end optimization to achieve a coarse level multi-organ segmentation (39). The modified network consists of 8 encoders with convolutional kernel size of $3 \times 3 \times 3$, batch normalization layers, and 10 decoders with deconvolutional kernel size of $2 \times 2 \times 2$. Skip connections are used to integrate and capture the small variant representations from encoder blocks. ReLU activation units are used in both encoder and decoder blocks. The global representations can then be abstracted among all organs with the integration of high-level features from encoder to decoder. For multiple classes of abdominal organs A, we define the output of the final layer

from 3D UNet as $d_0 \in R^{((C-1)\times S\times H\times W)}$, where H, W, S and C are the number of height, width, slices and the channel number of the predicted multi-organ segmentation. A softmax activation is used to compute the probability map of the predicted segmentation p(A)= softmax(d_0) for each voxel. Each value of p(A) is extracted and compared with the similarity with the ground truth multi-organ label with multi-source Dice loss (MSDL).

$$MSDL = -\frac{2}{A}\frac{\sum_{a=0}^{A} w \sum_{i=1}^{M}\sum_{j=1}^{N} S_{ij}V_{ij} + \phi}{\sum_{a=0}^{A} w \sum_{i=1}^{M}\sum_{j=1}^{N} S_{ij} + \sum_{a=0}^{A} w \sum_{i=1}^{M}\sum_{j=1}^{N} V_{ij} + \phi}, \tag{2.1}$$

where A is the number of organ anatomies, w represents the variance between labels set properties. S and V provides the segmentation probability and the intensity of voxel belongs to the classes of organs respectively. A function ϕ is created to compute the correlation of the prediction and voxel value.

2.3.3 Single Classifier with Local Anatomical Random Priors

To capture the local variation in feature representation, 50 volumetric patches are extracted randomly according to the corresponding anatomical prior extracted from the coarse model, to ensure the overlapping region covering the complete volume of organ. The extracted patches include neighboring organ voxels apart from the main organ and lead to the adverse effect of segmenting main organ in corresponding patches. Here, we integrate the global anatomical prior with the volumetric patches as two channels input and input all prior-integrated organ patches for end-to-end training. We adapt the 3D UNet model architecture with the same network configuration of the coarse segmentation model. However, we modified the number of classes label in patches. Only the corresponding organ patch label is extracted as binary mask using anatomical organ priors as the label input to the refine model. We define the output of the final layer from 3D UNet model as $d_1 \in R^{(2\times S\times H\times W)}$, where H, W, S and 2 are the number of height, width, slices and the number of channels.

Training as single model framework provides an opportunity to adapt significant variation of the morphological and contrastive characteristics from large organs to small organs. Large numbers of organ-corresponding patches for training increases the generalizability of the model to segment all organs with prior information. As binary labels are used for all organs, the shape variation from small to large organs are also adapted to the feature representation for model to encode. With the integration of morphological and intensity variation characteristics across organs, the abstracted representations increase the localization ability of model. The integration of anatomical prior preserves the global anatomical location and the boundaries of specific organs in abdominal regions, while the random prior patches capture the large variability of shape and voxel intensities in the local regions, generating connective linkage between the voxel intensity variability and the morphological characteristics to stabilize the segmentation performance with the integrated representations from all organs.

Table 2.1: Dice score comparison of the testing cohort between coarse model, Roth et. al. (182), coarse + 13 refine models (state-of-the-art), Zhu et. al. (259) and our proposed model. Our proposed pipeline outperformed the other three coarse-to-fine methods with an average Dice of 84.58% (p¡0.0001, Wilcoxon signed-rank test).

Methods	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	LAG	Avg
Coarse only	0.921	0.828	0.894	0.695	0.670	0.935	0.783	0.885	0.814	0.627	0.690	0.604	0.615	0.770
Roth et al. ()	0.926	0.884	0.889	0.531	0.724	0.953	0.819	0.884	0.823	0.687	0.720	0.664	0.693	0.784
Tang et al. ()	0.939	0.900	0.943	0.763	0.712	0.952	0.822	0.897	0.828	0.710	0.745	0.646	**0.754**	0.817
Zhu et al. ()	0.961	**0.928**	0.932	0.693	0.772	**0.964**	**0.849**	0.913	0.837	0.698	0.762	0.684	0.721	0.824
RAP-Net (Ours)	**0.965**	0.920	**0.945**	**0.793**	**0.783**	0.960	0.833	**0.916**	**0.856**	**0.762**	**0.766**	**0.741**	**0.746**	**0.846**

Single channel Dice loss is used as the loss function for optimizing the binary segmentation for organ patches. The predicted segmentation for each patch is a binary mask corresponding to the specific organ. All binary masks from each organ patch are fused to generate a 13 channel multi-organ segmentation mask with majority voting.

2.3.4 Implementation Details

We performed 4-fold cross-validation with our labeled clinical cohorts to ensure both coarse and refine level model can capture the anatomical information of each organs. In total, 80 volumes of the clinical research cohorts are used for training and validation. 80 volumes are randomly shuffled and split to 4 groups of combinations with 60 volumes for training and 20 volumes for validation. The optimized model is chosen with the best validation performance for segmentation across all folds. The testing cohort is the BTCV MICCAI 2015 Challenge testing dataset with 20 volumes. The batch size was set to 1 for coarse volume-based model, while it was set to 2 for refine patch-based model. Adam was used as the optimizer for both stages end-to-end training. We first trained the coarse segmentation model with 100 epochs with learning rate of 0.0001 and choose the model with the lowest validation loss. For the single refine model, we directly input 39000 patches and trained the model for 5 epochs with learning rate of 0.0001. To evaluate the segmentation performance of both models, Dice score is used as the evaluation metric and compute a quantitative measure of the overlapping similarity between the prediction and the ground truth label. Subject volumes without gall bladder are eliminated in calculating the quantitative measures of gall bladder organ only.

2.4 Results & Discussion

Table 2.1 shows the quantitative comparison of the segmentation performance with Roth et. al., the state-of-the-art method, Zhu et. al. and our proposed model. The average Dice coefficient of all organ segmentation is increased from 81.69% to 84.58% and the standard deviation of Dice decreased from an average of 13.0% to 11.5% comparing to the state-of-the-art. RAP-Net demonstrated a better segmentation performance than

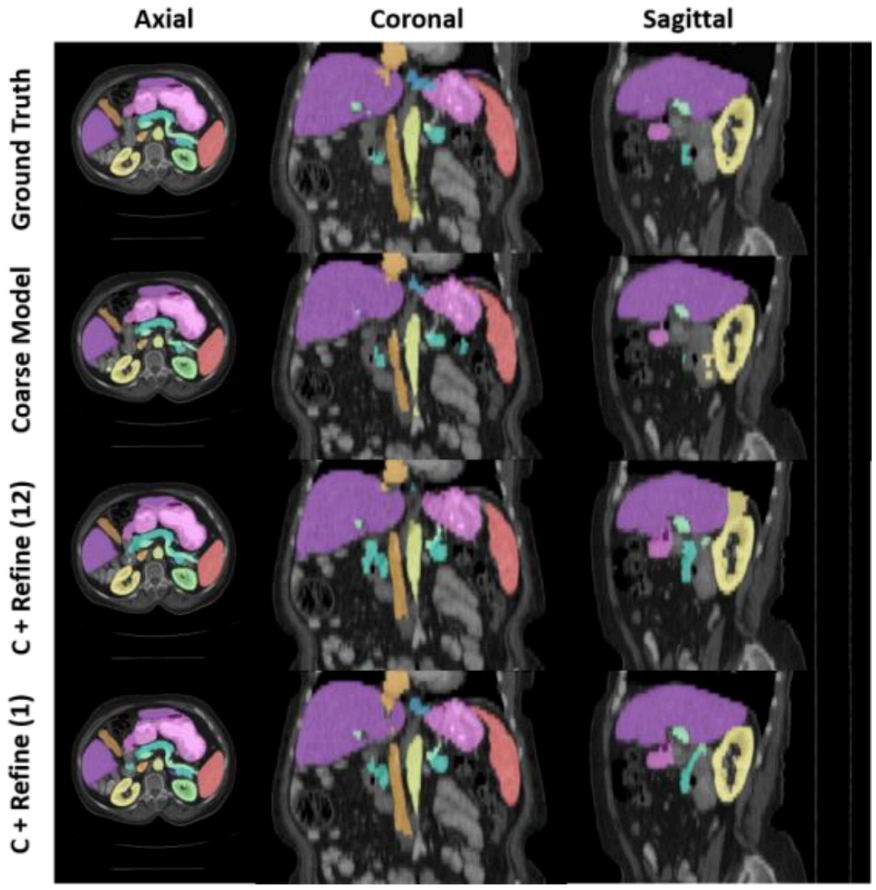


Figure 2.2: The qualitative representation of the multi-organ segmentation result with coarse level model, coarse (C) and 13 refine organ-specific (R) models, and coarse and 1 single refine model.

Zhu's method in 9 of 13 organs. The anatomical prior information provided a localization impact for improving the segmentation result. The single model encodes the variation across all organs with all organ patches directly training, while the 13 models only encode with the corresponding organ patches, the combination of intensity and morphology feature boost up the performance in the single model.

Figure 2.2 further demonstrated the confidence level of the quantitative result with qualitative representations. The segmentation result computed from coarse model can provide approximate anatomy localization information and the refine model can well adapt the organ regions for refining segmentation. The 13 organ corresponding models is well adapted with the anatomical information of the corresponding organ from the prior extracted in the coarse stage and abstracted the local intensity feature of each organ to refine the single target organ segmentation with the corresponding organ binary label. However, they only capture the corresponding morphological variation of the single target organ and learn the specific contrast characteristics within the prior preserved region, computing segmentations separately without anatomical linkage between

organs. Over-segmenting the neighboring organs still existed due to the similarity of neighboring organ voxel intensity. After we trained as a single model, the neighboring organ boundaries are substantially identified, as it encoded all organs' variation of the morphological changes in anatomical prior. The single refine framework provides an opportunity to control the refine segmentation in specific organ locations integrating with other organ localization information.

2.5 Conclusion

The proposed coarse-to-fine framework allows a single deep learning model to encode the integration of morphological and contrastive characteristics with multiple abdominal organs. Further studies will be performed by inputting multi-organ priors instead of one single corresponding organ prior, as other channels. The morphological linkage between organs can be evaluated in the future and innovate stratifying approach according to the anatomical characteristics between abdominal organs.

CHAPTER 3

Pseudo-Label Guided Multi-Contrast Generalization for Non-Contrast Organ-Aware Segmentation

3.1 Overview

[1]Non-contrast computed tomography (NCCT) is commonly acquired for assessment of general abdominal pain or suspected renal stones, trauma evaluation, and many other indications. However, the absence of contrast limits the ability of current deep learning (DL) models to distinguish organ in-between boundaries. Furthermore, current DL models are well pre-trained with public contrast-enhanced CT (CECT). Domain shift is introduced with NCCT due to the significant variation of contrast, thus demonstrates a decremental performance in organ segmentation. In this paper, we propose a novel unsupervised approach that leverages pairwise contrast-enhanced CT (CECT) context to compute non-contrast segmentation without ground-truth label. Unlike generative adversarial approaches, we compute a refine probabilistic mapping with the pairwise CECT to provide accurate anatomical correspondence, instead of generating fake anatomical context. Additionally, we further augment the intensity correlations in "organ-specific" settings and enhance the sensitivity in-between organ-aware boundaries. We validate our approach on multi-organ segmentation with paired non-contrast & contrast-enhanced CT scans (n=56) using five- fold cross-validation. Full external validations are performed on an independent non-contrast cohort (n=29) for aorta segmentation. Compared with current abdominal organs segmentation state-of-the-art in fully supervised setting, our proposed pipeline achieves a significantly higher Dice by 3.98% (internal multi-organ annotated, $p < 0.01$), and 8.00% (external aorta annotated, $p < 0.01$) for abdominal organs segmentation. The code and pretrained models are publicly available at https://github.com/MASILab/ContrastMix.

3.2 Introduction

Contrast-enhanced computed tomography (CECT) is a routine imaging modality to enhance the conspicuity of tissues of interest (80). Volumetric scans with variable contrast are reconstructed due to the injection of contrast agents and the retaining time of contrast agents in the blood. Specifically, CECT demonstrates a higher intensity contrast for several abdominal organs (e.g., aorta, liver, spleen, kidney) with better visibility. It further allows to map tissue permeability across anatomies (e.g. blood vessels), which could not be typically seen using regular non-contrast CT (NCCT) (in Figure 3.1) (223). Currently, significant efforts have been demonstrated to leverage deep learning (DL) models for abdominal organ segmentation on CECT (130;

[1]In submission at: Lee, Ho Hin, et al. "Pseudo-Label Guided Multi-Contrast Generalization for Non-Contrast Organ-Aware Segmentation.", 2023 IEEE Tranaction of Biomedical Engineering, 2023. (128)

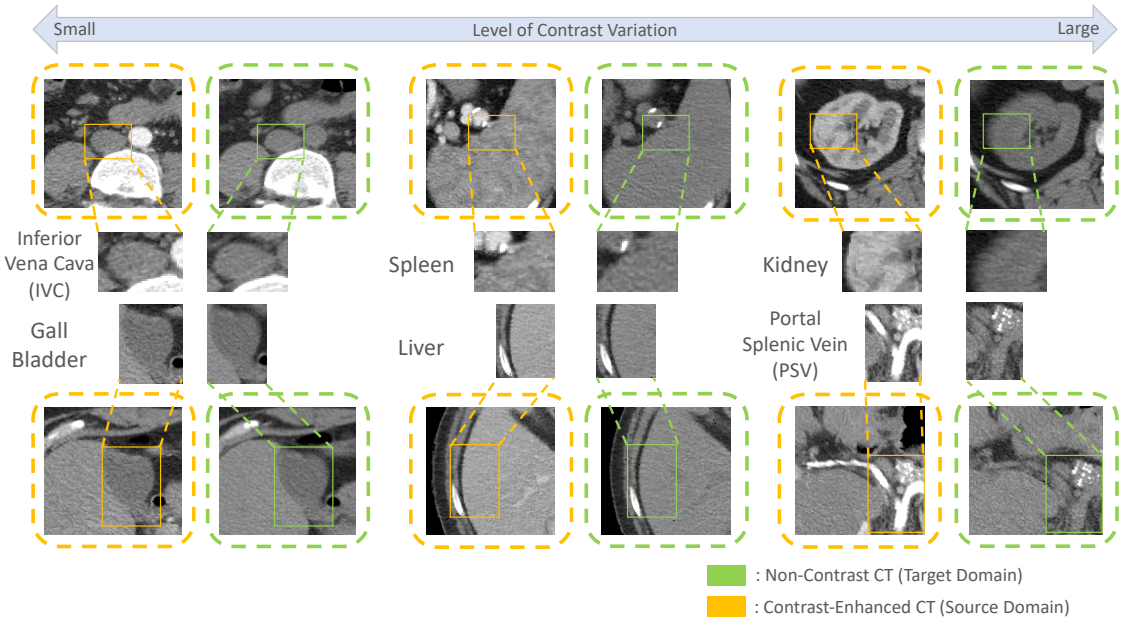


Figure 3.1: In multi-contrast phase CT, the contrast intensity varies according to the time of contrast agent retaining in the blood vessels. Different levels of contrast variation are demonstrated across organs. The robustness of model pretrained with CECT is limited to adapt the anatomical details in non-contrast phases.

125; 200). By contrast, relatively few prior studies have been proposed for abdominal organ segmentation on NCCT, especially those who boundaries are difficult to distinguish from the nearby tissues (e.g. aorta). Furthermore, the abdominal NCCT is more widely and conveniently available compared with CECT. For example, NCCT is broadly used and routinely acquired as a diagnostic modality for detecting renal stone or intramural hematoma in aorta (162; 62). Therefore, it is appealing to develop robust DL-based approaches and generate refine segmentation on NCCT in the absence of contrast between neighboring organs. However, a decrement of performance in NCCT scans is demonstrated by leveraging DL models that trained with CECT. This phenomenon is well-known as the limited generalizability of transferring DL models across different domains (14; 68).

In the natural imaging domain, the domain shift may contribute to different factors such as illumination, pose or image quality, which attribute to the degradation of model performance. However, such distribution shift is comparatively different than the modalities shift in medical imaging. The domain shift in medical imaging can be further classified into two main aspects: 1) **contrast intensity** and **shape morphology** (107; 56). By leveraging different protocols for image reconstruction, significant variations in intensity may exist across multiple CT scans. Previous works have been proposed to tackle the intensity shift into two main directions: 1) unsupervised domain adaptation (UDA) (63) and 2) multi-source domain generalization (MDG) approaches (158). Adversarial generative models, one of variants in the UDA approaches, are proposed to

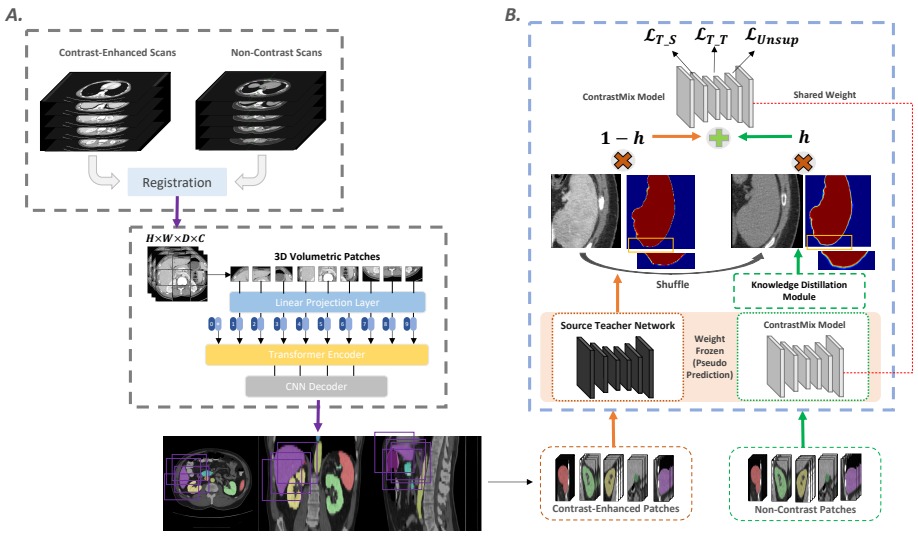


Figure 3.2: Overview of the proposed framework: **A. Minimize anatomical shift across domains:** Intra-subject registration is performed to minimize the anatomical variations between organs. We localize organ information with transformer-based network to adapt global correspondence of contrast intensity in each organ. **B. Generalize contrast variation across domains:** We leverage the contrast-enhanced prior context as teacher guidance to mix with distilled non-contrast context, thus to enhance the local boundary information in each organ.

perform image translation from the source domain to the target domain (97; 26). However, imbalanced training data exists between domains, which may lead to a significant loss of anatomical details and lack of stability in transferring fine-grained context in organ-level. It is also challenging to deploy the generative model in clinical practice due to its instability in adapting unseen domain samples. For MDG approaches, domain-invariant features such as shape topology are proposed to extract for better adaptation to unseen domain datasets (142). However, the anatomical context is not originally aligned in a well-defined spatial reference. It is challenging to provide dense correspondence of each independent organ across domains. With the above limitations to the current state-of-the-art (SOTA) approaches, we aim to have a segmentation framework, which can 1) **minimize the loss of anatomical context**, 2) **generalize the contrast variation across domains**; and 3) **learn the dense correspondence between organ anatomies across domains.**

In this work, we present a novel 3D anatomical-aware learning framework **ContrastMix**, to perform robust abdominal organ segmentation on NCCT scans (target domain) with CECT scans (source domain) in unsupervised setting. The backbone of the segmentation method is inspired by RAP-Net (125), which provides high flexibility to promote robust 3D abdominal multi-organ segmentation using single hierarchical network architecture. We further leverage pairwise CECT-NCCT samples to adapt the significant shifts in both contrast intensity and organ morphology. Specifically, intra-subject registration is applied to multi- minimize the morphological shift between multi-contrast scans. Meanwhile, organ-specific volumetric patches

are extracted from coarse attention maps computed from CECT-trained models as a spatial prior. We additionally leverage a source domain teacher network to generate refine probability mappings for soft supervision. For target domain, we perform knowledge distillation to sharpen the self-predicted non-contrast context and reduce the uncertainty in the boundary region during the training process. Furthermore, a beta-distributed weighting parameter is sampled randomly to mix both the contrast (image) and morphological (probability prior) variation from both domains to adapt the subtle boundary information with non-contrast characteristics. The experimental results demonstrated that ContrastMix outperforms the state-of-the-art (SOTA) supervised baselines with consistent improvements across internal and external testing cohorts. Our contributions are summarized with four-folds:

- We propose an unsupervised learning approach that leverage the pairwise contrast-enhanced domain context as pseudo supervision for refine non-contrast organ segmentation.
- We propose a knowledge distillation technique to sharpen the certainty of self-predicted non-contrast boundary and integrate the refined domain-invariant mapping from teacher network to preserve organ morphology.
- We propose a local mixing technique that efficiently adapts the diverse contrast appearance and enriches the boundary information of each organ across domains.
- We evaluate ContrastMix with two separate non-contrast clinical cohorts from various imaging scanners and protocols, and demonstrate consistent improvements comparing with current hierarchical and transformer-based state-of-the-art approaches.

3.3 Related Works

Medical Image Segmentation: Previous efforts have tackled organ segmentation in contrast-enhanced CT with significant usage of deep learning. A naive approach of aggregating target domain data is to directly train a deep network model with ground-truth label and has been validated with various regions of CT in the existing literature (40; 157). However, the supervised strategies require a large a number of high-quality ground truth labels and the inaccuracy of the boundary information still exists between neighboring anatomical structures. Meanwhile, volumetric images have to be downsampled to fit into deep neural networks for training due to the limited GPU memory (39; 182). To minimize the loss of anatomical context, patch-based approaches are proposed to adapt the high-resolution context for segmentation (99). Huo et al. proposed patch-based methods for whole-brain segmentation (98). However, the patch-based approaches only adapt the local representation and are limited in adapting the global spatial information in complete images. Therefore, hierarchical systems are proposed to adapt feature representation across scales. Roth et al. proposed

a coarse-to-fine method that roughly defines the local region to extract representation for refined segmentation (179). Roth et al. further extended the coarse-to-fine process into a multi-scale pyramid network to perform segmentation in high resolution (182). However, the performance is limited due to the inaccuracy prediction on the bounding box with the low-level models. In addition, upsampling is needed to resample the representation back to the original image resolution. Li et al. demonstrates a hybrid algorithm to integrate the 2D slice representation from the first stage with the 3D volumetric context in the second stage (139). Apart from adapting 2D information with 3D representation, Zhu et al. proposed an effective sliding window approach to extract the region of interest for refining segmentation (259). Additionally, an expanding bounding box is used to allocate the target regions accurately and minimize the outliers (out of target regions) for training models. To further integrate the global context for refining local segmentation, Tang et al. proposed a high-resolution approach to adapt the target-corresponding coarse segmentation as the additional channel with volumetric image patches for model training (200). However, Tang et al. and Zhu et al. targeted on the single organ segmentation in the refine stage and are limited to adapt the morphological variability across the organs in image patches. Lee et al. proposed RAP-Net to use the pseudo-binary organ prior as additional guidance and adapt all organ patches in a single network architecture (125).

Domain Adaptation for Segmentation: On top of deep supervised training, several studies performed domain adaptation techniques to find additional context for supervision from cross-modality data and compute segmentation on the target domain. With the investigation of generative adversarial approaches, several works proposed an image translation network (GAN-based) to align the contrast appearance from source to target domain and use the generated image for further processing. Huo et al. adapts the domain invariant feature across MRI to CT by using adversarial neural networks and further adds a segmentation module to increase the stability of the adaptation framework (97). Dou et al. investigates the extraction of domain-invariant features in feature space (55). In contrast, Chen et al. demonstrates a synergistic method to adapt domain-invariant features in both feature and image space (26). Chandrashekar et al. synthesizes images with contrast from the non-contrast domain and adjusts the contrast-characteristics for segmentation (25). However, the dataset from different domains may be unbalanced. It is challenging to translate the contrast characteristic of each organ interest into the target domain with stability. Apart from the generative approach, the multi-modal approach is used to leverage the modality-shared knowledge using feature fusion strategies. Valindia et al. (210) and Tulder et al. (211) demonstrate different parameter sharing strategies for unpaired multi-organ segmentation, while Dou et al. introduce the modality-specific normalization to adapt cross-modality unpaired context using knowledge distillation (54). Although promising progress is demonstrated, the adaptation framework is limited in 2D networks with a high demand for sufficient ground truth labels for volumetric image segmentation.

3.4 Method

A complete overview of ContrastMix is demonstrated in Figure 3.2. Our goal is to perform robust abdominal organs segmentation for non-contrast CT in an unsupervised setting. Let $S = \{(x_s, y_s)\}_{s=1}^{|S|}$ be the set of source domain contrast-enhanced cohorts and $T = \{(x_t, y_t)\}_{t=1}^{|T|}$ be the set of target domain non-contrast cohort, where each $x_s, x_t \in \mathscr{R}^(|\Phi|)$ is the corresponding domain paired image and $y_s, y_t \in \mathscr{R}^(|\Phi| \times |C|)$ is the corresponding one-hot paired ground truth segmentation mask. Here, we denote $\Phi \in \mathscr{R}^2 \times \mathscr{Z}$ as the dimension of images voxels and $C \in \mathscr{R}$ as the number channels for the segmentation classes. ContrastMix consists of three main hierarchical components: (1) intra-modal registration for generating anatomical context, (2) knowledge distillation of self-predicted non-contrast context and (3) cross-domain intensity-prior mixing with random weighting.

3.4.1 Intra-Modal Registration for Anatomical Prior Generation

As the organ anatomies are not well aligned in the initial state of both target and source domains, intra-modal registration is performed to minimize the anatomical shift between domains in an intra-subject setting. Here, we apply DEEDS (DEnsE Displacement Sampling) to perform hierarchical two-stage registration (84), which consists of 1) DEEDS affine registration and 2) DEEDS deformable registration (Fixed image: non-contrast domain, moving image: contrast-enhanced domain). The contrast correspondence between organs can be well demonstrated and extracted pairwise anatomical context as voxel-to-voxel. A transformer-based approach generates coarse segmentation with complete volumetric inputs from source domains (74). The volumetric images are downsampled to a specific resolution with tri-linear interpolation. Consider the coarse segmentation network F_c is parameterized with θ_c, the segmentation model aims to minimize as following:

$$\underset{\theta_c}{arg\,min}\, \underset{y}{} \mathscr{L}_V(\theta_c) \tag{3.1}$$

Here, we denote $\mathscr{L}_V$ as a combination of soft Dice loss and cross-entropy loss to train the low-resolution segmentation model with multiple semantic targets. The corresponding definition is as follows:

$$\mathscr{L}_V = 1 - \frac{2}{A}\sum_{a=1}^{A}\frac{\sum_{i=i}^{I} y_{i,j} h_{i,j}}{\sum_{i=1}^{I} y_{i,j}^2 + \sum_{i=1}^{I} h_{i,j}^2} - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{i,j} log(h_{i,j}) \tag{3.2}$$

where A denotes as the number of anatomical class; I is the number of voxels and $h_{i,j}$ is the probability output with softmax activation for class A at voxel i. The one-hot localization $a_s,\ a_t \in 0,\ 1^{|\Phi| \times |C|}$ provides spatial prior to extract organ-wise patches for refined segmentation. We then leverage the aligned spatial prior to randomly select voxels within each organ's one-hot bounded regions. Multiple bounding boxes are

placed as local views to extract volumetric patches with the voxels chosen as a center point. The extracted image patches are concatenated with the prior context as the additional channel guidance and perform organs segmentation refinement end-to-end in a single network architecture (125).

3.4.2 Self-Predicted Knowledge Distillation

Due to the non-contrast characteristic, it is challenging to identify subtle boundaries and may lead to over-segmentation across neighboring organs. To tackle such limitations, we propose a simple knowledge distillation module to enhance the probabilistic certainty in-between organ boundaries. We initially performed K augmentation for both images and coarse prior $\hat{x}_{t,k},\ \hat{x}_{s,k}, \hat{a}_{t,k},\ \hat{a}_{s,k} = Augment\,(x_t,\ x_s, a_s, a_t)\,,\ k \in (1,\ldots,K)$. Both augmented image and coarse prior are then concatenated as multi-channel input $m_{t,k},\ m_{s,k}$ to preserve the representation learned within the localized regions. For contrast-enhanced patches, a teacher model q_s is used to compute the probability map and provide sufficient boundary context to refine target segmentation. For non-contrast patches, the student segmentation network F is trained from scratch and generates self-predicted probability maps q_t as self-supervision. We denote N as the number of the data augmentations and compute the average prediction $\widetilde{q}_t$ across all self-predicted probability maps within a minibatch via softmax activation if $N > 1$:

$$\widetilde{q}_t = \frac{1}{K}\sum_{k=1}^{K} F(m_{t,k};\theta) \tag{3.3}$$

As the average prediction may extract inaccurate boundary information from the self-predicted context, we introduce a knowledge distillation module to increase the sensitivity towards the target organs' boundary. The knowledge distillation function is defined as follows:

$$S(\widetilde{q},T)_i = \frac{\widetilde{q}_i^{\frac{1}{T}}}{\sum_{j=1}^{C}\widetilde{q}_j^{\frac{1}{T}}} \tag{3.4}$$

where $\widetilde{q}_t$ is the average source domain probability map with C classes prediction over K augmentations with knowledge distillation, and T is a temperature scalar to sharpen the soft prediction labels and amplify the inter-class relationships in the knowledge distillation module.

3.4.3 Cross-Domain Intensity-Prior Mixing

Apart from the limitation of subtle boundary, the contrast variation between organ interests poses a great challenge to the segmentation generalizability across domains. Here, we propose an intensity-prior mixing module to random weight the contrast intensity and domain-invariant context (anatomical prior) of each domain, and sum it to generate an augmented version as the generalized representation. Specifically, we first concatenate both domains input and generate a shuffle version of the concatenated input $(\widetilde{m}_1, \widetilde{m}_2)$. Both

concatenated images and pseudo probability mapping are multiplied with a random weighting h, which is sampled from a beta distribution $h \sim Beta(\alpha, \beta)$ with hyperparameters α and β. The mix-and-match strategy provides another form of data augmentation for the segmentation model F to behave linearly in-between the cross-domain samples. Let $p = F(x;\theta) \in 0,\ 1^{|\Phi| \times |C|}$ be the class probability map prediction from the segmentation network. Both the contextual features extracted from both source and target domains are integrated as $(x',\ p')$ by:

$$\begin{aligned} x' &= h \cdot (m_s, m_t) + (1-h) \cdot (\widetilde{m}_1, \widetilde{m}_2) \\ p' &= h \cdot (p_s, p_t) + (1-h) \cdot (\widetilde{p}_{m_1}, \widetilde{p}_{m_2}) \end{aligned} \tag{3.5}$$

The context shuffling strategies increase the confidence level of predicting uncertain regions between the neighboring organs with the integration of source domain representation. Additionally, we use p' as soft guidance to preserve the organ anatomies and benefit to identify the subtle boundary with non-contrast characteristics. The weighted samples are then input into the segmentation network and computed to refine binary segmentation as the final output.

3.4.4 Loss Functions

Assume that the non-contrast segmentation network F is parameterized by the set of weights θ, both target and source domain patches are exploited jointly in the training process by minimizing the following loss function:

$$\mathscr{L}_{total} = \mathscr{L}_{T_s}(\theta;S) + \mathscr{L}_{T_t}(\theta;T) + \lambda \mathscr{L}_{unsup}(\theta;T) \tag{3.6}$$

To preserve the similarity of the organ anatomy between pairwise images, teacher knowledge supervision loss $\mathscr{L}_{T_s}(\cdot)$ and $\mathscr{L}_{T_t}(\cdot)$ are computed to constrain the segmentation network and generate predictions with a similar anatomical structure to the pseudo prior generated with random weighting p'. Besides optimizing the non-contrast target outputs, constraining the contrast-enhanced prediction can help adapt the shape invariant representation in the source domain imaging. The pseudo prior supervised loss is defined as follows:

$$\begin{aligned} \mathscr{L}_{T_s}(\theta;S) &= \sum_{x_s \in S} \ell_{T_s}(F(x_s;\theta),\ p') \\ \mathscr{L}_{T_t}(\theta;T) &= \sum_{x_t \in T} \ell_{T_t}(F(x_t;\theta),\ p') \end{aligned} \tag{3.7}$$

Both x_s and x_t are registered aligned in the preprocessing step and well adapt to the topological correspondence across organs in the abdominal region. Dice loss is used for the segmentation loss functions $\ell_{T_s}(\cdot)$ and

$\ell_{T_t}(\cdot)$. The dice loss function is defined as follows:

$$\begin{aligned} \ell_{T_t}\left(p_t, p'\right) &= \sum_{p_t \in T}\left(1 - 2 \cdot \frac{p_t \cdot p'}{p_t + p'}\right) \\ \ell_{T_s}\left(p_s, p'\right) &= \sum_{p_s \in S}\left(1 - 2 \cdot \frac{p_s \cdot p'}{p_s + p'}\right) \end{aligned} \tag{3.8}$$

Apart from the pseudo prior supervised loss, a soft unsupervised loss function is employed to further leverage the self-predicted context and minimize the probability of over-segmentation. We introduce $\mathscr{L}_{unsup}$ to combine the cross-entropy loss between p_s and the teacher prediction M_s, and the square L_2 loss on the final target domain prediction and the self-supervised intermediate prediction M_t. The integration with square L_2 loss reduces the sensitivity of incorrect prediction within the anatomical context bounded region. The unsupervised loss is defined as follows:

$$\mathscr{L}_{unsup} = -\sum_{i \in \Phi}\sum_{j \in C} M_{sij} \log\left(p_{sij}\right) + \lambda_t \sum_{p_t \in T} ||p_t - M_t||_2^2 \tag{3.9}$$

where λ_t is the weighting hyperparameter of the unsupervised loss computed with the self-predicted outputs. Overall, the training objectives including $\mathscr{L}_{T_s}$, $\mathscr{L}_{T_t}$ and $\mathscr{L}_{unsup}$ are optimized end-to-end concerning weighting parameters θ. In the testing phase, only non-contrast target domain scans are input for inference and obtain refined segmentation with a majority vote for joint label fusion.

3.5 Experiments

To evaluate the performance of the proposed multi-contrast domain adaptation pipeline, we perform internal and external validations on two organ segmentation tasks with non-contrast phase CT samples. In the following subsection, we describe more details about the datasets that we implement and our experimental setup. More details of preprocessing, training and evaluation metric are demonstrated in the supplementary material (U54DK120058).

3.5.1 Datasets

Contrast-Enhanced & Non-Contrast Pairwise Clinical Research Cohort (CENC):: We retrieved 56 de-identified splenomegaly subjects with pairwise portal venous phase CT and non-contrast phase CT for internal training and validation. Each contrast-enhanced and non-contrast volume is annotated and refined manually with 12 classes of multiple abdominal organs. The ground truth labels of 12 multiple organs are provided including 1) spleen, 2) right kidney, 3) left kidney, 4) gall bladder, 5) esophagus, 6) liver, 7) stomach, 8) aorta, 9) inferior vena cava (IVC), 10) portal splenic vein (PSV), 11) pancreas, and 12) right adrenal gland. The

Table 3.1: Comparison of the fully-supervised, unsupervised, semi-supervised and partially supervised state-of-the-art methods on the 2015 MICCAI BTCV challenge leaderboard. (We show 8 main organs Dice scores due to limited space, $*$: fully-supervised approach, $\star$: semi-supervised approach, $\triangle$: partially supervised approach.)

Method	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Aorta	IVC	Average Dice
3D U-Net (2016)	0.937	0.856	0.912	0.690	0.631	0.920	0.880	0.769	0.762
Roth et al. (2017)	0.940	0.870	0.923	0.701	0.674	0.925	0.891	0.772	0.770
Syn-Seg Net (2018)	0.941	0.868	0.910	0.654	0.652	0.927	0.895	0.780	0.764
Zhu et al. (2019)	0.950	0.880	0.918	0.710	0.643	0.932	0.890	0.802	0.781
Dou et al. (2020)	0.957	0.878	0.938	0.708	0.649	0.930	0.900	0.784	0.788
RAP-Net (2021)	0.954	0.874	0.928	0.701	0.653	0.928	0.897	0.790	0.784
UNETR (2021)	0.960	0.896	0.936	0.800	0.736	0.949	0.888	0.789	0.798
ContrastMix (Ours)	**0.971**	**0.926**	**0.950**	**0.820**	**0.736**	**0.960**	**0.915**	**0.821**	**0.820***

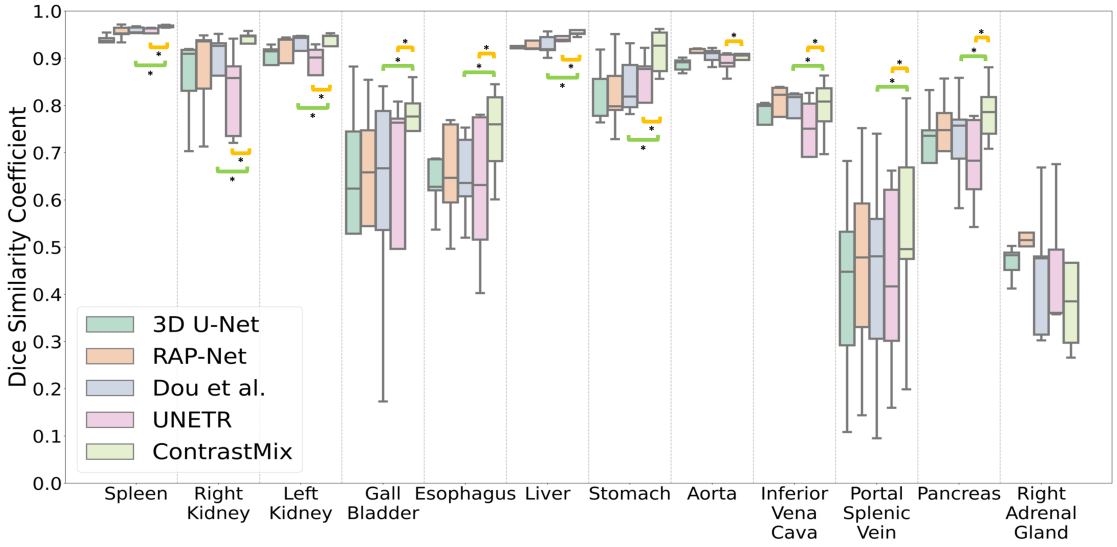


Figure 3.3: ContrastMix outperforms the current SOTA approaches on multi-organ segmentation with CENC dataset. (*:$p < 0.01$, with Wilcoxon signed-rank test)

axial-plane pixel dimension of each scan varies from 0.64 to 0.98 mm and the corresponding slice thickness (z-axis) is constantly 3mm. Each CT volume consist 75 -116 slices with $512 \times 512 pixels$.

Non-Contrast Healthy Clinical Research Cohort (NCH): We retrieved 29 de-identified subjects free from splenomegaly with non-contrast phase CT for external validation of non-contrast segmentation performance. Manual annotation of aorta volumes (the ground truth) was performed by expert image analysts under the supervision of a clinical radiologist (MD) from Vanderbilt University Medical Center. These scans have a large variance of the morphology for the aorta, with volumes varying from 39.3 cubic centimeters (cc) to 96.9 cc. Each CT volume consists of 70 slices of 512×512 pixels and has a constant resolution of $0.68mm \times 0.68mm \times 3.00mm$ for axial plane and slice thickness across all subjects.

3.5.2 Experimental Setup

We conducted experiments on two perspectives of analysis to evaluate the effectiveness of the cross-domain adaptation pipeline on non-contrast organs segmentation. We performed multi-organ segmentation on the non-contrast samples in CENC dataset as internal training and validation. The network is trained from scratch with five-fold cross-validations using CENC dataset: training subjects: n=44; validation subjects: n=6; testing subjects: n=6 (10% ratio of the samples are used for testing).

Moreover, we further evaluate the impact of the distillation module and the cross-domain mixing module on the segmentation performance. We performed ablation studies to analyze the correspondence of the hyperparameters T and (α, β) towards the robustness of the segmentation pipeline. Apart from the internal validation, we performed external testing with single organ aorta segmentation within NCH samples using the well-trained model to demonstrate the confidence level on the generalization ability to adapt non-contrast segmentation in a semi-supervised setting.

We compared our method against several volumetric supervised state-of-the-arts (SOTA). The fully-supervision SOTA considers all training samples with ground-truth labeled. Our pipeline consists of training the coarse volumetric segmentation network with source domain ground-truth labels; we consider our proposed method as semi-supervised setting without using non-contrast ground-truth label usage. Moreover, we compare our pipeline with fully supervised and multi-modal state-of-the-arts methods: 3D U-Net segmentation approach (39), Syn-Seg Net (97), multi-scale segmentation (181), pyramid scaling segmentation (259), RAP-Net (125), segmentation with knowledge distillation (54), and transformer network for segmentation (74). We constrain all testing inference with the same underlying network architecture, optimization procedure, and data augmentations for a fair comparison. We finally compared the average Dice score and average mean surface distance across organs segmentation and computed the statistical significance with Wilcoxon signed-rank test.

3.6 Results

3.6.1 Internal Testing Performance

As shown in Figure 3.3 and Table 3.1, the performance increases consistently from a single U-Net architecture network to a coarse-to-fine approach adapting the non-contrast scans. With the knowledge distillation guidance (54), performance in particular organs such as the spleen and left kidney, improves by reducing the possibility of over-segmentation across organ boundaries. However, Figure 3.3 illustrates that Dou et al. predicts the organ boundaries with less confidence and limits to preserve the core anatomical context of the organs only. UNETR outperforms all the supervised state-of-the-arts with an average Dice score of 0.798 across 12 organs. It further increases the encoder's power to abstract the non-contrast representation context

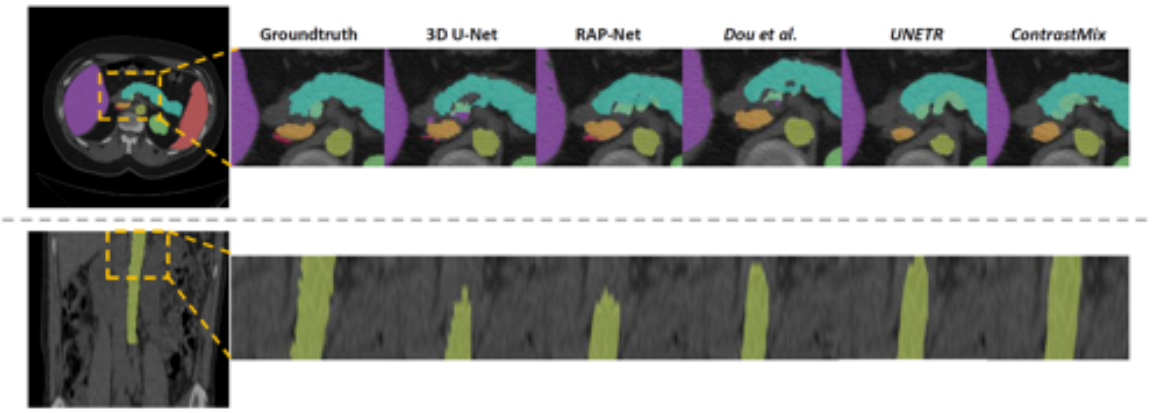


Figure 3.4: Qualitative Representations with different state-of-the-art strategies yields incremental improvement in segmentation performance. ContrastMix results in smooth boundaries and accurate morphological information for each organ in both internal and external inferences.

across organs. With the use of ContrastMix, the performance of the non-contrast segmentation significantly boosts from the average Dice score of 0.798 to 0.820 across all organs. The performance on all organs segmentation outperforms significantly to all state-of-the-art methods with statistical significance. The qualitative representation in Figure 3.4 further shows that ContrastMix refines organ details and preserves the boundary information between neighboring organs.

3.6.2 External Testing Performance

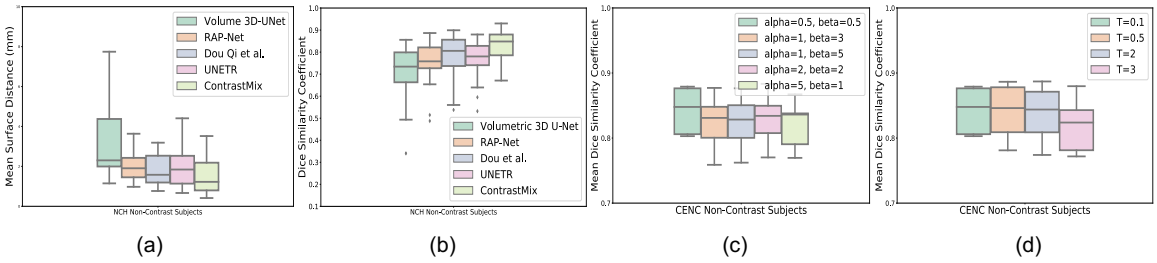
Figure 3.5(a) & 3.5(b) and Table 3.2 present the quantitative performance of single organ aorta segmentation with NCH. The trending of the segmentation performance is similar to that of multi-organ segmentation with CENC. Interestingly, the knowledge distillation pipeline significant improves the performance of the state-of-the-art hierarchical RAP-Net and achieves the minimal MSD across all state-of-the-arts. In addition to the transformer network, UNETR shows improvement to a small extent in Dice score and limits to establish an additional advantage in MSD. By using ContrastMix, the segmentation performance outperforms all state-of-the-art approaches with 8.00% leverage in Dice score and 33.6% decrease in MSD. In the qualitative perspective in Figure 3.4, the segmentation mask of the aorta organ is incrementally refined and ContrastMix best preserves the anatomical and boundary details across all the state-of-the-art approaches.

3.6.3 Ablation Studies

We further evaluate the contribution of each key module adapting non-contrast segmentation in our model architecture with the internal testing cohort. We optimize each module's performance by adapting the variation of hyperparameters 1) temperature scaling T and 2) the random weighting distribution (α, β).

Table 3.2: SOTA approaches comparison for aorta segmentation on external testing dataset NCH (*:$p < 0.01$, with Wilcoxon signed-rank test)

Methods	Average Dice	Average MSD
3D U-Net (2016)	0.712 ± 0.111	4.14 ± 3.94
Roth et al. (2017)	0.735 ± 0.100	3.43 ± 3.45
Syn-Seg Net (2018)	0.721 ± 0.124	3.89 ± 2.57
Zhu et al. (2017)	0.746 ± 0.0978	3.01 ± 2.21
Dou et al. (2017)	0.770 ± 0.114	2.34 ± 1.88
RAP-Net (2020)	0.755 ± 0.0944	2.67 ± 1.74
UNETR (2021)	0.775 ± 0.0820	2.41 ± 1.84
ContrastMix (Ours)	$\mathbf{0.837 \pm 0.0662}$	$\mathbf{1.60 \pm 1.08}$



(a) (b) (c) (d)

Figure 3.5: ContrastMix outperforms the current SOTA approaches and demonstrates significant improvement in both a) Dice score and b) mean surface distance in external datatsets. We further perform ablations studies on the contribution of adaptation mdoules with hyperparmeters c) α & β for beta-distribution and d) temperature.

Variation of hyperparameters T: From Figure 3.5(c), interestingly, constant robustness in performance is demonstrated across the variability of T and the median Dice with $T = 0.1$ is slightly higher than the others. However, the segmentation performance did not substantially vary and may be due to the limited data augmentation performed for non-contrast patches in each batch ($A = 1$) with limited GPU memory.

Variation of weighting distribution (α, β): We further investigate the influence of the hyperparameter $h\prime$ and adapt random intensity augmentation with different weighting on the contrasting context in the training process. Figure 3.5(d) illustrates the effect on the segmentation performance of extracting $h\prime$ with multiple shape parameters combination of a beta distribution. The segmentation performance with beta-distribution of shape parameters $\alpha = 0.5$ and $\beta = 0.5$, demonstrates significantly better generalizability across the dataset compared with the extracted $h\prime$ from other beta-distribution. The model with such beta distribution adopts the contrasting domain knowledge by heavily randomizing the weighting to either one of the domains. The increase of sensitivity is demonstrated by encoding the variability across the domain shift and is beneficial in

Table 3.3: Quantitative measures on ablation studies of multi-organ segmentation performance.

ContrastMix	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Aorta	IVC	Average Dice
T=0.1	**0.971**	**0.926**	**0.950**	**0.820**	0.736	**0.960**	**0.915**	**0.821**	**0.820**
T=0.5	0.968	0.925	0.944	0.805	0.713	0.958	0.907	0.797	0.814
T=2.0	0.969	0.921	0.945	0.790	0.716	0.957	0.909	0.795	0.812
T=3.0	0.968	0.924	0.944	0.797	**0.741**	0.957	0.903	0.794	0.810
α=0.5, β=0.5	**0.971**	0.926	**0.950**	**0.820**	**0.736**	**0.960**	**0.915**	**0.821**	**0.820**
α=1.0, β=3.0	0.970	0.926	0.943	0.790	0.684	0.957	0.905	0.799	0.803
α=1.0, β=5.0	0.970	**0.928**	0.946	0.776	0.710	0.959	0.907	0.800	0.810
α=2.0, β=2.0	0.970	0.926	0.947	0.784	0.704	0.959	0.907	0.800	0.805
α=5.0, β=1.0	0.969	0.926	0.944	0.693	0.695	0.958	0.907	0.792	0.803

adapting the structural information of the abdominal organs.

3.7 Disucssions

In this work, we present a novel 3D anatomical-aware semi-supervised learning scheme ContrastMix to adapt non-contrast imaging for robust abdominal organs segmentation. One of the challenges in non-contrast segmentation is distinguishing the subtle boundary between organ interests. Unlike traditional generative adversarial approaches, we use pairwise-registered contrast-enhanced imaging to generate refined boundary context and provide morphological constraints as additional supervision. First, we provide a 3D coarse-to-fine pipeline, which refines the morphological context in organ-aware regions. Next, we sharpen the probabilistic context of the non-contrast boundary through the self-predicted prior and additionally provide teacher context from contrast-enhanced domain as supervision. Finally, we randomly weight the organ-wise contrast correlation between domains and adapt the generalized contrast context for robust segmentation. Moreover, a large scale of experiments is performed, including the comparison between current learning state-of-the-arts and ablation studies on the proposed innovations. We demonstrate that ContrastMix outperforms all fully supervised approaches for multi-organ segmentation. Additionally, we deploy our trained model on another unseen dataset for single organ segmentation and demonstrate the generalizability across different cohorts. The ablation studies provide a better understanding of the impact on distilling self-predicted context and the random intensity mixing strategy for non-contrast segmentation.

Although ContrastMix tackles the current challenges of non-contrast segmentation, limitations still exist in the process of ContrastMix. One limitation is the dependency of the teacher prediction quality in the source domain. As we leverage the morphological context in contrast-enhanced domain as supervision, low-quality predictions of organ-aware regions may also be possible to compute and use as guidance. Inaccurate prior information may thus be introduced into the training process. Another limitation is performance in an

organ-centric setting with well aligned anatomical references. We aim to innovate a single stage pipeline with end-to-end optimization as our future work.

3.8 Conclusion

In summary, the proposed ContrastMix network achieved consistent performance on organ segmentation with non-contrast scans, compared with the current state-of-the-art approaches. The core innovations of ContrastMix are to 1) sharpen the certainty of the non-contrast boundary context with knowledge distillation and 2) adapt the contrast and morphology variations by generating samples with randomly weighted contrast and prior knowledge across domains. As our proposed pipeline focuses on adapting the contrast variation across domains, a potential extension can be focused on adapting significant domain shift of other imaging modalities for organ segmentation in unsupervised setting.

CHAPTER 4

Semi-Supervised Multi-Organ Segmentation through Quality Assurance Supervision

4.1 Overview

[1]Human in-the-loop quality assurance (QA) is typically performed after medical image segmentation to ensure that the systems are performing as intended, as well as identifying and excluding outliers. By performing QA on large-scale, previously unlabeled testing data, categorical QA scores (e.g. "successful" versus "unsuccessful") can be generated. Unfortunately, the precious use of resources for human in-the-loop QA scores are not typically reused in medical image machine learning, especially to train a deep neural network for image segmentation. Herein, we perform a pilot study to investigate if the QA labels can be used as supplementary supervision to augment the training process in a semi-supervised fashion. In this paper, we propose a semi-supervised multi-organ segmentation deep neural network consisting of a traditional segmentation model generator and a QA involved discriminator. An existing 3-D abdominal segmentation network is employed, while the pre-trained ResNet-18 network is used as discriminator. A large-scale dataset of 2027 volumes are used to train the generator, whose 2-D montage images and segmentation mask with QA scores are used to train the discriminator. To generate the QA scores, the 2-D montage images were reviewed manually and coded 0 (success), 1 (errors consistent with published performance), and 2 (gross failure). Then, the ResNet-18 network was trained with 1623 montage images in equal distribution of all three code labels and achieved an accuracy 94% for classification predictions with 404 montage images withheld for the test cohort. To assess the performance of using the QA supervision, the discriminator was used as a loss function in a multi-organ segmentation pipeline. The inclusion of QA-loss function boosted performance on the unlabeled test dataset from 714 patients to 951 patients over the baseline model. Additionally, the number of failures decreased from 606 (29.90%) to 402 (19.83%). The contributions of the proposed method are threefold: We show that (1) the QA scores can be used as a loss function to perform semi-supervised learning for unlabeled data, (2) the well trained discriminator is learnt by QA score rather than traditional "true/false", and (3) the performance of multi-organ segmentation on unlabeled datasets can be fine-tuned with more robust and higher accuracy than the original baseline method. The use of QA-inspired loss functions represents a promising area of future research and may permit tighter integration of supervised and semi-supervised learning.

[1]Published at: Lee, Ho Hin, et al. "Semi-Supervised Multi-Organ Segmentation through Quality Assurance Supervision.", Medical Imaging 2020: Image Processing. Vol. 11313. SPIE, 2020. (130)

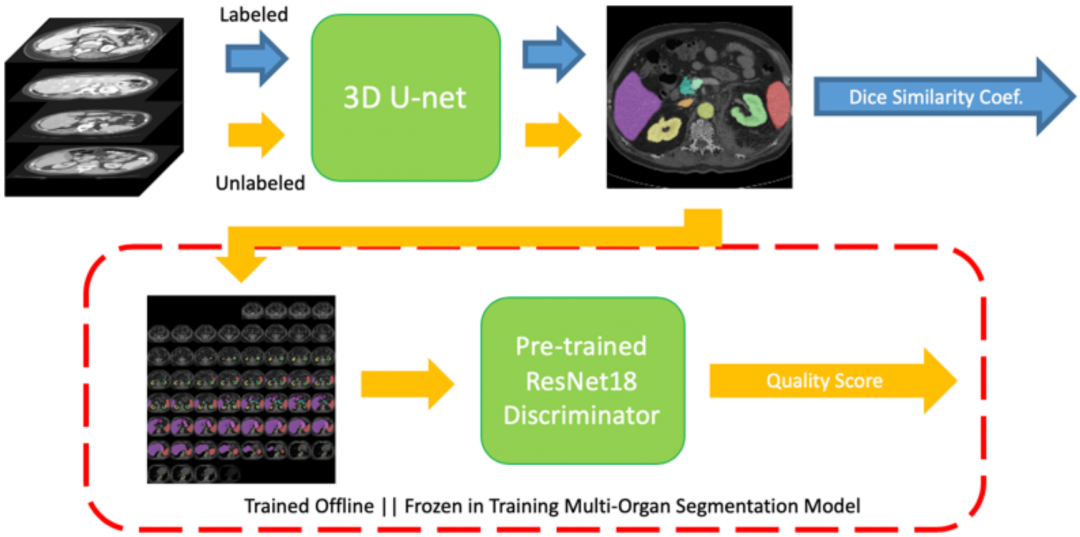


Figure 4.1: The full pipeline of semi-supervised learning-based segmentation is divided into two main processes: multi-organ segmentation and a quality discriminator module. All labeled and unlabeled data entered a 3D U-Net to predict organ segmentation masks. Since Dice loss cannot be calculated from the unlabeled datasets due to the lac of ground truth, 2D slice alignment montage images are formed instead and predict segmentation quality score as the loss function. The quality scores backpropagated to the 3D U-Net model to fine-tune the performance of segmentation on unlabeled datasets.

4.1.1 Introduction

The quality of medical image processing (e.g., segmentation) is affected by imaging quality, which can be influenced by both hardware-related and human-related artifacts (16). The effectiveness and accuracy of image processing algorithms play an essential role in the usefulness and quality of medical image processing outcomes. To assess the image processing performance on a previously unseen dataset, quality assurance (QA) is typically performed to ensure the accuracy of the results. QA is a rich area of study and broadly consists of two main directions: subjective assessment, which is judged by human interaction, and objective assessment, which is decided by mathematical algorithms (115). In medical image segmentation, human involved QA is still the de facto standard process to decide if the segmentation results are acceptable for the intended purposes. For instance, CT images with segmentation masks can be labeled with scores 0, 1 and 2 in a subjective manner and provide general information about the overlay image's quality. This quality score can be used to provide supplementary information in diagnosing abdominal organ disease via automatic algorithms such as deep learning.

With the onset of deep learning, large datasets are needed for training deep learning models; extensive engineering efforts are put into labelling and excluded outliers from large populations of medical images

manually, especially CT images with low contrast (222). Manually annotating organs with human expertise is preferred to ensure that the organs are labeled in the correct locations before used for training with supervised and semi-supervised learning methods (222). However, manually performing annotations on medical images is time-consuming. Also, predictive models are dependent on the quality of the labels, and therefore suffer from outliers with low image quality. Image quality has become an important aspect to focus on and is shown to have a great impact on deep learning model prediction. In order to enhance the accuracy and efficiency of medical images analysis, previous research has been done for determining the image quality necessary to extract valuable information in medical perspectives.

Previously, assessing image quality of cardiac magnetic resonance (MR) images was proposed to detect the missing slice of the 3D MR images in cardiac scans and extract meaningful biomarkers from the missing slice with the use of convolutional neural networks (242). The training dataset was extended with mis-triggering artifacts, using different levels of corruption in image quality to enhance the effect of image artifacts in training for data augmentation. On the other hand, an automated deep learning system has been created which uses retinal image quality as a feature to provide accurate diagnosis of diabetic retinography (240). Also, uncertain chest x-ray image quality was leveraged to perform disease classification and lesion detection with deep Bayesian neural networks (236). Producing uncertain predictions on medical cases leads to significant clinical value, while some of the cases need to be evaluated with physical exam or surgery to confirm (240; 236). For medical image segmentation, an image-specific fine-tuning algorithm was proposed to make the convolutional neural network model adaptive to specific testing images and increase the generalizability of previously unseen data with the use of image quality (215). New directions offer the potential to assess image quality to enhance the efficiency and accuracy of evaluating medical images. In recent years, with the use of deep learning, networks have been proposed to extract deep features and perform diagnosis in a robust manner, such as 3D U-net (39) and ResNet (79). These networks have shown consistent interest and a great variety of usage in medical image segmentation (66; 57).

In this paper, we propose a semi-supervised learning method with a human-guided discriminator to determine the generalized quality of CT image segmentation and fine-tune the segmentation accuracy with the score prediction from the discriminator. The original 80 labeled 3D abdominal CT images along with the 2027 QA-only datasets were used for training the multi-organ segmentation model, along with the prediction of segmentation labels in each training epoch overlaid with the corresponding input image. Each overlaid 3D volume in the training phase was converted into a montage image and used as input into the discriminator to predict the segmentation quality at each training epoch. The prediction score from the discriminator was included in the loss and backpropagated to increase the model performance from QA- only data. 80 labeled datasets and 2027 unlabeled datasets are used in training the multi-organ segmentation model. The 3D Dice

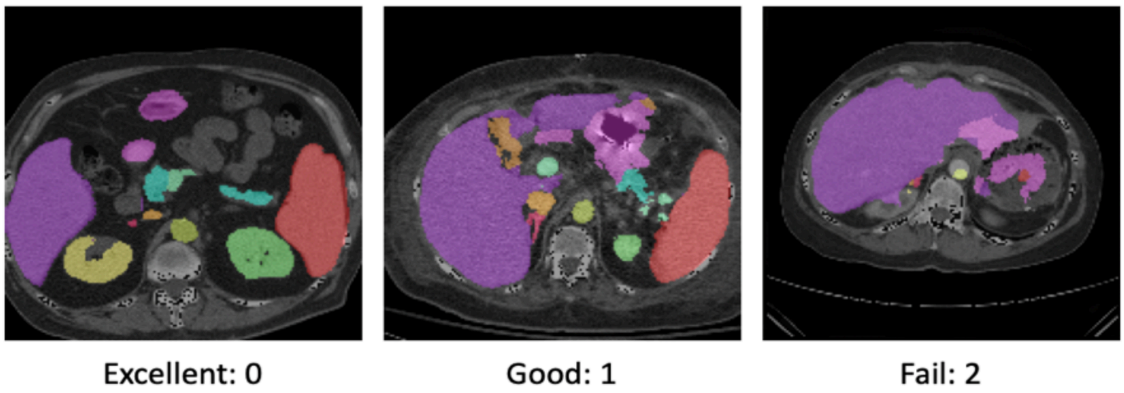


Figure 4.2: Respective images show the segmentation quality coded with 0 (success), 1 (errors consistent with published performance), and 2 (gross failure).

loss function is used for determining the difference between the predicted segmentation and the labels.

4.2 Methods

In this paper, we propose a semi-supervised learning method for multi-organ segmentation with a 3D U-net and create a segmentation quality discriminator with ResNet-18 (79), which is presented in Figure 4.1.

4.2.1 Preprocessing

3D CT abdomen volumes are used as the raw datasets, and the volume part near the middle kidney was chosen via body part regression algorithms. Each value from the body part regression results represented the approximate location of the slice in the whole volume of 3D images. Slices with value near +12 are located around the duodenum, while slices with value -12 are near the heart. The slices with value -6 to +6 are extracted as a 3D volume and normalized to a resolution of $2 \times 2 \times 6$ mm with dimensions of 168x168x64.

4.2.2 Network

Two independent networks are presented in Figure 4.1 that form the combined solution. For the multi-organ segmentation, a 3D U-net is used to extract deep 3D features with the encoder in U-net. Another network, ResNet18, is used as the discriminator to predict the segmentation quality score from score-labeled montage images.

4.2.3 Segmentation Quality Discriminator

For unlabeled datasets, no ground truth masks were given for multi-channels to calculate the Dice similarity coefficients. After the segmentation mask prediction from 3D U-net, the segmentation mask was overlaid

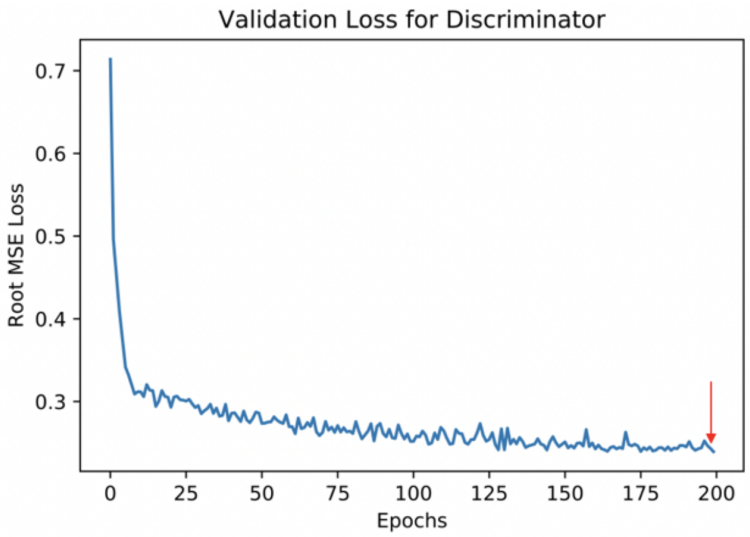


Figure 4.3: Validation loss as a function of traiing epoch and epochs 198 model is chosen with the least MSE error on the validation cohort.

with its original input image and decomposed into 2D slices; these slices were aligned into a single montage image with the dimension 1344×1344. The discriminator was trained to determine the general quality of the segmentation predictions with the input of montage images. The input montages were downsampled to 256×256 for training to reduce the memory used. The discriminator obtained a high accuracy of 94% on 407 labeled test images with a ResNet18 trained with 1620 score-labeled montage images. A quality score was predicted and coded with 0 (success), 1 (errors consistent with published performance), and 2 (gross failure). After the model was trained, the discriminator model was frozen and put into the multi-organ segmentation pipeline to predict quality scores for unlabeled datasets. The score acted as the loss function for the multi-organ segmentation and was backpropagated from the frozen pre-trained network to fine-tune the performance on the datasets without labels.

4.2.4 Loss Functions for Semi-Supervised Multi-Organ Segmentation

In the multi-organ segmentation pipeline, all labeled images and unlabeled images are input into the 3D U-Net to generate color segmentation prediction masks for different organs. For labeled images, 3D Dice loss was calculated for 12 organ channels, and all 12 anatomies were pushed through average Dice loss in the backpropagation and optimization process. We calculated the Dice loss with batch size 1 and all inliers and outliers were included to perform segmentation. The loss functions were:

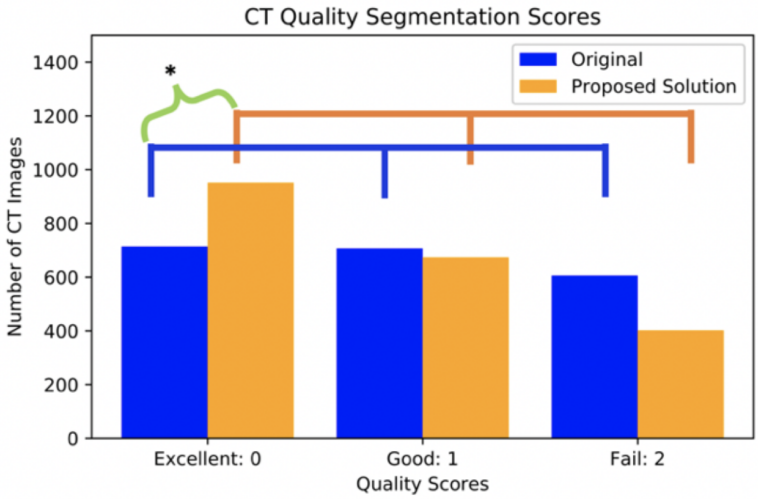


Figure 4.4: Significant increase in segmentation quality for unlabeled datasets is shown and reduce failure rate. (*significant at p<0.0001)

Multi-Sourced Dice Loss (MSDL):

$$MSDL = -\frac{2}{A}\frac{\sum_{a=0}^{A} w \sum_{i=1}^{M}\sum_{j=1}^{N} S_{ij}V_{ij} + \phi}{\sum_{a=0}^{A} w \sum_{i=1}^{M}\sum_{j=1}^{N} S_{ij} + \sum_{a=0}^{A} w \sum_{i=1}^{M}\sum_{j=1}^{N} V_{ij} + \phi}, \tag{4.1}$$

where A is the number of anatomies, w represents the variance between labels set properties and P is the segmentation mapping for various organs. $\in$ ensures the stability of the loss function. Hence, $\in$ was used in computing the prediction and voxel value correlation. In the segmentation, 12 anatomies are adopted and the Dice loss function was iteratively optimized using Adam optimization.

Mean Square Error Loss (MSEL):

$$MSEL = \frac{\sum_{i=1}^{N}(y_i - y_i^P)^2}{n} \tag{4.2}$$

For unlabeled data, the Dice loss function cannot be used due to the lack of ground truth segmentation masks. Therefore, the human-guided discriminator is used as a loss function and predicts a score for image segmentation quality. Montage images are created by slicing the 3D volume and aligning all 2D slices into one single color image with the segmentation overlaid. MSE loss is used to calculate the difference between the human-labeled score and the predicted score.

4.3 Data and Experiments

4.3.1 Data and Platform

2107 total (both labeled and unlabeled) 3D abdomen CT images were used in the multi-organ segmentation pipeline. The data was retrieved in de-identified form from ImageVU under IRB approval. The volume of all datasets has dimensions of 168x168x64. 80 images from the datasets were manually labeled with all 12 anatomies. The remaining 2027 3D images were sliced and each of them was aligned as one single 2D montage image, which was used for training and validation in the discriminator module. Before inputting into the discriminator module, the montages are downsampled to 256×256 and overlaid with the mask prediction at each epoch. All 2027 montages were shuffled, and 80% of the datasets were used for training the discriminator, while another 20% of the montages were used for validation to assure the accuracy of the ResNet18 in comparing with score labels manually reviewed. For multi-organ segmentation, 80% of the original 3D volumes were randomly picked for training, and the remaining 20% were used for testing the performance of the segmentation model on labeled and unlabeled datasets.

4.3.2 Experiment Design

4.3.2.1 Multi-Organ Segmentation

The multi-organ segmentation is performed as a baseline by changing the resolution from 1x1x3 mm to 2x2x6 mm and show the original state of art result for the 80 labeled and the 2027 unlabeled 3D datasets. 12 anatomies are segmented including spleen, kidney, gallbladder, esophagus, liver, stomach and pancreas (96). After performing the segmentations, their montages are manually reviewed by experts, who determine the segmentation quality score (0,1,2) for unlabeled datasets. For the labeled datasets, Dice similarity coefficient were calculated to assess the performance of the segmentation.

4.3.2.2 3D U-Net

For the multi-organ segmentation, a 3D U-net is used for extracting the deep features and increase the model capacity to reconstruct the image for mask prediction (39). The learning rate was set at 0.0001 and the network parameters are presented with the original paper published (39). Multi or sourced Dice loss (MSDL) are used because of the multiple anatomies and segmentation mapping are the final output prediction from the network (202).

4.3.2.3 Discriminator Module

After performing prediction of the segmentation mask, organ segmentation masks were overlaid with the original 3D volume and sliced into 2D. Each slice was aligned in a single montage image and down-sampled

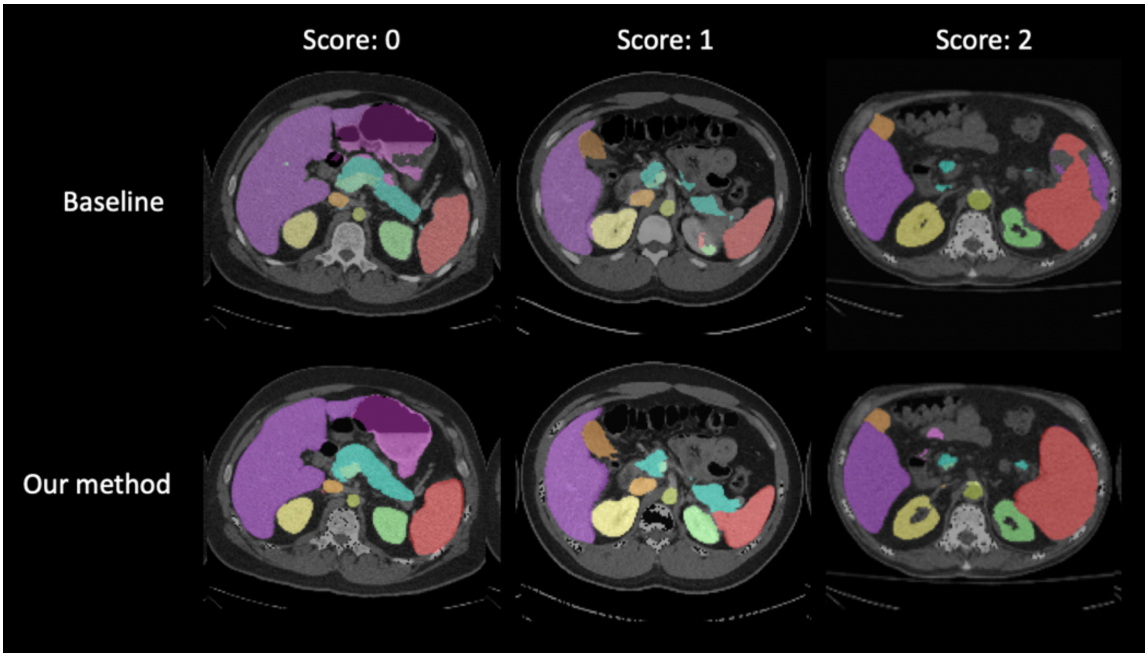


Figure 4.5: The segmentation mask overlay improved over the baseline method with our proposed method.

to 256x256, in order to input into the discriminator for determining the segmentation quality. The quality score was backpropagated through the ResNet, and montage processing back to 3D U-net model to fine-tune the model performance. The discriminator module was separately trained with 80% of the 2027 score labeled montage images. The remaining 20% of 2027 volumes were randomly separated into equal portions for testing and validation. A ResNet with 18 layers is used and performed regression to output a score for the segmentation quality. The learning rate was 0.0001 and the network parameters are presented with the original paper published. MSE Loss was used to see the difference between the score label and the prediction score. For the model, validation was performed and epoch 198 model had the least root mean square error value.

4.3.2.4 Visual Quality Assessment

The quality score for each montage image was determined for labeling under several conditions: (1) The accuracy of the segmentation of liver, (2) kidney and (3) spleen. If the segmentation of liver, kidney and spleen were well located, the montage image were coded with score 0. Score 1 was coded for inaccurate segmentation of liver, spleen and kidney, as long as the segmentation mask could roughly define the location of each organ. For score 2, bad segmentation were allocated for each organ as presented in Figure 4.2.

4.3.3 Results

Testing loss was evaluated with the discriminator module as a function of epoch and was presented in Figure 4.3. The testing loss is essentially constant after 175 epochs, which is consistent with the validation performance. Based on the lowest validation loss value from validation curve, epoch 198 model was chosen with the red arrow shown in Figure 4.3. For the performance of multi-organ segmentation, a significant increase in segmentation quality for the unlabeled datasets is shown in Figure 4.4 and 951 patients compared to the baseline of 714 patients ($p<0.0001$). Additionally, the number of outliers has been reduced from 29.90% to 19.83% of all testing datasets. Quality of segmentation with each code are shown in Figure 4.5 and the accuracy for code 0 segmentation are increased comparing with the baseline. For the previously failed scans, the segmentation becomes more detailed, and the incorrect locations for the liver and spleen are corrected as shown in Figure 4.5.

4.4 Conclusion

The proposed semi-supervised learning method with the integration of using QA scores to train the discriminator leads to more robust and effective segmentation performance. With the quality score predicted in each epoch for unlabeled data, the quality score acts as a loss function to provide feedback information for the original segmentation pipeline, improving the performance in outliers for segmentation. Hence, the use of QA-inspired loss functions represents an important perspective in supervised and semi-supervised learning, particularly for datasets with limited labels in medical image analysis.

CHAPTER 5

3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation

5.1 Overview

[1]The recent 3D medical ViTs (e.g., SwinUNETR) achieve the state-of-the-art performances on several 3D volumetric data benchmarks, including 3D medical image segmentation. Hierarchical transformers (e.g., Swin Transformers) reintroduced several ConvNet priors and further enhanced the practical viability of adapting volumetric segmentation in 3D medical datasets. The effectiveness of hybrid approaches is largely credited to the large receptive field for non-local self-attention and the large number of model parameters. We hypothesize that volumetric ConvNets can simulate the large receptive field behavior of these learning approaches with fewer model parameters using depth-wise convolution. In this work, we propose a lightweight volumetric ConvNet, termed 3D UX-Net, which adapts the hierarchical transformer using ConvNet modules for robust volumetric segmentation. Specifically, we revisit volumetric depth-wise convolutions with large kernel (LK) size (e.g. starting from $7 \times 7 \times 7$) to enable the larger global receptive fields, inspired by Swin Transformer. We further substitute the multi-layer perceptron (MLP) in Swin Transformer blocks with pointwise depth convolutions and enhance model performances with fewer normalization and activation layers, thus reducing the number of model parameters. 3D UX-Net competes favorably with current SOTA transformers (e.g. SwinUNETR) using three challenging public datasets on volumetric brain and abdominal imaging: 1) MICCAI Challenge 2021 FLARE, 2) MICCAI Challenge 2021 FeTA, and 3) MICCAI Challenge 2022 AMOS. 3D UX-Net consistently outperforms SwinUNETR with improvement from 0.929 to 0.938 Dice (FLARE2021) and 0.867 to 0.874 Dice (Feta2021). We further evaluate the transfer learning capability of 3D UX-Net with AMOS2022 and demonstrates another improvement of 2.27% Dice (from 0.880 to 0.900). The source code with our proposed model are available at https://github.com/MASILab/3DUX-Net.

5.2 Introduction

Significant progress has been made recently with the introduction of vision transformers (ViTs) (52) into 3D medical downstream tasks, especially for volumetric segmentation benchmarks (218; 74; 253; 229; 29). The characteristics of ViTs are the lack of image-specific inductive bias and the scaling behaviour, which are enhanced by large model capacities and dataset sizes. Both characteristics contribute to the significant

[1]Published at: Lee, Ho Hin, et al. "3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation.", The Eleventh International Conference on Learning Representations (ICLR), 2023. (122)

improvement compared to ConvNets on medical image segmentation (203; 12; 76; 4). However, it is challenging to adapt 3D ViT models as generic network backbones due to the high complexity of computing global self-attention with respect to the input size, especially in high resolution images with dense features across scales. Therefore, hierarchical transformers are proposed to bridge these gaps with their intrinsic hybrid structure (245; 145). Introducing the "sliding window" strategy into ViTs termed Swin Transformer behave similarly with ConvNets (145). SwinUNETR adapts Swin transformer blocks as the generic vision encoder backbone and achieves current state-of-the-art performance on several 3D segmentation benchmarks (73; 203). Such performance gain is largely owing to the large receptive field from 3D shift window multi-head self-attention (MSA). However, the computation of shift window MSA is computational unscalable to achieve via traditional 3D volumetric ConvNet architectures. As the advancement of ViTs starts to bring back the concepts of convolution, the key components for such large performance differences are attributed to the **scaling behavior** and **global self-attention with large receptive fields**. As such, we further ask: **Can we leverage convolution modules to enable the capabilities of hierarchical transformers?**

The recent advance in LK-based depthwise convolution design (e.g., Liu et al. (146)) provides a computationally scalable mechanism for large receptive field in 2D ConvNet. Inspired by such design, this study revisits the 3D volumetric ConvNet design to investigate the feasibility of (1) **achieving the SOTA performance via a pure ConvNet architecture**, (2) **yielding much less network complexity compared with 3D ViTs**, and (3) **providing a new direction of designing 3D ConvNet on volumetric high resolution tasks**. Unlike SwinUNETR, we propose a lightweight volumetric ConvNet 3D UX-Net to adapt the intrinsic properties of Swin Transformer with ConvNet modules and enhance the volumetric segmentation performance with smaller model capacities. Specifically, we introduce volumetric depth-wise convolutions with LK sizes to simulate the operation of large receptive fields for generating self-attention in Swin transformer. Furthermore, instead of linear scaling the self-attention feature across channels, we further introduce the pointwise depth convolution scaling to distribute each channel-wise feature independently into a wider hidden dimension (e.g., $4\times$input channel), thus minimizing the redundancy of learned context across channels and preserving model performances without increasing model capacity. We evaluate 3D UX-Net on supervised volumetric segmentation tasks with three public volumetric datasets: 1) MICCAI Challenge 2021 FeTA (infant brain imaging), 2) MICCAI Challenge 2021 FLARE (abdominal imaging), and 3) MICCAI Challenge 2022 AMOS (abdominal imaging). Surprisingly, 3D UX-Net, a network constructed purely from ConvNet modules, demonstrates a consistent improvement across all datasets comparing with current transformer SOTA. We summarize our contributions as below:

- We propose the 3D UX-Net to adapt transformer behavior purely with ConvNet modules in a volumet-

ric setting. To our best knowledge, this is the first large kernel block design of leveraging 3D depthwise convolutions to compete favorably with transformer SOTAs in volumetric segmentation tasks.

- We leverage depth-wise convolution with LK size as the generic feature extraction backbone, and introduce pointwise depth convolution to scale the extracted representations effectively with less parameters.
- We use three challenging public datasets to evaluate 3D UX-Net in 1) direct training and 2) fine-tuning scenarios with volumetric multi-organ/tissues segmentation. 3D UX-Net achieves consistently improvement in both scenarios across all ConvNets and transformers SOTA with fewer model parameters.

5.3 Related Work

5.3.1 Transformer-based Segmentation

Significant efforts have been put into integrating ViTs for dense predictions in medical imaging domain (74; 29; 253; 218). With the advancement of Swin Transformer, SwinUNETR equips the encoder with the Swin Transformer blocks to compute self-attention for enhancing brain tumor segmentation accuracy in 3D MRI Images (73). Tang et al. extends the SwinUNETR by adding a self-supervised learning pre-training strategy for fine-tuning segmentation tasks. Another Unet-like architecture Swin-Unet further adapts Swin Transformer on both the encoder and decoder network via skip-connections to learn local and global semantic features for multi-abdominal CT segmentation (21). Similarly, SwinBTS has the similar intrinsic structure with Swin-Unet with an enhanced transformer module for detailed feature extraction (104). However, the transformer-based volumetric segmentation frameworks still require lengthy training time and are accompanied by high computational complexity associated with extracting features at multi-scale levels (229; 187). Therefore, such limitations motivate us to rethink if ConvNets can emulate transformer behavior to demonstrate efficient feature extraction.

5.3.2 Depthwise convolution based Segmentation

Apart from transformer-based framework, previous works began to revisit the concept of depthwise convolution and adapt its characteristics for robust segmentation. It has been proved to be a powerful variation of standard convolution that helps reduce the number of parameters and transfer learning (71). Zunair et al. introduced depthwise convolution to sharpen the features prior to fuse the decode features in a UNet-like architecture (261). 3D U^2-Net leveraged depthwise convolutions as domain adaptors to extract domain-specific features in each channel (93). Both studies demonstrate the feasibility of using depthwise con-

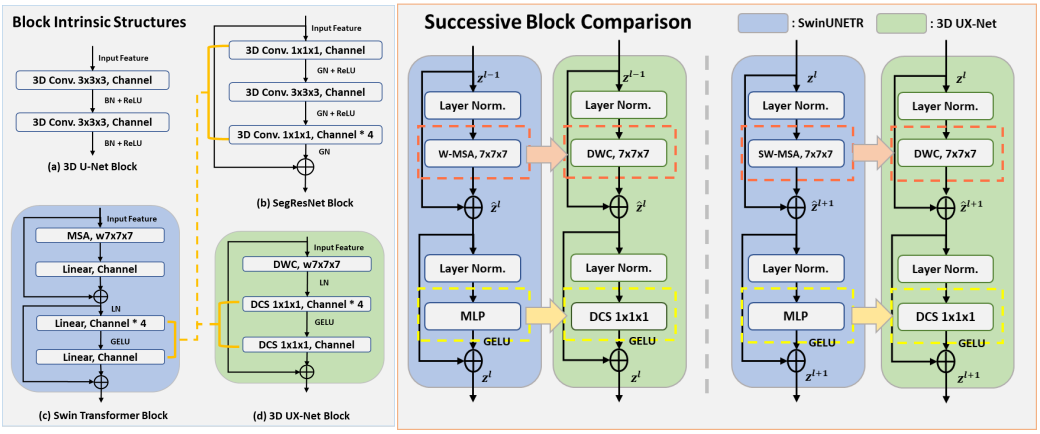


Figure 5.1: Overview of our proposed designed convolution blocks to simulate the behaviour of swin transformers. We leverage depthwise convolution and pointwise scaling to adapt large receptive field and enrich the features through widening independent channels. We further compare different backbones of volumetric ConvNets and Swin Transformer block architecture. The yellow dotted line demonstrates the differences in spatial position of widening feature channels in the network bottleneck.

volution in enhancing volumetric tasks. However, only a small kernel size is used to perform depthwise convolution. Several prior works have investigated the effectiveness of LK convolution in medical image segmentation. For instance, Huo et al. leveraged LK (7x7) convolutional layers as the skip connections to address the anatomical variations for splenomegaly spleen segmentation (96); Li et al. proposed to adapt LK and dilated depthwise convolutions in decoder for volumetric segmentation (137). However, significant increase of FLOPs is demonstrated with LKs and dramatically reduces both training and inference efficiency. To enhance the model efficiency with LKs, Liu et al. proposed ConvNeXt as a 2D generic backbone that simulate ViTs advantages with LK depthwise convolution for downstream tasks with natural image (146), while ConvUNeXt is proposed to extend for 2D medical image segmentation and compared only with 2D CNN-based networks (e.g., ResUNet (192), UNet++ (257)) (72). However, limited studies have been proposed to efficiently leverage depthwise convolution with LKs in a volumetric setting and compete favorably with volumetric transformer approaches. With the large receptive field brought by LK depthwise convolution, we hypothesize that LK depthwise convolution can potentially emulate transformers' behavior and efficiently benefits for volumetric segmentation.

5.4 3D UX-Net: Intuition

Inspired by (146), we introduce 3D UX-Net, a simple volumetric ConvNet that adapts the capability of hierarchical transformers and preserves the advantages of using ConvNet modules such as inductive biases. The basic idea of designing the encoder block in 3D UX-Net can be divided into 1) block-wise and 2) layer-wise perspectives. First, we discuss the block-wise perspective in three views:

- **Patch-wise Features Projection**: Comparing the similarities between ConvNets and ViTs, there is a common block that both networks use to aggressively downscale feature representations into particular patch sizes. Here, instead of flattening image patches as a sequential input with linear layer (52), we adopt a LK projection layer to extract patch-wise features as the encoder's inputs.

- **Volumetric Depth-wise Convolution with LKs**: One of the intrinsic specialties of the swin transformer is the sliding window strategy for computing non-local MSA. Overall, there are two hierarchical ways to compute MSA: 1) window-based MSA (W-MSA) and 2) shifted window MSA (SW-MSA). Both ways generate global receptive field across layers and further refine the feature correspondence between non-overlapping windows. Inspired by the idea of depth-wise convolution, we have found similarities between the weighted sum approach in self-attention and the convolution per-channel basis. We argue that using depth-wise convolution with a LK size can provide a large receptive field in extracting features similar to the MSA blocks. Therefore, we propose compressing the window shifting characteristics of the Swin Transformer with a volumetric depth-wise convolution using a LK size (e.g., starting from $7 \times 7 \times 7$). Each kernel channel is convolved with the corresponding input channel, so that the output feature has the same channel dimension as the input.

- **Inverted Bottleneck with Depthwise Convolutional Scaling**: Another intrinsic structure in Swin Transformer is that they are designed with the hidden dimension of the MLP block to be four times wider than the input dimension, as shown in Figure 5.1. Such a design is interestingly correlated to the expansion ratio in the ResNet block (79). Therefore, we leverage the similar design in ResNet block and move up the depth-wise convolution to compute features. Furthermore, we introduce depthwise convolutional scaling (DCS) with $1 \times 1 \times 1$ kernel to linearly scale each channel feature independently. We enrich the feature representations by expanding and compressing each channel independently, thus minimizing the redundancy of cross-channel context. We enhance the cross-channel feature correspondences with the downsampling block in each stage. By using DCS, we further reduce the model complexity by 5% and demonstrates a comparable results with the block architecture using MLP.

The macro-design in convolution blocks demonstrates the possibility of adapting the large receptive field and leveraging similar operation of extracting features compared with the Swin Transformer. We want to further investigate the variation between ConvNets and the Swin Transformer in layer-wise settings and refine the model architecture to better simulate ViTs in macro-level. Here, we further define and adapt layer-wise differences into another three perspectives:

- **Applying Residual Connections**: From Figure 5.1, the golden standard 3D U-Net block demonstrates

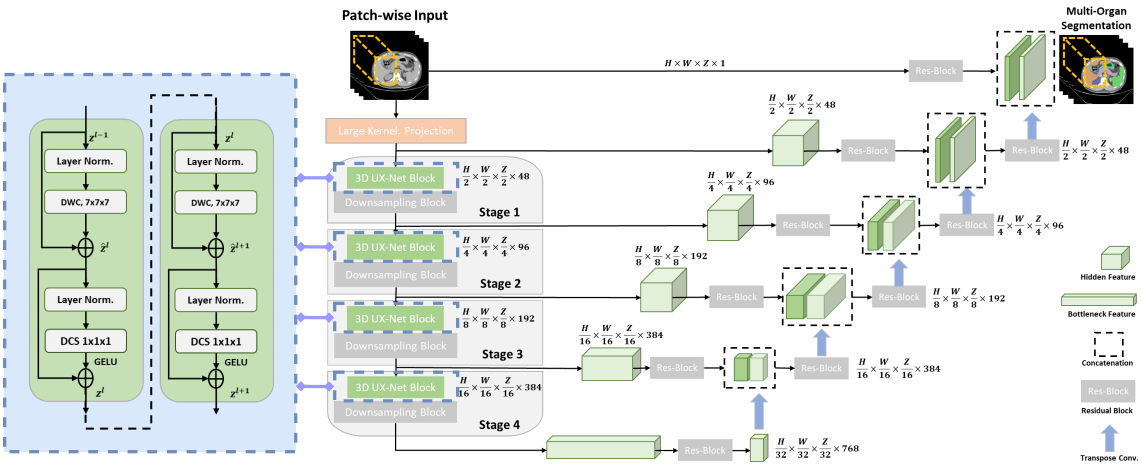


Figure 5.2: Overview of the proposed 3D UX-Net with our designed convolutional block as the encoder backbone. LK convolution is used to project features into patch-wise embeddings. A downsampling block is used in each stage to mix and enrich context across all channels, while our designed blocks extract meaningful features in depth-wise setting.

the naive approach of using small kernels to extract local representations with increased channels (39), while the SegResNet block applies the residual similar to the transformer block (160). Here, we also apply residual connections between the input and the extracted features after the last scaling layer. However, we do not apply any normalization and activation layers before and after the summation of residual to be equivalent with the swin transformer structure.

- **Adapting Layer Normalization (LN)**: In ConvNets, batch normalization (BN) is a common strategy that normalizes convolved representations to enhance convergence and reduce overfitting. However, previous works have demonstrated that BN can lead to a detrimental effect in model generalizability (228). Although several approaches have been proposed to have an alternative normalization techniques (186; 208; 227), BN still remains as the optimal choice in volumetric vision tasks. Motivated by vision transformers and (146), we directly substitute BN with LN in the encoder block and demonstrate similar operations in ViTs (6).

- **Using GELU as the Activation Layer**: Many previous works have used the rectified linear unit (ReLU) activation layers (161), providing non-linearity in both ConvNets and ViTs. However, previously proposed transformer models demonstrate the Gaussian error linear unit (GELU) to be a smoother variant, which tackle the limitation of sudden zero in the negative input range in ReLU (87). Therefore, we further substitute the ReLU with the GELU activation function.

5.5 3D UX-Net: Complete Network Description

3D UX-Net comprises multiple re-designed volumetric convolution blocks that directly utilize 3D patches. Skip connections are further leveraged to connect the multi-resolution features to a convolution-based decoder network. Figure 5.2 illustrates the complete architecture of 3D UX-Net. We further describe the details of the encoder and decoder in this section.

5.5.1 Depth-wise Convolution Encoder

Given a set of 3D image volumes $V_i = X_i, Y_{i i=1,\ldots,L}$, random sub-volumes $P_i \in \mathscr{R}^{H \times W \times D \times C}$ are extracted to be the inputs for the encoder network. Instead of flattening the patches and projecting it with linear layer (74), we leverage a LK convolutional layer to compute partitioned feature map with size $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$ that are projected into a $C = 48$-dimensional space. To adapt the characteristics of computing local self-attention, we use the depthwise convolution with kernel size starting from $7 \times 7 \times 7$ (DWC) with padding of 3, to act as a "shifted window" and evenly divide the feature map. As global self-attention is generally not computationally affordable with a large number of patches extracted in the Swin Transformer (145), we hypothesize that performing depthwise convolution with a LK size can effectively extract features with a global receptive field. Therefore, we define the output of encoder blocks in layers l and $l+1$ as follows:

$$\begin{aligned} \hat{z}^{l} &= \mathrm{DWC}(\mathrm{LN}(z^{l-1})) + z^{l-1} \\ z^{l} &= \mathrm{DCS}(\mathrm{LN}(\hat{z}^{l})) + \hat{z}^{l} \\ \hat{z}^{l+1} &= \mathrm{DWC}(\mathrm{LN}(z^{l})) + z^{l} \\ z^{l+1} &= \mathrm{DCS}(\mathrm{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{5.1}$$

where $\hat{z}_l$ and $\hat{z}_{l+1}$ are the outputs from the DWC layer in different depth levels; LN and DCS denote as the layer normalization and the depthwise convolution scaling, respectively (see. Figure 5.1). Compared to the Swin Transformer, we substitute the regular and window partitioning multi-head self-attention modules, W-MSA and SW-MSA respectively, with two DWC layers.

Motivated by SwinUNETR (203; 73), the complete architecture of the encoder consists of 4 stages comprising of 2 LK convolution blocks at each stage (*i.e.* L=8 total layers). Inside the block, the DCS layer is followed by the DWC layer in each block. The DCS layer helps scale the dimension of the feature map (4 times of the input channel size) without increasing model parameters and minimizes the redundancy of the learned volumetric context across channels. To exchange the information across channels, instead of using MLP, we leverage a standard convolution block with kernel size $2 \times 2 \times 2$ with stride 2 to downscale the feature resolution by a factor of 2. The same procedure continues in stage 2, stage 3 and stage 4 with

Table 5.1: Comparison of transformer and ConvNet SOTA approaches on the Feta 2021 and FLARE 2021 testing dataset. (*: $p < 0.01$, with Wilcoxon signed-rank test to all SOTA approaches)

			FeTA 2021								FLARE 2021				
Methods	#Params	FLOPs	ECF	GM	WM	Vent.	Cereb.	DGM	BS	Mean	Spleen	Kidney	Liver	Pancreas	Mean
3D U-Net (39)	4.81M	135.9G	0.867	0.762	0.925	0.861	0.910	0.845	0.827	0.857	0.911	0.962	0.905	0.789	0.892
SegResNet (160)	1.18M	15.6G	0.868	0.770	0.927	0.865	0.911	0.867	0.825	0.862	0.963	0.934	0.965	0.745	0.902
RAP-Net (125)	38.2M	101.2G	0.880	0.771	0.927	0.862	0.907	0.879	0.832	0.865	0.946	0.967	0.940	0.799	0.913
nn-UNet (99)	31.2M	743.3G	0.883	0.775	0.930	0.868	**0.920**	0.880	0.840	0.870	0.971	0.966	0.976	0.792	0.926
TransBTS (218)	31.6M	110.4G	**0.885**	0.778	0.932	0.861	0.913	0.876	0.837	0.868	0.964	0.959	0.974	0.711	0.902
UNETR (74)	92.8M	82.6G	0.861	0.762	0.927	0.862	0.908	0.868	0.834	0.860	0.927	0.947	0.960	0.710	0.886
nnFormer (253)	149.3M	240.2G	0.880	0.770	0.930	0.857	0.903	0.876	0.828	0.863	0.973	0.960	0.975	0.717	0.906
SwinUNETR (73)	62.2M	328.4G	0.873	0.772	0.929	0.869	0.914	0.875	0.842	0.867	0.979	0.965	0.980	0.788	0.929
3D UX-Net (Ours)	53.0M	639.4G	0.882	**0.780**	**0.934**	**0.872**	0.917	**0.886**	**0.845**	**0.874***	**0.981**	**0.969**	**0.982**	**0.801**	**0.934***

Table 5.2: Comparison of Finetuning performance with transformer SOTA approaches on the AMOS 2021 testing dataset.(*: $p < 0.01$, with Wilcoxon signed-rank test to all SOTA approaches)

Methods	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg
nn-UNet	0.965	0.959	0.951	0.889	0.820	0.980	0.890	0.948	0.901	0.821	0.785	0.739	0.806	0.869	0.839	0.878
TransBTS	0.885	0.931	0.916	0.817	0.744	0.969	0.837	0.914	0.855	0.724	0.630	0.566	0.704	0.741	0.650	0.792
UNETR	0.926	0.936	0.918	0.785	0.702	0.969	0.788	0.893	0.828	0.732	0.717	0.554	0.658	0.683	0.722	0.762
nnFormer	0.935	0.904	0.887	0.836	0.712	0.964	0.798	0.901	0.821	0.734	0.665	0.587	0.641	0.744	0.714	0.790
SwinUNETR	0.959	0.960	0.949	0.894	0.827	0.979	0.899	0.944	0.899	0.828	0.791	**0.745**	0.817	0.875	0.841	0.880
3D UX-Net	**0.970**	**0.967**	**0.961**	**0.923**	**0.832**	**0.984**	**0.920**	**0.951**	**0.914**	**0.856**	**0.825**	0.739	**0.853**	**0.906**	**0.876**	**0.900***

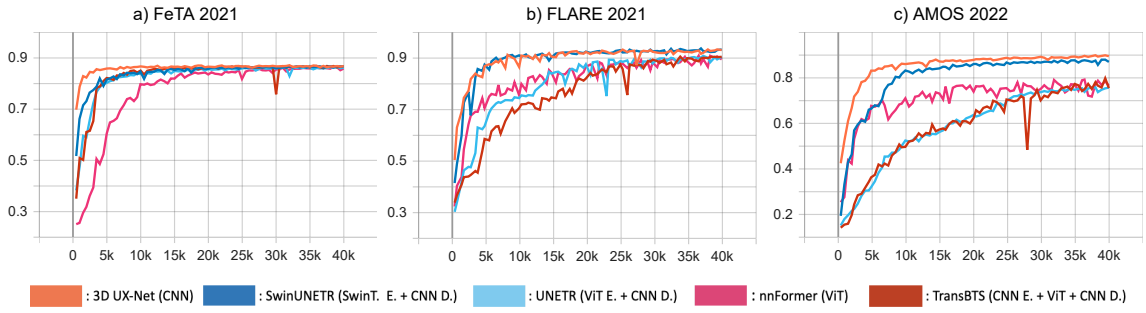


Figure 5.3: Validation Curve with Dice Score for FeTA2021 (a), FLARE2021 (b) and AMOS2022 (c). 3D UX-Net demonstrates the fastest convergence rate with limited samples training (FeTA2021) and transfer learning (AMOS2022) scenario respectively, while the convergence rate is comparable to SwinUNETR with the increase of sample size training (FLARE2021).

the resolutions of $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$ and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ respectively. Such hierarchical representations in multi-scale setting are extracted in each stage and are further leveraged for learning dense volumetric segmentation.

5.5.2 Decoder

The multi-scale output from each stage in the encoder is connected to a ConvNet-based decoder via skip connections and form a "U-shaped" like network for downstream segmentation task. Specifically, we extract the output feature mapping of each stage $i (i \in 0, 1, 2, 3, 4)$ in the encoder and further leverage a residual block comprising two post-normalized $3 \times 3 \times 3$ convolutional layers with instance normalization to stabilize the

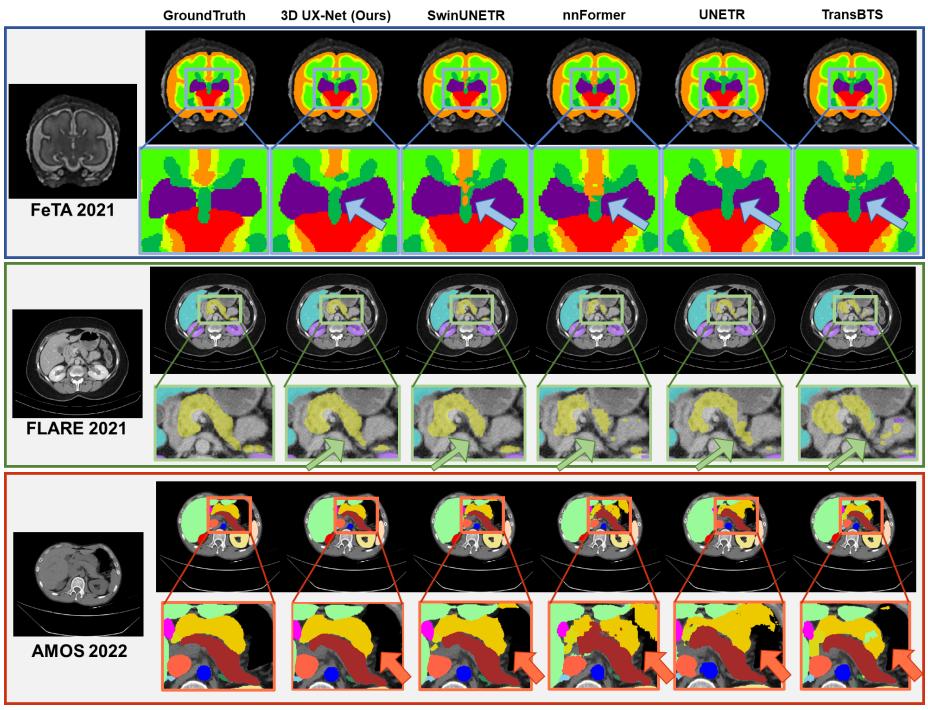


Figure 5.4: Qualitative representations of tissues and multi-organ segmentation across three public datasets. Boxed are further zoomed in and visualize the significant differences in segmentation quality. 3D UX-Net shows the best segmentation quality compared to the ground-truth.

extracted features. The processed features from each stage are then upsampled with a transpose convolutional layer and concatentated with the features from the preceding stage. For downstream volumetric segmentation, we also concatenate the residual features from the input patches with the upsampled features and input the features into a residual block with $1 \times 1 \times 1$ convolutional layer with a softmax activation to predict the segmentation probabilities.

5.6 Experimental Setup

Datasets We conduct experiments on three public multi-modality datasets for volumetric segmentation, which comprising with 1) MICCAI 2021 FeTA Challenge dataset (FeTA2021) (171), 2) MICCAI 2021 FLARE Challenge dataset (FLARE2021) (151), and 3) MICCAI 2022 AMOS Challenge dataset (AMOS2022) (102). For the FETA2021 dataset, we employ 80 T2-weighted infant brain MRIs from the University Children's Hospital with 1.5 T and 3T clinical whole-body scanners for brain tissue segmentation, with seven specific tissues well-annotated. For FLARE2021 and AMOS2022, we employ 511 multi-contrast abdominal CT from FLARE2021 with four anatomies manually annotated and 200 multi-contrast abdominal CT from AMOS 2022 with sixteen anatomies manually annotated for abdominal multi-organ segmentation. More details of the three public datasets can be found in appendix A.2.

Table 5.3: Ablation studies of different architecture on FeTA2021 and FLARE2021

Methods	#Params (M)	FeTA2021 Mean Dice	FLARE2021 Mean Dice
SwinUNETR	62.2	0.867	0.929
Use Standard Conv.	186.9	0.875	0.937
Use Depth Conv.	53.0	0.874	0.934
Kernel=$3 \times 3 \times 3$	52.5	0.867	0.928
Kernel=$5 \times 5 \times 5$	52.7	0.869	0.931
Kernel=$7 \times 7 \times 7$	53.0	0.874	0.934
Kernel=$9 \times 9 \times 9$	53.6	0.870	0.934
Kernel=$11 \times 11 \times 11$	54.4	0.871	0.936
Kernel=$13 \times 13 \times 13$	55.7	0.871	0.938
No MLP	51.1	0.869	0.915
Use MLP	56.3	0.872	0.933
Use DCS $1 \times 1 \times 1$	53.0	0.874	0.934

Implementation Details We perform evaluations on two scenarios: 1) direct supervised training and 2) transfer learning with pretrained weights. FeTA2021 and FLARE2021 datasets are leverage to evaluate in direct training scenario, while AMOS dataset is used in transfer learning scenario. We perform five-fold cross-validations to both FeTA2021 and FLARE2021 datasets. More detailed information of data splits are provided in Appendix A.2. For the transfer learning scenario, we leverage the pretrained weights from the best fold model trained with FLARE2021, and finetune the model weights on AMOS2022 to evaluate the fine-tuning capability of 3D UX-Net. The complete preprocessing and training details are available at the appendix A.1. Overall, we evaluate 3D UX-Net performance by comparing with current volumetric transformer and ConvNet SOTA approaches for volumetric segmentation in fully-supervised setting. We use the Dice similarity coefficient as an evaluation metric to compare the overlapping regions between predictions and ground-truth labels. Furthermore, we performed ablation studies to investigate the effect on different kernel size and the variability of substituting linear layers with depthwise convolution for feature extraction.

5.7 Results

5.7.1 Evaluation on FeTA & FLARE

Table 5.1 shows the result comparison of current transformers and ConvNets SOTA on medical image segmentation in volumetric setting. With our designed convolutional blocks as the encoder backbone, 3D UX-Net demonstrates the best performance across all segmentation task with significant improvement in Dice score (FeTA2021: 0.870 to 0.874, FLARE2021: 0.929 to 0.934). From Figure 5.2, we observe that 3D UX-Net demonstrates the quickest convergence rate in training with FeTA2021 datasets. Interestingly, when

the training sample size increases, the efficiency of training convergence starts to become compatible between SwinUNETR and 3D UX-Net. Apart from the quantitative representations, Figure 5.3 further provides additional confidence of demonstrating the quality improvement in segmentation with 3D UX-Net. The morphology of organs and tissues are well preserved compared to the ground-truth label.

5.7.2 Transfer Learning with AMOS

Apart from training from scratch scenario, we further investigate the transfer learning capability of 3D UX-Net comparing to the transformers SOTA with AMOS 2022 dataset. We observe that the finetuning performance of 3D UX-Net significantly outperforms other transformer network with mean Dice of 0.900 (2.27% enhancement) and most of the organs segmentation demonstrate a consistent improvement in quality. Also, from Figure 5.2, although the convergence curve of each transformer network shows the comparability to that of the FLARE2021-trained model, 3D UX-Net further shows its capability in adapting fast convergence and enhancing the robustness of the model with finetuning. Furthermore, the qualitative representations in Figure 5.2 demonstrates a significant improvement in preserving boundaries between neighboring organs and minimize the possibility of over-segmentation towards other organ regions.

5.7.3 Ablation Analysis

After evaluating the core performance of 3D UX-Net, we study how the different components in our designed architecture contribute to such a significant improvement in performance, as well as how they interact with other components. Here, both FeTA2021 and FLARE2021 are leveraged to perform ablation studies towards different modules. All ablation studies are performed with kernel size $7 \times 7 \times 7$ scenario except the study of evaluating the variability of kernel size.

Comparing with Standard Convolution: We investigate the effectiveness of both standard convolution and depthwise convolution for initial feature extraction. With the use of standard convolution, it demonstrates a slight improvement with standard convolution. However, the model parameters are about 3.5 times than that of using depthwise convolution, while the segmentation performance with depthwise convolution still demonstrates a comparable performance in both datasets.

Variation of Kernel Size: From Table 5.3, we observe that the convolution with kernel size $7 \times 7 \times 7$ optimally works for FeTA2021 dataset, while the segmentation performance of FLARE2021 demonstrates the best with kernel size of $13 \times 13 \times 13$. The significant improvement of using $13 \times 13 \times 13$ kernel for FLARE2021 may be due to the larger receptive field provided to enhance the feature correspondence between multiple neighboring organs within the abdominal region. For FeTA2021 dataset, only the small infant brains are well localized as foreground and $7 \times 7 \times 7$ kernel demonstrates to be optimal receptive field to

extract the tissues correspondence.

Adapting DCS: We found that a significant decrement is performed without using MLP for feature scaling. With the linear scaling, the performance enhanced significantly in FLARE2021, while a slight improvement is demonstrated in FeTA2021. Interestingly, leveraging depthwise convolution with $1 \times 1 \times 1$ kernel size for scaling, demonstrates a slightly enhancement in performance for both FeTA2021 and FLARE2021 datasets. Also, the model parameters further drops from 56.3M to 53.0M without trading off the model performance.

5.8 Discussion

In this work, we present a block-wise design to simulate the behavior of Swin Transformer using pure ConvNet modules. We further adapt our design as a generic encoder backbone into "U-Net" like architecture via skip connections for volumetric segmentation. We found that the key components for improved performance can be divided into two main perspectives: 1) the sliding window strategy of computing MSA and 2) the inverted bottleneck architecture of widening the computed feature channels. The W-MSA enhances learning the feature correspondence within each window, while the SW-MSA strengthens the cross-window connections at the feature level between different non-overlapping windows. Such strategy integrates ConvNet priors into transformer networks and enlarge receptive fields for feature extraction. However, we found that the depth convolutions can demonstrate similar operations of computing MSA in Swin Transformer blocks. In depth-wise convolutions, we convolve each input channel with a single convolutional filter and stack the convolved outputs together, which is comparable to the patch merging layer for feature outputs in Swin Transformers. Furthermore, adapting the depth convolutions with LK filters demonstrates similarities with both W-MSA and SW-MSA, which learns the feature connections within a large receptive field. Our design provides similar capabilities to Swin Transformer and additionally has the advantage of reducing the number of model parameters using ConvNet modules.

Another interesting difference is the inverted bottleneck architecture. Figure 5.1 shows that both Swin Transformer and some standard ConvNets have their specific bottleneck architectures (yellow dotted line). The distinctive component in Swin Transformer's bottleneck is to maintain the channels size as four times wider than the input dimension and the spatial position of the MSA layer. We follow the inverted bottleneck architecture in Swin Transformer block and move the depthwise convolution to the top similar to the MSA layer. Instead of using linear scaling, we introduce the idea of depthwise convolution in pointwise setting to scale the dense feature with wider channels. Interestingly, we found a slight improvement in performance is shown across datasets (FeTA2021: 0.872 to 0.874, FLARE2021: 0.933 to 0.934), but with less model parameters. As each encoder block only consists of two scaling layers, the limited number of scaling blocks may affect the performance to a small extent. We will further investigate the scalability of linear scaling layer

in 3D as the future work.

5.9 Conclusion

We introduce 3D UX-Net, the first volumetric network adapting the capabilities of hierarchical transformer with pure ConvNet modules for medical image segmentation. We re-design the encoder blocks with depth-wise convolution and projections to simulate the behavior of hierarchical transformer. Furthermore, we adjust layer-wise design in the encoder block and enhance the segmentation performance across different training settings. 3D UX-Net outperforms current transformer SOTAs with fewer model parameters using three challenging public datasets in both supervised training and transfer learning scenarios.

5.10 Supplementary

5.10.1 Data Preprocessing & Model Training

We apply hierarchical steps for data preprocessing: 1) intensity clipping is applied to further enhance the contrast of soft tissue (FLARE2021 & AMOS2022:{min:-175, max:250}). 2) Intensity normalization is performed after clipping for each volume and use min-max normalization: $(X - X_1)/(X_{99} - X_1)$ to normalize the intensity value between 0 and 1, where X_p denote as the p_{th} percentile of intensity in X. We then randomly crop sub-volumes with size $96 \times 96 \times 96$ at the foreground and perform data augmentations, including rotations, intensity shifting, and scaling (scaling factor: 0.1). All training processes with 3D UX-Net are optimized with an AdamW optimizer. We trained all models for 40000 steps using a learning rate of 0.0001 on an NVIDIA-Quadro RTX 5000 for both FeTA2021 and FLARE2021, while we perform training for AMOS2022 using NVIDIA-Quadro RTX A6000. One epoch takes approximately about 1 minute for FeTA2021, 10 minutes for FLARE2021, and 7 minutes for AMOS2022, respectively. We further summarize all the training parameters with Table 5.4.

5.10.2 Public Datasets Details

5.10.3 Further Discussions Comparing to nn-UNet

In Table 5.1 & 5.2, we compare our proposed network with multiple CNN-based SOTA networks and the golden standard approach nn-UNet. We observe that the performance of nn-UNet nearly outperform most of the transformer state-of-the-arts in both FeTA 2021 and FLARE 2021 datasets. Such improvement may mainly contribute to its innovation of self-configuration training strategies and ensembling outputs as post-processing technique, while the network used in nn-UNet is only the plain 3D U-Net architecture. To further characterize the ability of our proposed network, we further substitute the plain 3D U-Net architecture with our proposed 3D UX-Net and adapt the self-configuring hyperparameters for training. We demonstrate a

Table 5.4: Hyperparameters of both directly training and finetuning scenarios on three public datasets

Hyperparameters	**Direct Training**	**Finetuning**
Encoder Stage	4	
Layer-wise Channel	$48, 96, 192, 384$	
Hidden Dimensions	768	
Patch Size	$96 \times 96 \times 96$	
No. of Sub-volumes Cropped	2	1
Training Steps	40000	
Batch Size	2	1
AdamW ε	$1e-8$	
AdamW β	$(0.9, 0.999)$	
Peak Learning Rate	$1e-4$	
Learning Rate Scheduler	ReduceLROnPlateau	N/A
Factor & Patience	0.9, 10	N/A
Dropout	X	
Weight Decay	0.08	
Data Augmentation	Intensity Shift, Rotation, Scaling	
Cropped Foreground	✓	
Intensity Offset	0.1	
Rotation Degree	-30° to $+30^\circ$	
Scaling Factor	x: 0.1, y: 0.1, z: 0.1	

Table 5.5: Complete Overview of three public MICCAI Chanllenge Datasets

MICCAI Challenge	FeTA 2021	FLARE 2021	AMOS 2022
Imaging Modality	1.5T & 3T MRI	Multi-Contrast CT	Multi-Contrast CT
Anatomical Region	Infant Brain	Abdomen	Abdomen
Dimensions	$256 \times 256 \times 256$	$512 \times 512 \times \{37-751\}$	$512-768 \times 512-768 \times \{68-353\}$
Resolution	$\{0.43-0.70\} \times \{0.43-0.70\} \times \{0.43-0.70\}$	$\{0.61-0.98\} \times \{0.61-0.98\} \times \{0.50-7.50\}$	$\{0.45-1.07\} \times \{0.45-1.07\} \times \{1.25-5.00\}$
Sample Size	80	361	200
Anatomical Label	External Cerebrospinal Fluid (ESF), Grey Matter (GM), White Matter (WM), Ventricles, Cerebellum, Deep Grey Matter (DGM) Brainstem	Spleen, Kidney, Liver, Pancreas	Spleen, Left & Right Kidney, Gall Bladder, Esophagus, Liver, Stomach, Aorta, Inferior Vena Cava (IVC) Pancreas, Left & Right Adrenal Gland (AG), Duodenum, Bladder, Prostates/uterus
Data Splits	5-Fold Cross-Validation Train: 50 / Validation: 12 / Test: 18	5-Fold Cross-Validation Train: 272 / Validation: 69 / Test: 20	1-Fold Train: 160 / Validation: 20 / Test: 20

Table 5.6: Albation Studies of Adapting nn-UNet architecture on the Feta 2021 and FLARE 2021 testing dataset. (*: $p < 0.01$, with Wilcoxon signed-rank test to all SOTA approaches, D.S: Deep Supervision)

	FeTA 2021								FLARE 2021				
Methods	ECF	GM	WM	Vent.	Cereb.	DGM	BS	Mean	Spleen	Kidney	Liver	Pancreas	Mean
nn-UNet (99)	0.883	0.775	0.930	0.868	0.920	0.880	0.840	0.870	0.971	0.966	0.976	0.792	0.926
3D UX-Net (Plain)	0.882	0.780	0.934	0.872	0.917	0.886	0.845	0.874	0.981	0.969	0.982	0.801	0.934
3D UX-Net (nn-UNet struct., w/o D.S.)	0.885	0.784	0.937	0.872	0.921	0.887	0.849	0.876	**0.983**	**0.972**	**0.983**	**0.821**	**0.940***
3D UX-Net (nn-UNet struct., D.S.)	**0.890**	**0.791**	**0.939**	**0.877**	**0.922**	**0.891**	**0.854**	**0.881***	**0.986**	**0.974**	**0.983**	**0.833**	**0.944***

significant improvement of performance in FeTA 2021 and FLARE 2021 datasets with mean organ Dice from 0.874 to 0.881 and from 0.934 to 0.944 respectively, as shown in Table 5.6. To further investigate the difference in the network architecture, we observed that the convolution blocks in nn-UNet leverage the combination of instance normalization and leakyReLU. Such design allows to normalize channel-wise feature independently and mix the channel context with small kernel convolutional layers. In our design, we provide an alternative thought of extracting channel-wise features independently with depthwise convolution and mix the channel information during the downsampling layer only. Therefore, layer normalization is leveraged in our scenario and we want to further enhance the feature correspondence with large receptive field across channels efficiently. Furthermore, we found that the deep supervision strategy in nn-UNet, which compute an auxiliary loss with each stages' intermediate output, also demonstrates its effectiveness to further improve the performance (FeTA 2021: from 0.876 to 0.881; FLARE 2021: from 0.940 to 0.944).

For the training scenarios, instead of using the proposed initial learning rate 0.01, we reduce the initial learning rate to 0.002 to train with 150 epochs (40000 steps $\approx$ 150 epochs) for FLARE 2021 and 850 epochs (40000 steps $\approx$ 850 epochs) for FeTA 2021 respectively, with the batch size of 2. For the finetuning scenario with AMOS 2022, we only train the nn-UNet model with 250 epochs (40000 steps $\approx$ 250 epochs), instead of the default settings (1000 epochs) to ensure the fair network comparison with similar steps.

5.10.4 Further Discussions on Training and Inference Efficiency

Table 5.7: Albation Studies of Optimizing 3D U-XNet architecture on the Feta 2021 and FLARE 2021 testing dataset. (SD: Stage Depth, HDim: Hidden Dimension in the Bottleneck Layer.)

Methods	#Params (M)	FLOPs (G)	FeTA2021 Mean Dice	FLARE2021 Mean Dice
nn-UNet	31.2M	743.3G	0.870	0.926
SwinUNETR	62.2M	328.4G	0.867	0.929
SD: 2,2,2,2, HDim: 768	53.0M	639.4G	0.874	0.934
SD: 2,2,8,2, HDim: 384	32.1M	536.8G	0.873	0.932

Apart from the advantage of quantitative performance, we further leverage the LK depthwise convolutions to reduce the model parameters from 62.2M to 53.0M, compared to SwinUNETR in Table 5.3. However, although the training efficiency of 3D UX-Net is already better than nn-UNet (FLOPs: 743.3G to 639.4G), we observed that the FLOPs of 3D UX-Net still remains at a high value. Inspired by the architectures of both Swin Transformer (145) and ConvNeXt (146) used in the natural image domain, we further remove the bottleneck layer (ResNet block with 768 channels) and increase the block depth of stage 3 (e.g., 8 blocks). Such optimized design further significantly reduces both the model parameters (from 53.0M to 32.1M, nn-UNet: 31.2M) and FLOPs (from 639.4G to 536.1G, nn-UNet: 743.3M), while preserving the performance (shown in Table 5.7). Additional validation studies is needed to investigate the effectiveness of both MLP and pointwise DCS, and optimizing 3D UX-Net architecture, which will be the next steps of our future work. Another observation in Table 5.3 is the subtle differences in model parameters between kernel size of $3 \times 3 \times 3$ and $7 \times 7 \times 7$. We found that the increase of both model parameters and FLOPs is also attributed to the design of decoder network. Our decoder block design further add a 3D ResNet block after the transpose convolution to further resample and mix the channel context, instead of directly perform transpose convolution in nn-UNet. A efficient block design in decoder network is demanded to be further investigated and using depthwise convolution may be another potential solution to reduce the low efficiency burden.

To further reduce the burden of low training and inference efficiency, re-parameterization of LK convolutional blocks may be another promising direction to focus. Prior works have demonstrated to scale up few convolutional blocks with LK size (31×31) and propose the idea of parallel branches with small kernels for residual shortcuts (48; 49; 47). The parallel branch can then be mutually converted through equivalent transformation of parameters. For example. a branch of 1×1 convolution and a branch of 7×7 convolution, can be transferred into a single branck of 7×7 convolution (49). Furthermore, Hu et al. proposed online convolutional re-parameterization (OREPA) to leverage a linear scaling at each branch to diversify the optimization directions, instead of applying non-linear normalization after convolution layer (91). Also, stack of small kernels are leveraged to generate similar receptive field of view as LKs with better training and inference efficiency. The effectiveness of leveraging small kernels stack and multiple parallel branches design will be further investigated as another directions of our future work.

CHAPTER 6

Scaling Up 3D Kernels with Bayesian Frequency Re-parameterization for Medical Image Segmentation

6.1 Overview

[1]With the inspiration of vision transformers, the concept of depth-wise convolution revisits to provide a large Effective Receptive Field (ERF) using Large Kernel (LK) sizes for medical image segmentation. However, the segmentation performance might be saturated and even degraded as the kernel sizes scaled up (e.g., $21 \times 21 \times 21$) in a Convolutional Neural Network (CNN). We hypothesize that convolution with LK sizes is limited to maintain an optimal convergence for locality learning. While Structural Re-parameterization (SR) enhances the local convergence with small kernels in parallel, optimal small kernel branches may hinder the computational efficiency for training. In this work, we propose RepUX-Net, a pure CNN architecture with a simple large kernel block design, which competes favorably with current network state-of-the-art (SOTA) (e.g., 3D UX-Net, SwinUNETR) using 6 challenging public datasets. We derive an equivalency between kernel re-parameterization and the branch-wise variation in kernel convergence. Inspired by the spatial frequency in the human visual system, we extend to vary the kernel convergence into element-wise setting and model the spatial frequency as a Bayesian prior to re-parameterize convolutional weights during training. Specifically, a reciprocal function is leveraged to estimate a frequency-weighted value, which rescales the corresponding kernel element for stochastic gradient descent. From the experimental results, RepUX-Net consistently outperforms 3D SOTA benchmarks with internal validation (FLARE: 0.929 to 0.944), external validation (MSD: 0.901 to 0.932, KiTS: 0.815 to 0.847, LiTS: 0.933 to 0.949, TCIA: 0.736 to 0.779) and transfer learning (AMOS: 0.880 to 0.911) scenarios in Dice Score. Both codes and pretrained models are available at: https://github.com/MASILab/RepUX-Net

6.2 Introduction

With the introduction of Vision Transformers (ViTs), CNNs have been greatly challenged as seen with the leading performance in multiple volumetric data benchmarks, especially for medical image segmentation (74; 73; 203; 253). The key contribution of ViTs is largely credited to the large Effective Receptive Field (ERF) with a multi-head self-attention mechanism (52). Note the attention mechanism is computationally unscalable with respect to the input resolutions (145; 146). Therefore, the concept of depth-wise convolution

[1]Accepted at: Lee, Ho Hin, et al. "Scaling Up 3D Kernels with Bayesian Frequency Re-parameterization for Medical Image Segmentation.", International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2023. (123)

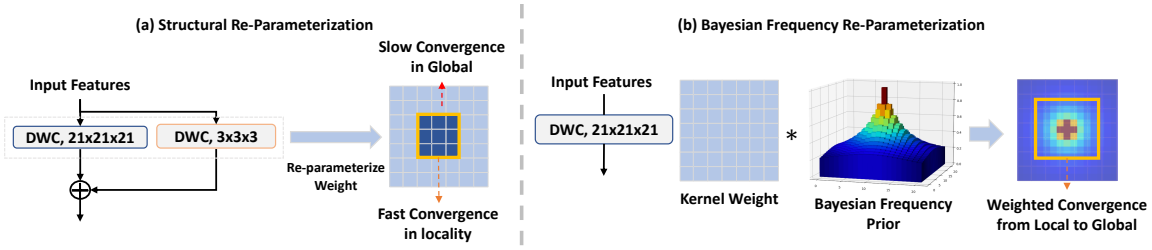


Figure 6.1: With the fast convergence in small kernels, SR merges the branches weights and enhances the locality convergence with respect to the kernel size (deep blue region), while the global convergence is yet to be optimal (light blue region). By adapting BFR, the learning convergence can rescale in an element-wise setting and distribute the learning importance from local to global.

is revisited to provide a scalable and efficient feature computation with large ERF using large kernel sizes (e.g., $7 \times 7 \times 7$) (146; 121). However, either from prior works or our experiments, the model performance becomes saturated or even degraded when the kernel size is scaled up in encoder blocks (48; 143). We hypothesize that scaling up the kernel size in convolution may limit the optimal learning convergences across local to global scales. Recently, the feasibility of leveraging large kernel convolutions (e.g., 31×31 (48), 51×51(143)) has been shown with natural image domain with Structural Re-parameterization (SR), which adapts Constant-Scale Linear Addition (CSLA) block (Figure 6.2(b)) and re-parameterizes the large kernel weights during inference (48). As convolutions with small kernel sizes converge more easily, the convergence of small kernel regions enhances in the re-parameterized weight, as shown in Figure 6.1(a). With such observation, we further ask: **Can we adapt variable convergence across elements of the convolution kernel during training, instead of regional locality only?**

In this work, we first derive and extend the theoretical equivalency of the weight optimization in the CSLA block. We observe that the kernel weight of each branch can be optimized with variable convergence using branch-specific learning rates. Furthermore, the ERF with SR is visualized to be more widely distributed from the center element to the global surroundings (48), demonstrating a similar behavior to the spatial frequency in the human visual system (114). Inspired by the reciprocal characteristics of spatial frequency, we model the spatial frequency as a Bayesian prior to adapt variable convergence of each kernel element with stochastic gradient descent (Figure 6.1(b)). Specifically, we compute a scaling factor with respect to the distance from the kernel center and multiply the corresponding element for re-parameterization during training. Furthermore, we simplify the encoder block design into a plain convolution block only to minimize the computation burden in training and achieve State-Of-The-Art (SOTA) performance. We propose RepUX-Net, a pure 3D CNN with the large kernel size (e.g., $21 \times 21 \times 21$) in encoder blocks, to compete favorably with current SOTA segmentation networks. We evaluate RepUX-Net on supervised multi-organ segmentation

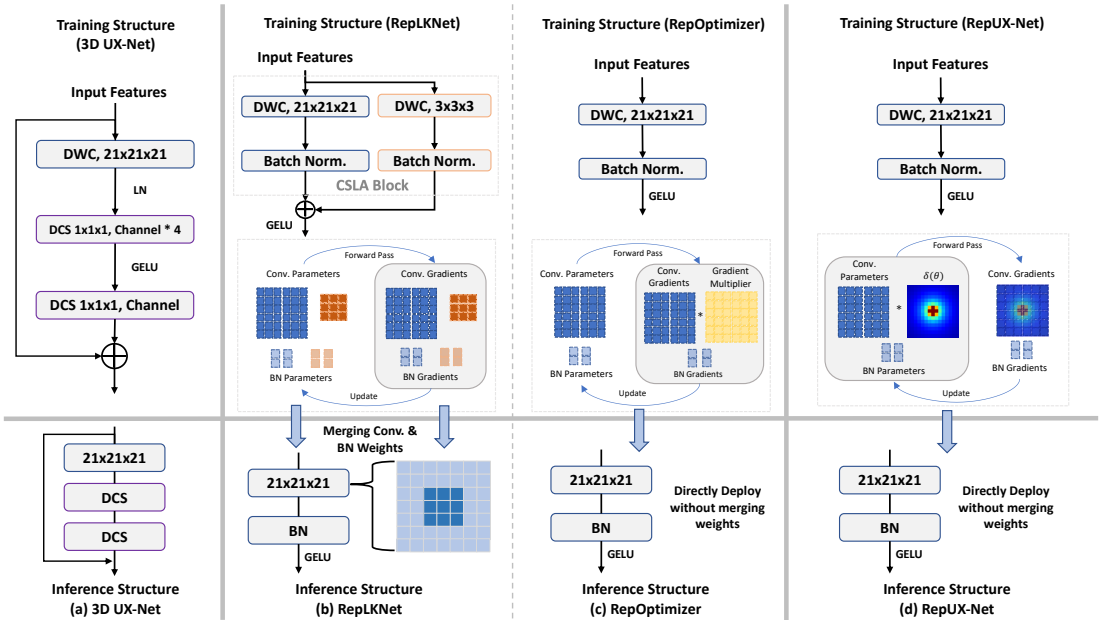


Figure 6.2: Overview of RepUX-Net. Unlike performing SR to merge branches weight or performing GR within optimizers, we propose to multiply a Bayesian function δ and scale the element-wise learning importance in each large kernel. We then put the scaled weights back into the convolution layer for training.

with 6 different public volumetric datasets. RepUX-Net demonstrates significant improvement consistently across all datasets compared to all SOTA networks. We summarize our contributions as below:

- We propose RepUX-Net with better adaptation in large kernel convolution than 3D UX-Net, achieving SOTA performance in 3D segmentation. To our best knowledge, this is the first network that effectively leverages large kernel convolution with plain design in the encoder for 3D segmentation.
- We propose a novel theory-inspired re-parameterization strategy to scale the element-wise learning convergence in large kernels with Bayesian prior knowledge. To our best knowledge, this is the first re-parameterization strategy to adapt 3D large kernels in the medical domain.
- We leverage six challenging public datasets to evaluate RepUX-Net in 1) direct training and 2) transfer learning scenarios with 3D multi-organ segmentation. RepUX-Net achieves significant improvement consistently in both scenarios across all SOTA networks.

6.3 Related Works

Weights Re-parameterization: SR is a methodology of equivalently converting model structures via transforming the parameters in kernel weights. For example, RepVGG demonstrates to construct one extra ResNet-style shortcut as a 1×1 convolution, parallel to 3×3 convolution during training (49). Such par-

allel branch design is claimed to enhance the learning efficiency during training, in which the 1x1 branch is then merged into the parallel 3×3 kernel via a series of linear transformation in the inference stage. OREPA further adds more parallel branches with linear scaling modules to enhance training efficiency (91). Inspired by the parallel branches design, RepLKNet is proposed to scale up the 2D kernel size (e.g., 31x31) with a 3x3 convolution as the parallel branch (48). SLaK further extends the kernel size to 51x51 by decomposing the large kernel into two rectangular parallel kernels with sparse groups and training the model with dynamic sparsity (143). However, the proposed models' FLOPs remain at a high-level with the parallel branch design and demonstrates to have a trade-off between model performance and training efficiency. To tackle the trade-off, RepOptimizer provides an alternative to re-parameterize the back-propagate gradient, instead of the structural parameters of kernel weights, to enhance the training efficiency with plain convolution block design (47). Significant efforts have been demonstrated to enlarge the 2D kernel size in the natural image domain, while limited studies have been proposed for 3D kernels in medical domain. As 3D kernels have a larger number of parameters than 2D, it is challenging to directly leverage the parallel branch design and maintain an optimal convergence of learning large kernel convolution without trading off the computation efficiency significantly.

6.4 Methods

Instead of changing the gradient dynamics during training (47), we introduce RepUX-Net, a pure 3D CNN architecture that performs element-wise scaling in large kernel weights to enhance the learning convergence and effectively adapts large receptive field for volumetric segmentation. To design such behavior, we adapt a two-step pipeline: 1) we define the theoretical equivalency of variable learning convergence in convolution branches; 2) we simulate the behavior of spatial frequency to re-weight the learning importance of each element in kernels for stochastic gradient descent. Note the theoretical derivation depends on the optimization with first-order gradient-driven optimizer (e.g., SGD, AdamW) (47).

6.4.1 Variable Learning Convergence in Multi-Branch Design

From Figure 6.2, the learning convergence of the large kernel convolution can be improved by either adding up the encoded outputs of parallel branches weighted by diverse scales with SR (RepLKNet (48)) or performing Gradient Re-parameterization (GR) by multiplying with constant values (RepOptimizer (47)) in a Single Operator (SO). Inspired by the concepts of SR and GR, we extend the equivalency proof in RepOptimizer to adapt variable learning convergence in branches. Here, we only showcase the conclusion with two convolutions and two constant scalars as the scaling factors for simplicity. The complete proof of equivalency is demonstrated in Supplementary 1.1. Let $\{\alpha_L, \alpha_S\}$ and $\{W_L, W_S\}$ be the two constant scalars and two convo-

lution kernels (Large & Small) respectively. Let X and Y be the input and output features, the CSLA block is formulated as $Y_{CSLA} = \alpha_L(X \star W_L) + \alpha_S(X \star W_S)$, where $\star$ denotes as convolution. For SO blocks, we train the plain structure parameterized by W' and $Y_{SO} = X \star W'$. Let i be the number of training iterations, we ensure that ${Y^{(i)}}_{CSLA} = {Y^{(i)}}_{SO}, \forall i \geq 0$ and derive the stochastic gradient descent of parallel branches as follows:

$$\alpha_L W_{L(i+1)} + \alpha_S W_{S(i+1)} = \alpha_L W_{L(i)} - \lambda_L \alpha_L \frac{\partial \mathscr{L}}{\partial W_{L_i}} + \alpha_S W_{S(i)} - \lambda_S \alpha_S \frac{\partial \mathscr{L}}{\partial W_{S_i}}, \tag{6.1}$$

where $\mathscr{L}$ is the objective function; λ_L and λ_S are the Learning Rate (LR) of each branch respectively. We observe that the optimization of each branch can be different by adjusting the branch-specific LR. The locality convergence in large kernels enhance with the quick convergence in small kernels. Additionally from our experiments, a significant improvement is demonstrated with different branch-wise LR using SGD (Table 6.2). With such observation, we further hypothesize that **the convergence of each large kernel element can be optimized differently by linear scaling with prior knowledge**.

6.4.2 Bayesian Frequency Re-parameterization (BFR)

With the visualization of ERF in RepLKNet (48), the diffused distribution (from local to global) in ERF demonstrates similar behavior with the spatial frequency in the human visual system (114). High spatial frequency (small ERF) allows to refine and sharpen details with high acuity, while global details are demonstrated with low spatial frequency. Inspired by the reciprocal characteristics in spatial frequency, we first generate a Bayesian prior distribution to model the spatial frequency by computing a reciprocal distance function between each element and the central point of the kernel weight as follows:

$$\begin{aligned} d(x,y,z,c) &= \sqrt{(x-c)^2 + (y-c)^2 + (z-c)^2} \\ \delta(x_k, y_k, z_k, c, \alpha) &= \frac{\alpha}{d(x_k, y_k, z_k, c) + \alpha} \end{aligned} \tag{6.2}$$

where k and c are the element and central index of the kernel weight, α is the hyperparameter to control the shape of the generated frequency distribution. Instead of adjusting the LR in parallel branches, we propose to re-parameterize the convolution weights by multiplying the scaling factor δ to each kernel element and apply a static LR λ for stochastic gradient descent in single operator setting as follows:

$$W_{i+1}^{'} = \delta W_i^{'} - \lambda \frac{\partial L}{\partial \delta W_i^{'}} \tag{6.3}$$

With the multiplication with δ, each element in the kernel weight is rescaled with respect to the frequency level and allow to converge differently with a static LR in stochastic gradient descent. Such design demon-

strates to influence the weighted convergence diffused from local to global in theory, thus tackling the limitation of enhancing the local convergence only in branch-wise setting.

6.4.3 Model Architecture

The backbone of RepUX-Net is based on 3D UX-Net (121), which comprises multiple volumetric convolution blocks that directly utilize 3D patches and leverage skip connections to transfer hierarchical multi-resolution features for end-to-end optimization. Inspired by (136), we choose a kernel size of $21 \times 21 \times 21$ for DepthWise Convolution (DWC-21) as the optimal choice without significant trade-off between model performance and computational efficiency in 3D. We further simplify the block design as a plain convolution block design to minimize the computational burden from additional modules. The encoder blocks in layers l and $l+1$ are defined as follows:

$$\hat{z}^{l} = \text{GeLU}(\text{DWC-21}(\text{BN}(z^{l-1}))), \hat{z}^{l+1} = \text{GeLU}(\text{DWC-21}(\text{BN}(z^{l}))) \tag{6.4}$$

where $\hat{z}_l$ and $\hat{z}_{l+1}$ are the outputs from the DWC layer in each depth level; BN denotes as the batch normalization layer.

6.5 Experimental Setup

Datasets We perform experiments on six public datasets for volumetric segmentation, which comprise with 1) Medical Segmentation Decathlon (MSD) spleen dataset (2), 2) MICCAI 2017 LiTS Challenge dataset (LiTS) (15), 3) MICCAI 2019 KiTS Challenge dataset (KiTS) (86), 4) NIH TCIA Pancreas-CT dataset (TCIA) (180), 5) MICCAI 2021 FLARE Challenge dataset (FLARE) (151), and 6) MICCAI 2022 AMOS challenge dataset (AMOS) (102). More details of each dataset (including data split for training and inference) are described in Supplementary Material (SM) Table 6.4.

Implementation We evaluate RepUX-Net with three different scenarios: 1) internal validation with direct supervised learning, 2) external validation with the unseen datasets, and 3) transfer learning with pretrained weights. All preprocessing and training details including baselines, are followed with (121) for benchmarking. For external validations, we leverage the AMOS-pretrained weights to evaluate 4 unseen datasets. In summary, we evaluate the segmentation performance of RepUX-Net by comparing current SOTA networks in a fully-supervised setting. Furthermore, we perform ablation studies to investigate the effect on Bayesian frequency distribution with different scales generated by α and the variability of branch-wise learning rates with first-order gradient optimizers (e.g., SGD, AdamW) for volumetric segmentation. Dice similarity coefficient is leveraged as an evaluation metric to measure the overlapping regions between the model predictions

Table 6.1: Comparison of SOTA approaches on the five different testing datasets. (*: $p < 0.01$, with Paired Wilcoxon signed-rank test to all baseline networks)

			Internal Testing FLARE					External Testing MSD	KiTS	LiTS	TCIA
Methods	#Params	FLOPs	Spleen	Kidney	Liver	Pancreas	Mean	Spleen	Kidney	Liver	Pancreas
nn-UNet (99)	31.2M	743.3G	0.971	0.966	0.976	0.792	0.926	0.917	0.829	0.935	0.739
TransBTS (218)	31.6M	110.4G	0.964	0.959	0.974	0.711	0.902	0.881	0.797	0.926	0.699
UNETR (74)	92.8M	82.6G	0.927	0.947	0.960	0.710	0.886	0.857	0.801	0.920	0.679
nnFormer (253)	149.3M	240.2G	0.973	0.960	0.975	0.717	0.906	0.880	0.774	0.927	0.690
SwinUNETR (73)	62.2M	328.4G	0.979	0.965	0.980	0.788	0.929	0.901	0.815	0.933	0.736
3D UX-Net (k=7) (121)	53.0M	639.4G	0.981	0.969	0.982	0.801	0.934	0.926	0.836	0.939	0.750
3D UX-Net (k=21) (121)	65.9M	757.6G	0.980	0.968	0.979	0.795	0.930	0.908	0.808	0.929	0.720
RepOptimizer (47)	65.8M	757.4G	0.981	0.969	0.981	0.822	0.937	0.913	0.833	0.934	0.746
3D RepUX-Net (Ours)	65.8M	757.4G	**0.984**	**0.970**	**0.983**	**0.837**	**0.944***	**0.932***	**0.847***	**0.949***	**0.779***

Table 6.2: Ablation studies with quantitative Comparison on Block Designs with/out frequency modeling using different optimizer

Optimizer	Main Branch	Para. Branch	BFR	Train Steps	Main LR	Para. LR	Mean Dice
SGD	$21 \times 21 \times 21$	$\times$	$\times$	40000	0.0003	$\times$	0.898
AdamW	$21 \times 21 \times 21$	$\times$	$\times$	40000	0.0001	$\times$	0.906
SGD	$21 \times 21 \times 21$	$3 \times 3 \times 3$	$\times$	40000	0.0003	0.0006	0.917
AdamW	$21 \times 21 \times 21$	$3 \times 3 \times 3$	$\times$	40000	0.0001	0.0001	0.929
AdamW	$21 \times 21 \times 21$	$\times$	✓	40000	0.0001	$\times$	**0.938**
SGD	$21 \times 21 \times 21$	$3 \times 3 \times 3$	$\times$	60000	0.0003	0.0006	0.930
AdamW	$21 \times 21 \times 21$	$3 \times 3 \times 3$	$\times$	60000	0.0001	0.0001	0.938
AdamW	$21 \times 21 \times 21$	$\times$	✓	60000	0.0001	$\times$	**0.944**

and the manual ground-truth labels.

6.6 Results

Different Scenarios Evaluations. Table 6.1 shows the result comparison of current SOTA networks on medical image segmentation in a volumetric setting. With our designed convolutional blocks as the encoder backbone, RepUX-Net demonstrates the best performance across all segmentation task with significant improvement in Dice score (FLARE: 0.934 to 0.944, AMOS: 0.891 to 0.902). Furthermore, RepUX-Net demonstrates the best generalizability consistently with a significant boost in performance across 4 different external datasets (MSD: 0.926 to 0.932, KiTS: 0.836 to 0.847, LiTS: 0.939 to 0.949, TCIA: 0.750 to 0.779). For transfer learning scenario, the performance of RepUX-Net significantly outperforms the current SOTA networks with mean Dice of 0.911 (1.22% enhancement), as shown in Table 6.2. RepUX-Net demonstrates its capabilities across the generalizability of unseen datasets and transfer learning ability. The qualitative representations (in Figure 6.3) further provides additional confidence of the quality improvement in segmentation predictions with RepUX-Net.

Ablation studies with block designs & optimizers. With the plain convolution design, a mean dice score of 0.906 is demonstrated with AdamW optimizer and perform slightly better than that with SGD. With the addi-

Table 6.3: Evaluations on the AMOS testing split in different scenarios.(*: $p < 0.01$, with Paired Wilcoxon signed-rank test to all baseline networks)

Train From Scratch Scenario																
Methods	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg
nn-UNet	0.951	0.919	0.930	0.845	0.797	0.975	0.863	0.941	0.898	0.813	0.730	0.677	0.772	0.797	0.815	0.850
TransBTS	0.930	0.921	0.909	0.798	0.722	0.966	0.801	0.900	0.820	0.702	0.641	0.550	0.684	0.730	0.679	0.783
UNETR	0.925	0.923	0.903	0.777	0.701	0.964	0.759	0.887	0.821	0.687	0.688	0.543	0.629	0.710	0.707	0.740
nnFormer	0.932	0.928	0.914	0.831	0.743	0.968	0.820	0.905	0.838	0.725	0.678	0.578	0.677	0.737	0.596	0.785
SwinUNETR	0.956	0.957	0.949	0.891	0.820	0.978	0.880	0.939	0.894	0.818	0.800	0.730	0.803	0.849	0.819	0.871
3D UX-Net (k=7)	0.966	0.959	0.951	0.903	0.833	0.980	0.910	0.950	0.913	0.830	0.805	0.756	**0.846**	0.897	0.863	0.890
3D UX-Net (k=21)	0.963	0.959	0.953	**0.921**	0.848	0.981	0.903	0.953	0.910	0.828	0.815	0.754	0.824	0.900	0.878	0.891
RepOptimizer	0.968	**0.964**	0.953	0.903	0.857	0.981	0.915	0.950	0.915	0.826	0.802	0.756	0.813	0.906	0.867	0.892
RepUX-Net (Ours)	**0.972**	0.963	**0.964**	0.911	**0.861**	**0.982**	**0.921**	**0.956**	**0.924**	**0.837**	**0.818**	**0.777**	0.831	**0.916**	**0.879**	**0.902***
Transfer Learning Scenario																
Methods	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg
nn-UNet	0.965	0.959	0.951	0.889	0.820	0.980	0.890	0.948	0.901	0.821	0.785	0.739	0.806	0.869	0.839	0.878
TransBTS	0.885	0.931	0.916	0.817	0.744	0.969	0.837	0.914	0.855	0.724	0.630	0.566	0.704	0.741	0.650	0.792
UNETR	0.926	0.936	0.918	0.785	0.702	0.969	0.788	0.893	0.828	0.732	0.717	0.554	0.658	0.683	0.722	0.762
nnFormer	0.935	0.904	0.887	0.836	0.712	0.964	0.798	0.901	0.821	0.734	0.665	0.587	0.641	0.744	0.714	0.790
SwinUNETR	0.959	0.960	0.949	0.894	0.827	0.979	0.899	0.944	0.899	0.828	0.791	0.745	0.817	0.875	0.841	0.880
3D UX-Net (k=7)	0.970	0.967	0.961	0.923	0.832	0.984	0.920	0.951	0.914	0.856	0.825	0.739	0.853	0.906	0.876	0.900
3D UX-Net (k=21)	0.969	0.965	0.962	0.910	0.824	0.982	0.918	0.949	0.915	0.850	0.823	0.740	0.843	0.905	0.877	0.898
RepOptimizer	0.967	0.967	0.957	0.908	0.847	0.983	0.913	0.945	0.914	0.838	0.825	0.780	0.836	0.915	0.864	0.897
RepUX-Net	**0.973**	**0.968**	**0.965**	**0.933**	**0.865**	**0.985**	**0.930**	**0.960**	**0.923**	**0.859**	**0.829**	**0.793**	**0.869**	**0.918**	**0.891**	**0.911***

tional design of a parallel small kernel branch, the segmentation performance significantly improved (SGD: 0.898 to 0.917, AdamW: 0.906 to 0.929) with the optimized parallel branch LR using SR. The performance is further enhanced (SGD: 0.917 to 0.930, AdamW: 0.929 to 0.937) without being saturated with the increase of the training steps. By adapting BFR, the segmentation performance outperforms the parallel branch design significantly with a Dice score of 0.944.

Effectiveness on Different Frequency Distribution. From Figure 6.4 in SM, RepUX-Net demonstrates the best performance when $\alpha = 1$, while comparable performance is demonstrated in both $\alpha = 0.5$ and $\alpha = 8$. A possible family of Bayesian distributions (different shapes) may need to further optimize the learning convergence of kernels across each channel.

Limitations. The shape of the generated Bayesian distribution is fixed across all kernel weights with an unlearnable distance function. Each channel in kernels is expected to extract variable features with different distributions. Exploring different families of distributions to rescale the element-wise convergence in kernels will be our potential future direction.

6.7 Conclusion

We introduce RepUX-Net, the first 3D CNN adapting extreme large kernel convolution in encoder network for medical image segmentation. We propose to model the spatial frequency in the human visual system as a reciprocal function, which generates a Bayesian prior to rescale the learning convergence of each element in kernel weights. By introducing the frequency-guided importance during training, RepUX-Net outperforms

Table 6.4: Complete overview of six public MICCAI challenge datasets

Challenge	FLARE	AMOS	MSD	KiTS	LiTS	TCIA
Imaging Modality	Multi-Contrast CT	Multi-Contrast CT	Venous CT	Arterial CT	Venous CT	Venous CT
Anatomical Region	Abdomen	Abdomen	Spleen	Kidney	Liver	Pancreas
Sample Size	361	200	41	300	131	89
Anatomical Label	Spleen, Kidney, Liver, Pancreas	Spleen, Left & Right Kidney, Gall Bladder, Esophagus, Liver, Stomach, Aorta, Inferior Vena Cava (IVC) Pancreas, Left & Right Adrenal Gland (AG), Duodenum	Spleen	Kidney, Tumor	Liver, Tumor	Pancreas
Data Splits	5-Fold Cross-Validation (Internal) Train: 272 / Validation: 69 / Test: 20	1-Fold (Internal) Train: 160 / Validation: 20 / Test: 20	All (External) Test: 41	Test: 300	Test: 131	Test: 89

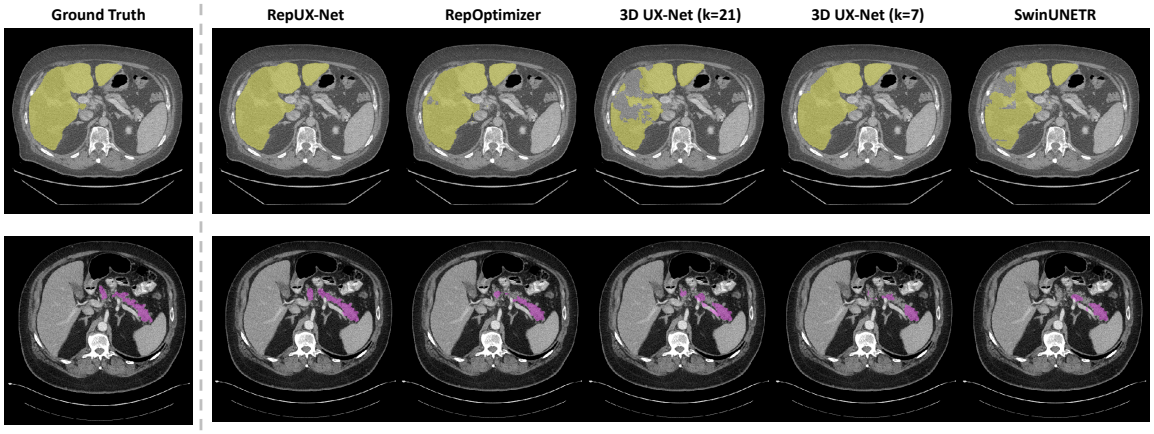


Figure 6.3: Qualitative Representations of organ segmentation in LiTS and TCIA datasets

current SOTA networks on six challenging public datasets via both direct training and transfer learning scenarios.

6.8 Supplementary Material

6.8.1 Derivation of Variable Convergence in Multi-Branch Design

The parallel structural design is referred to the CSLA block. Each branch only comprises on differentiable linear operator with trainable parameters (e.g., Convolution (Conv), Fully-Connected (FC) layer, scaling layer) and no training-time non-linearity. We begin with a simple case where the CSLA block has two parallel Conv kernels with same dimensions in kernel weights by padding and scaled by constant values. Let α_L, α_S and W_L, W_S be the constant scalars and the weights of two Conv kernels (Large & Small), and X and Y be the input and the output features. The computation flow of the CSLA block is formulated as following:

$$Y_{CSLA} = \alpha_L(X * W_L) + \alpha_S(X * W_S), \tag{6.5}$$

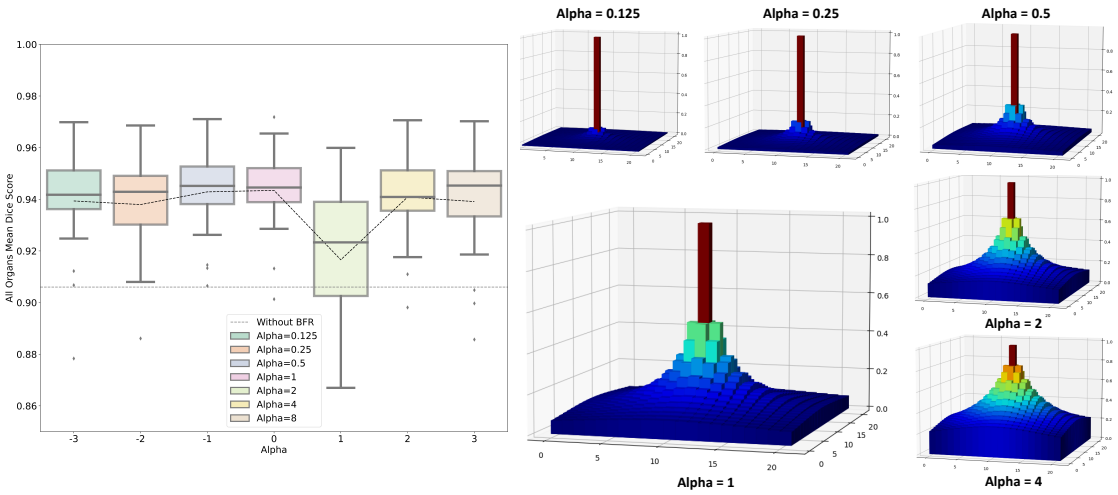


Figure 6.4: Quantitative evaluation of the ablation study with different frequency distribution.

where $*$ denotes convolution. To optimize the gradient in parallel structure as a Single Operator (SO), we first derive the gradient flow scenario in a single operator structure, which is parameterized by $W^{'}$ as follows:

$$Y_{SO} = X * W^{'} \tag{6.6}$$

Our goal is to ensure generating same outputs in both multi-branch and single operator setting $Y_{CSLA} = Y_{SO}$ during training. Since only linear operations are performed within branches, we derive the kernel weights relationship as follows:

$$W^{'} = \alpha_L W_{L(i)} + \alpha_S W_{S(i)}. \tag{6.7}$$

With the above theoretical relationship in kernel weights, we apply the stochastic gradient descent rule and update the parallel branches gradient as follows:

$$W^{'}_{i+1} = W^{'}_{i} - \lambda \frac{\partial \mathscr{L}}{\partial W^{'}_{i}} \tag{6.8}$$

$$\alpha_L W_{L(i+1)} + \alpha_S W_{S(i+1)} = \alpha_L W_{L(i)} + \alpha_S W_{S(i)} - \lambda (\alpha_L \frac{\partial \mathscr{L}}{\partial W_{L_i}} + \alpha_S \frac{\partial \mathscr{L}}{\partial W_{S_i}}) \tag{6.9}$$

where $\mathscr{L}$ is the differentiable loss function, i is the index number of training iterations and λ is the Learning Rate (LR). We further expand equation 5 and observe that each conv branch can be optimized with different convergence rate by adjusting the corresponding LR as follows:

$$\alpha_L W_{L(i+1)} + \alpha_S W_{S(i+1)} = \alpha_L W_{L(i)} - \lambda_L \alpha_L \frac{\partial \mathscr{L}}{\partial W_{L_i}} + \alpha_S W_{S(i)} - \lambda_S \alpha_S \frac{\partial \mathscr{L}}{\partial W_{S_i}}, \tag{6.10}$$

where λ_L and λ_S are the LR for the large kernel branch and small kernel branch respectively. After training, the small kernel parameters are merged onto the central point of the large kernels, which is equivalent to enhance the learning convergence in locality with single operator setting.

CHAPTER 7

Semantic-Aware Contrastive Learning for Multi-object Medical Image Segmentation

7.1 Overview

[1]Medical image segmentation, or computing voxel-wise semantic masks, is a fundamental yet challenging task in medical imaging domain. To increase the ability of encoder-decoder neural networks to perform this task across large clinical cohorts, contrastive learning provides an opportunity to stabilize model initialization and enhance downstream tasks performance without labels. However, multiple target objects (with different semantic meanings) with different contrast level may exist in a single image, which poses a problem for adapting traditional contrastive learning methods from prevalent "image-level classification" to "pixel-level segmentation". In this paper, we propose a simple semantic-aware contrastive learning approach leveraging attention masks and image-wise labels to advance multi-object semantic segmentation. Briefly, we embed different semantic objects to different clusters rather than the traditional image-level embeddings. We evaluate our proposed method on a multi-organ medical image segmentation task with both in-house data and MICCAI Challenge 2015 BTCV datasets. Compared with current state-of-the-art training strategies, our proposed pipeline yields a substantial improvement of 5.53% and 6.09% on Dice score for both medical image segmentation cohorts respectively (p-value$<$0.01). The performance of the proposed method is further assessed on external medical image cohort via MICCAI Challenge FLARE 2021 dataset, and achieves a substantial improvement from Dice 0.922 to 0.933 (p-value$<$0.01). The code is available at: https://github.com/MASILab/DCC_CL

7.2 Introduction

Contrastive learning is a variant of self-supervised learning (SSL) that has advanced to learn useful representation for an image classification task (33). Traditional contrastive learning approach consists of two primary concepts: 1) the learning process pulls the target image (anchor) and a matching sample close to each other as a "positive pair" in the embedding space, and 2) the learning process pushes the anchor from non-matching samples away from each other as "negative pairs" in the embedding space. Data augmentation is used to generate the positive samples from a training sample, while the negative pairs are formed from the remaining samples of non-matching objects. Previous studies demonstrate the advantages of contrastive learning in image-level classification tasks (38; 238; 36). We posit that contrastive learning can also leverage the

[1]In submission at: Lee, Ho Hin, et al., "Semantic-Aware Contrastive Learning for Multi-object Medical Image Segmentation.", IEEE Journal of Biomedical and Health Informatics, 2023 (133)

capability of sub-image-level feature encoding, to advance pixel-level segmentation tasks. However, some gaps need to be filled to achieve the latter goal, especially for multi-organ segmentation tasks in medical imaging (24; 144; 248). For example, multiple semantic objects may exist in medical images (e.g., abdomen, organs, brain tissues), while each element in the convoluted/downsampled feature may correlate to multiple objects. Thus, it is difficult to align the object-wise semantics with the learned representation in the latent space for multi-organ segmentation as the downstream task. moreover, it is challenging to provide the semantic explainability, which classifies the quality of the learned feature for multi-object segmentation in medical images.

In this work, we propose a semantic-aware attention-guided contrastive learning (AGCL) framework to advance multi-object medical image segmentation with contrastive learning. We integrate object-corresponding attention maps as additional input channels to adapt representations into corresponding semantic embeddings (as shown in Figure 7.1). To further stabilize the latent space, we propose a multi-class conditional contrastive loss that increase the arbitrary number of positive pairs within the same sub-classes for contrastive learning. Instead of leveraging pixel-wise label, radiological conditions such as modality and organ information, are provided as image-wise multi-class label to constrain the normalized embedding. Figure 7.2 provides a visual explanation of our proposed framework. Our proposed contrastive learning strategy AGCL is evaluated with three medical imaging datasets (two public contrast-enhanced CT dataset (117), in-house non-contrast dataset). The results demonstrate that consistent improvements are achieved on both ResNet-50 and ResNet-101 architectures (79). Our main contributions are summarized as below: 1. We propose a semantic-aware contrastive learning framework to advance multi-object pixel-level semantic segmentation.

2. We propose a multi-conditional contrastive loss to integrate meta information as an additional constraint for classifying representations into sub-class embeddings.

3. We demonstrate that the proposed AGCL generalizes the CT contrast phase variation in each organ and significantly boosts the segmentation performance.

7.3 Related Works

Contrastive Learning: Self-supervised representation learning approaches have recently been proposed to learn useful representation from unlabeled data. Some approaches propose learning embeddings directly in lower-dimensional representation spaces instead of computing a pixel-wise predictive loss (243). Self-distillation with a teacher-student network is further proposed to enrich the semantic correspondence using pseudo-label predictions (22), while the masked autoencoder provides an alternative to learn the spatial feature correspondence with the image reconstruction task (76). Contrastive learning is one such state-of-the-art

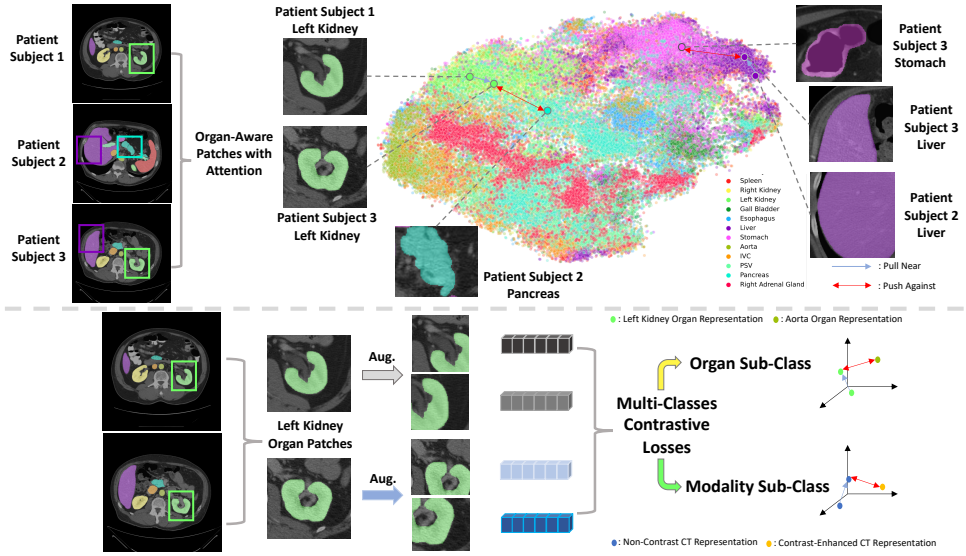


Figure 7.1: With multiple organs located in a single image, organ attention maps guide their representations into corresponding embeddings and adapt contrastive learning for multi-object segmentation. Categorical information can be used for supervisory context to constrain the separation of clusters (grey arrow: pull the matching representations together, red arrow: push the non-matching representations apart).

methods for self-supervised learning to model the semantic-wise relationships in the latent space (247; 33). It employs a loss function to pull latent representations closer together for positive pairs, while pushing them apart for negative pairs. Maximizing mutual information between embeddings has also been proposed as an alternative to extract the similar information between targets (7). Adapting with memory bank and momentum contrastive approaches have been proposed to increase the batch sizes and generate more dissimilar pairs in a minibatch for contrastive learning (154). Additionally, to constrain and stabilize the embedding spaces, class label information has been added to provide additional supervision to stabilize contrastive learning process (108).

However, most of the prior works in contrastive learning focused on improving the "image-wise classification", while relatively fewer methods have been proposed for the "pixel-wise segmentation". Pixel-wise contrastive loss was proposed to adapt the representation from the ground-truth label information (248), while dense contrastive loss was also proposed to minimize the discrepancy of image-level prediction and pixel-level prediction (220). Furthermore, one-stage contrastive learning framework was proposed to enforce the pixel embeddings belonging to a same semantic class to be more similar than embeddings from different classes (219; 90; 1). In the medical domain, the contrastive learning framework was extended to leverage the structural similarity and learn the distinctive representations of local regions without using pixel-wise ground truth labels (24) and image-wise labels (241). Similarly, limited ground-truth labels were adapted into the pretraining step for contrastive learning and enhanced the segmentation performance (92; 147). However,

such method typically needs pixel-wise segmentation labels. Therefore, our proposed method identifies the semantic-aware regions with one-hot attentional guidance and leverages multi-class image-level labels to define region-bounded representations as an arbitrary number of positive pairs for contrastive learning, without using pixel-wise labels.

Medical Image Segmentation & Multi-Organ Segmentation: Fully-supervised deep learning methods have been developed to enhance both the segmentation performance and the generalizability across different datasets (39; 65; 257; 95). However, the supervised learning strategies are limited to the quality of pixel/voxel-wise labels and the resolution of volumes (62). Thus, hierarchical approaches and patch-wise approaches have been proposed to perform segmentation across scales and resolutions (182; 259; 200). Another study further enhanced the segmentation accuracy with the statistical fusion (222). Apart from multi-view attention, shape-aware network was proposed to consistently smoothen the label prediction by learning the signed distance function as additional constraints (233). Furthermore, RAP-Net was proposed to leverage one-hot shape-aware mappings to provide additional localization context as additional input channel and refine the segmentation mapping hierarchically (125). nn-UNet further enhanced the generalizability with self-configuring structure to diversely predict segmentation for multi-modality imaging (99). Meanwhile, vision transformer was introduced as the encoder network to extract attention features with large receptive field for robust segmentation (74). On the other hand, partially-supervised, semi-supervised and self-supervised learning have also been explored to adapt unlabeled data in the medical imaging domain. Multiple single organ-labeled datasets are used to provide structural prior knowledge during the training with multiple organ-labeled dataset to enhance multi-organ segmentation performance (255). A quality assurance module have been proposed to adapt the segmentation quality as the supervision from unlabeled data (130). Pretext tasks such as colorization, deformation and image rotation, have been used as pre-training features to initialize the segmentation networks (258). Self-supervised context has been explored by predicting the relative patch location and the degree of rotation (8; 260). Contrastive learning has been used to extract global and local representations for domain-specific MRI images (24). A contrastive predictive network has been used to summarize the latent vectors in a minibatch and predict the latent representation of adjacent patches (195).

7.4 Method

We present our co-training approach AGCL that integrates one-hot organ attention into contrastive learning by adapting radiological context labels (modality and organ) to classify representations into sub-classes embeddings, as presented in Figure 7.3.

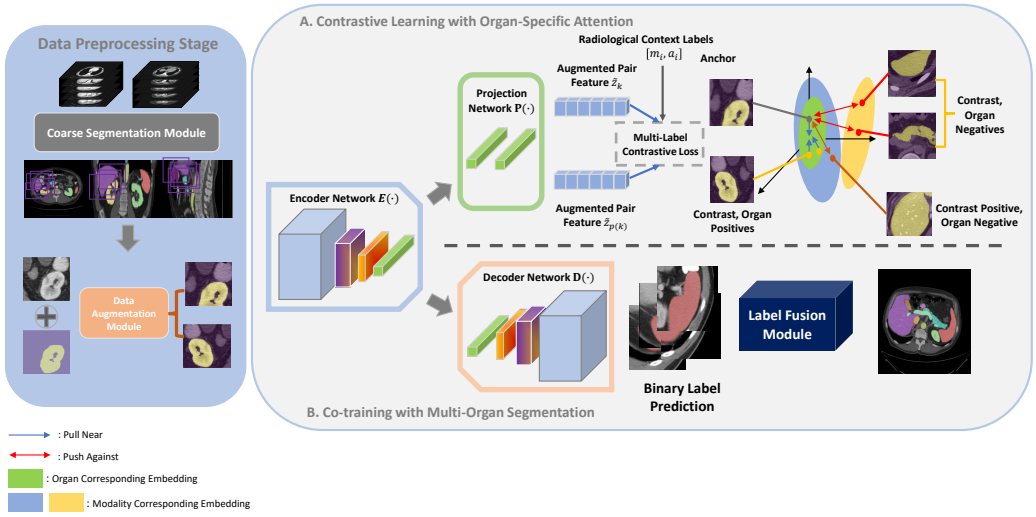


Figure 7.2: A 2D/3D segmentation pipeline (2D for natural image, 3D for medical image) is used to generate attention map for organ localization. 2D organ-corresponding query patches are randomly extracted and concatenated with the regional attention maps as an additional channel to guide embeddings of the organ targets. Data augmented pairs of the attention queries are constrained into corresponding radiological embeddings (such as organs and modalities) with additional label supervision in the proposed contrastive loss. The encoder is co-trained with decoder to generate refined segmentation with label fusion.

7.4.1 Hierarchical Coarse Segmentation

The input data of the entire pipeline is a multi-contrast 3D image volume $V_i = \{X_i, Y_i\}_{i=1,\ldots,L}$, where L is the number of all imaging samples, X is the volumetric image and Y is the corresponding multi-organ label. The corresponding outcomes of the preprocessing stage are coarse segmentation masks (attention maps) $A_i = RAP(X_i)$ from a hierarchical segmentation network $RAP(\cdot)$ (125). We define $A \in R^{H \times W \times C}$, where H and W denote as the axial dimension of the image, and C denotes as the number of label classes. The coarse segmentation network RAP-Net consists of two hierarchical stages: 1) low-resolution whole volume segmentation and 2) organ-specific patch segmentation refinement. The low-resolution model generates a rough segmentation map and provide anatomical context as additional channel input to refine the segmentation in patch-wise setting as the second step. Both low-resolution model and patch-wise model are trained in supervised setting with 5-fold cross-validations.

7.4.2 Data Preprocessing

The goal of the data preprocessing step is to randomly sample 3D training patches for downstream contrastive learning. In our design, we utilize the attention maps (1) as spatial restrictions of the organ-specific sampling process, and (2) highlight the current organ of interest to define semantic-wise embeddings for segmentation

refinement. Briefly, organ-specific patches $p_i = \{x_{C,i}, y_{C,i}, s_{C,i}\}_{i=1,\ldots,N}$ are randomly sampled within each organ class C in attention maps. The center point is randomly sampled from attention maps to crop the region of interest (ROI), N denotes as the total number of query patches, $x_{C,i}$ is the organ-corresponding image patch, $y_{C,i}$ is the binary ground-truth label patch, and $s_{C,i}$ is the coarse organ-specific attention map from A_i in binary setting. As the significant difference between $y_{C,i}$ and $s_{C,i}$ is the variation of segmentation quality, the model trained aims to refine the segmentation with the prior knowledge of $s_{C,i}$ as additional input channel. asaAs a standard process in data augmentation, random cropping, rotation (-30 to 30 degrees), scaling has been applied to augment the size of training samples.

7.4.3 Contrastive learning with Organ-Specific Attention

After generating augmented image pairs, pairwise images are then used as the inputs for contrastive learning. Specifically, a convolutional encoder network $E(\cdot)$ is used to extract high dimensional features. We further project each high-level feature map into 1D vector $\widetilde{z}_i$ using a multi-layer perceptron network $P(\cdot)$, $\widetilde{z}_i = P(E(\widetilde{a}_i)), \widetilde{z}_i \in R^{O_E}$ (pink box in Figure 7.2), where O_E is the size of the output vector. Then, the standard self-supervised contrastive loss (SSCL) (33) can be defined as the following:

$$\mathscr{L}_{self} = -\sum_{k=1}^{2N} \log \frac{\exp(\widetilde{z}_k \cdot \widetilde{z}_{p(k)} / \mathscr{T})}{\sum_{j \in J(k)} \exp(\widetilde{z}_k \cdot \widetilde{z}_j / \mathscr{T})} \tag{7.1}$$

where T is a hyperparameter indicating temperature scaling to control the radius weighting on the positive/negative pairs. Both k and $p(k)$ represent the index of the anchor sample and the corresponding positive sample, respectively. $J(k)$ represents the number of remaining negative samples. To incorporate the attention with modality and organ semantic meanings, we extend SSCL to adapt an arbitrary number of positive pairs by introducing multi-class image-level labels into contrastive loss. Here, modalities indicate the different contrast types in CT (we have utilized both contrast-enhanced and non-contrast CT scans in the training set). As the organ-specific attention only provides one-hot voxel-wise context to preserve organ regions, the multi-class labels represent different modalities/organs for the learned representations under the attention regions. It provides flexibility to further constrain the representations into semantic-aware clusters, which is conditional to the image-level label. In each batch, pairwise patches with the same organ and modality label are defined as a positive pair, while the remaining pairs are specified as negative pairs. With such positive-negative pairs definition, we further extend the contrastive loss with conditional constraints as follows:

$$\mathscr{L}_{MT} = \sum_{k=1}^{2N} \frac{-1}{|L(k)|} \sum_{l \in L(k)} \log \frac{\exp(\widetilde{z}_k \cdot \widetilde{z}_l / \mathscr{T})}{\sum_{j \in J(k)} \exp(\widetilde{z}_k \cdot \widetilde{z}_j / \mathscr{T})} \tag{7.2}$$

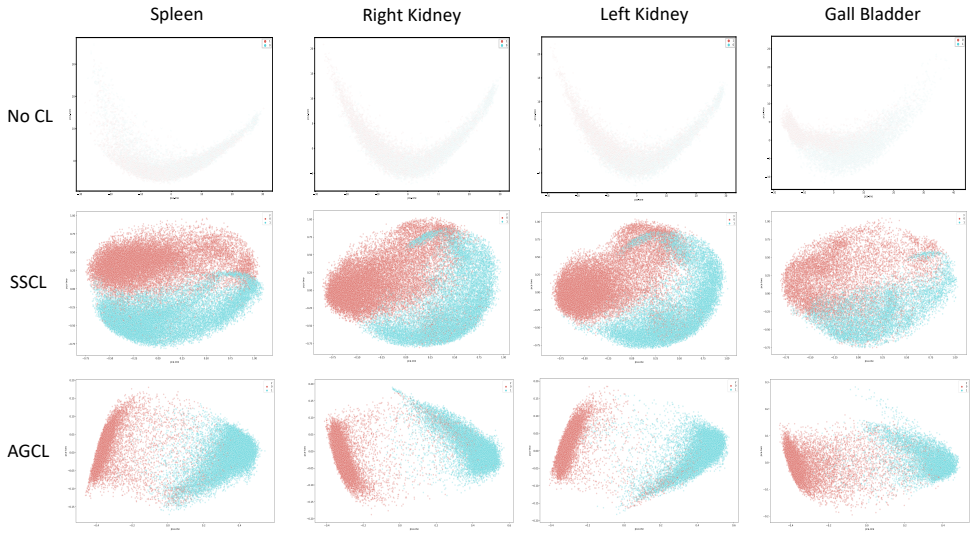


Figure 7.3: The latent distributions of four randomly selected organs using principal component analysis (PCA) are plotted with their corresponding modality (Blue: contrast-enhanced phase CT, red: non-contrast phase CT). The first two components are plotted as a visualization. With AGCL, the organ representation can be well separated into specific modal clusters.

where $L(k) \equiv \{l \in J(k) : m_k = m_l, o_p = o_l\}$, m and o denote as the corresponding modality and organ label, respectively. l and $\hat{z}_l$ are the index number and the projected feature representation of the corresponding positive sample with same organ and modality label. The feature vector output with 256 channels is directly used to compute the contrastive loss of modality and organ class respectively. By classifying learned representations into multi-classes embeddings, the model is initially learned the attention-bounded representations with semantic meanings, which hypothesize to be beneficial for downstream segmentation refinement task.

7.4.4 Co-training with Multi-Organ Segmentation

The ultimate goal of our framework is to achieve a robust patch-wise contrastive learning without using pixel-wise labels, which benefits for downstream segmentation task. The native two-stage strategy is to train both contrastive loss and downstream segmentation loss independently. Here, we attempt to have a co-training strategy, by training the contrastive loss and segmentation tasks simultaneously. The encoder network is followed with an atrous spatial pyramid pooling (ASPP) module as the decoder network to resample the bottleneck feature with multiple effect Field Of Views (FOV) (31). The DeepLabV3+ is employed as the segmentation part with the shared encoder structure for contrastive learning (31). The distinctiveness of adapting ASPP is to obtain multi-scale features during upsampling. One 1×1 convolution and three 3×3 convolution layers with different dilation rate (e.g., 6, 12, 18) are leveraged. With the increased number of

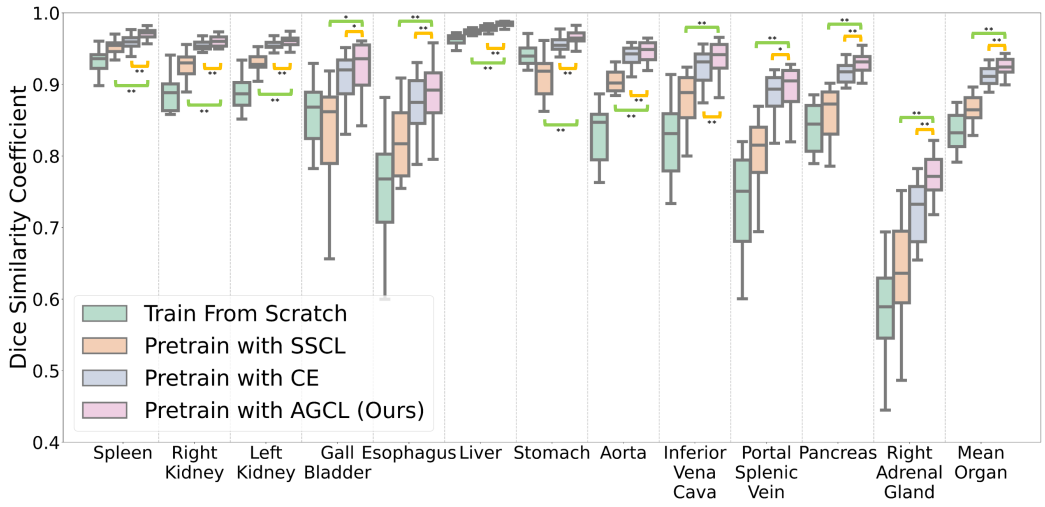


Figure 7.4: Comparison of different supervised / self-supervised pre-training strategies using multi-class labels for multi-organ segmentation. AGCL outperforms the current state-of-the-art pre-training methods with SSCL and classification pre-training with CE across all organs. (*: $p < 0.05$, **: $p < 0.01$, with Wilcoxon signed-rank test)

dilation rate, kernel stride is constrained while a larger FOV is accomplished without increasing the number of model parameters. Furthermore, image pooling is also performed in parallel to extract the global features. Features from different FOV are finally concatenated. The channel-wise features are mixed using a 1×1 convolution layer before passing through the final layer for high-resolution prediction. The rationale of such design is to adapt the multi-view behavior and search the optimal tradeoff between the localized features (small FOV) and the global-assimilated features (large FOV). Dice loss is used to compute the predicted output with the ground truth label in binary setting for the co-training segmentation task. After computing all organ-specific patches predictions, we fuse the organ-wise patches according to the center point recorded in the data preprocessing stage. We employ majority voting (254) as the label fusion module to fuse predictions into multi-organ labels.

7.5 Experiments

Datasets: To evaluate our proposed learning approach, one in-house research cohorts and two publicly available datasets in medical imaging are used with multi-organ segmentation as the downstream task.

MICCAI 2015 Challenge Beyond The Cranial Vault (BTCV) dataset is comprised of 100 de-identified unpaired 3D contrast-enhanced CT scans with 7,968 axial slices in total. 20 scans are publicly available for

Table 7.1: Comparison of the fully-supervised, unsupervised, semi-supervised and partially supervised state-of-the-art methods on the 2015 MICCAI BTCV challenge leaderboard. (We show 8 main organs Dice scores due to limited space, $*$: fully-supervised approach, $\circ$: unsupervised approach, $\triangle$: partially supervised approach, $\diamond$: contrastive learning approach.)

Method	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Aorta	IVC	Average Dice	Mean Surface Distance	Hausdorff Distance
Heinrich et al. $\circ$(84)	0.920	0.894	0.915	0.604	0.692	0.948	0.857	0.828	0.790	2.262	25.504
Cicek et al. $*$(39)	0.906	0.857	0.899	0.644	0.684	0.937	0.886	0.808	0.784	2.339	15.928
Roth et al. $*$(182)	0.935	0.887	0.944	0.780	0.712	0.953	0.880	0.804	0.816	2.018	17.982
Pawlowski et al. $*$(170)	0.939	0.895	0.915	0.711	0.743	0.962	0.891	0.826	0.815	1.861	62.872
Zhu et al. $*$(259)	0.935	0.886	0.944	0.764	0.714	0.942	0.879	0.803	0.814	1.692	18.201
Lee et al. $*$(125)	0.959	0.920	0.945	0.768	0.783	0.962	0.910	0.847	0.842	1.501	16.433
Isensee et al. $*$(99)	0.956	0.923	0.940	0.760	0.764	0.965	0.905	0.850	0.839	1.523	16.201
Hat. et al. $*$(74)	0.959	0.912	0.940	0.724	0.746	0.968	0.905	0.840	0.836	1.602	16.355
Zhou et al. $\triangle$(255)	0.968	0.920	0.953	0.729	0.790	0.974	0.925	0.847	0.850	1.450	18.468
Chaitanya et al. $\diamond$(24)	0.956	0.935	0.946	0.920	0.854	0.970	0.915	0.893	0.874	1.236	15.281
Alonso et al. $\diamond$(1)	0.954	0.933	0.932	0.903	0.858	0.973	0.918	0.904	0.890	1.291	15.032
Wang et al. $\diamond$(220)	0.963	0.939	0.900	0.815	0.838	0.976	0.922	0.907	0.882	1.303	14.759
Khosla et al. $\diamond$(108)	0.959	0.939	0.947	**0.932**	0.867	0.978	0.922	0.911	0.907	0.978	14.136
Wang et al. $\diamond$(219)	0.966	0.942	0.955	0.886	0.860	0.975	0.930	0.908	0.913	0.991	13.785
Ours (SSCL w/o Attention)	0.940	0.909	0.918	0.740	0.790	0.962	0.890	0.854	0.838	1.967	17.773
Ours (SSCL w/ GT Attention)	0.947	0.910	0.928	0.805	0.815	0.967	0.890	0.861	0.858	2.013	18.101
Ours (SSCL Pretraining)	0.953	0.922	0.930	0.830	0.822	0.972	0.899	0.874	0.863	1.899	17.073
Ours (SSCL Co-training)	0.957	0.930	0.935	0.843	0.827	0.974	0.895	0.869	0.869	1.803	16.714
Ours (AGCL w/o Attention)	0.950	0.933	0.935	0.803	0.805	0.967	0.903	0.862	0.852	1.923	17.215
Ours (AGCL w/ GT Attention)	0.962	0.937	0.940	0.857	0.846	0.973	0.915	0.890	0.891	1.312	14.852
Ours (AGCL Pretraining)	0.971	0.955	**0.963**	0.910	0.886	0.984	**0.941**	**0.932**	0.923	0.932	**13.024**
Ours (AGCL Co-training)	**0.975**	**0.958**	0.960	0.914	**0.890**	**0.985**	0.937	0.920	**0.926**	**0.927**	13.102

the testing phase in the MICCAI 2015 BTCV challenge. All CT scans are in portal venous phase. Peak enhancement of contrast is observed in several organs, such as liver, kidney, spleen, and portal splenic vein. For each scan, 12 organ anatomical structures are well-annotated, including spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava (IVC), portal splenic vein (PSV), pancreas and right adrenal gland. Each volume consists of $47 \sim 133$ slices of 512×512 pixels, with resolution of $([0.54 \sim 0.98] \times [0.54 \sim 0.98] \times [2.5 \sim 7.0])mm^3$.

Non-contrast clinical cohort is retrieved in de-identified form from ImageVU database of Vanderbilt University Medical Center. It consists of 56 unpaired 3D CT scans with 3,687 axial slices and expert-refined annotations for the same 12 organs as the MICCAI 2015 BTCV challenge dataset. All volumetric scans are generated without contrast enhancement procedures. Each volume consists of $49 \sim 174$ slices of 512×512 pixels, with resolution of $([0.64 \sim 0.98] \times [0.64 \sim 0.98] \times [1.5 \sim 5.0])mm^3$.

MICCAI 2021 Challenge Fast and Low GPU memory Abdominal Organ Segmentation (FLARE) dataset leverage large scales of abdominal contrast-enhanced CT with 511 unpaired cases from 11 medical centers in multi-contrast phases (including both portal venous phase and non-contrast phase CTs). It consists

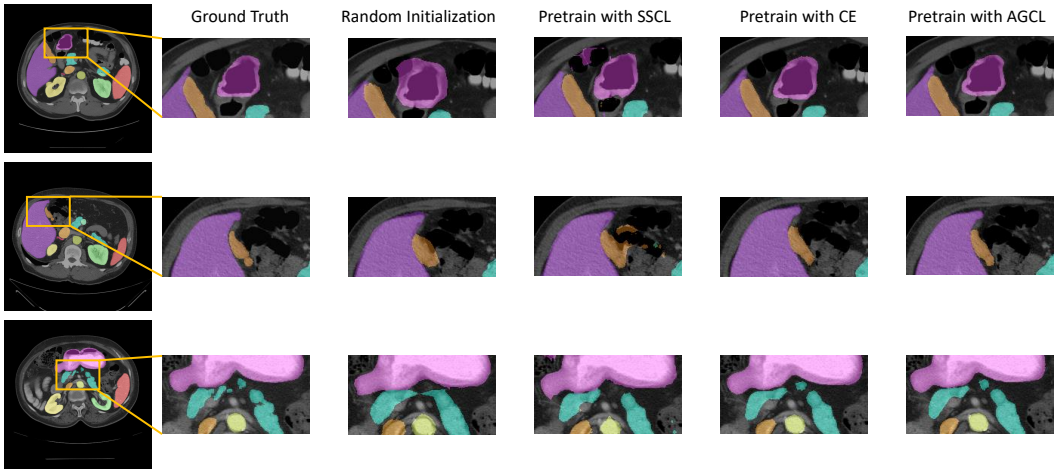


Figure 7.5: Qualitative representations of different pretraining strategies are demonstrated with ResNet-50 encoder backbone. Incremental improvement on segmentation quality is shown and AGCL demonstrates smooth boundaries and accurate morphological information between neighboring organs.

of 361 3D CT scans with four organ-specific labels including spleen, kidney, liver and pancreas. Each volume consists of $43 \sim 384$ slices of 512×512 pixels, with resolution of $([0.64 \sim 0.98] \times [0.64 \sim 0.98] \times [1.0 \sim 5.0])mm^3$.

Preprocessing: We apply the preprocessing steps as follows: (i) applying soft tissue windowing within the range of -175 to 250 Hu and perform intensity normalization of each 3D volume, v with min-max normalization: $(v - v_1)/(v_{99} - v_1)$, where v_p denote as the p^{th} intensity percentile in v, and (ii) apply volume-wise cropping in z-axis with body part regression algorithm to extract the abdominal region only for segmentation and ensure the similar field of view between scans (198).

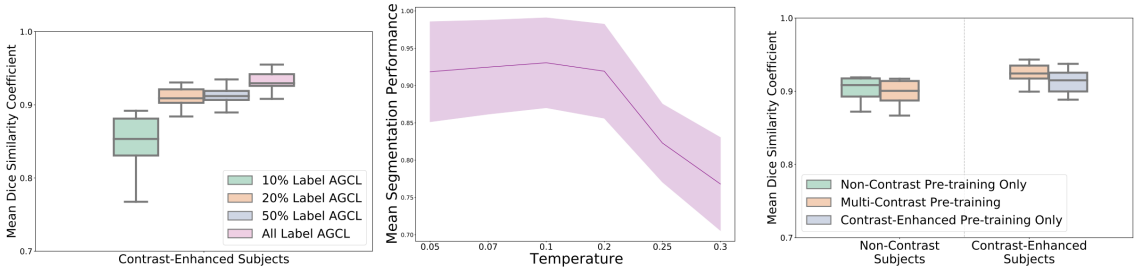
Network Training: Our proposed framework AGCL is trained with unpaired samples in our scenario, while it also allowed to train with paired samples. 5-fold cross-validation is performed for both contrast-enhanced phase and non-contrast phase CT (Training: 60 volumes (contrast-enhanced) and 44 volumes (non-contrast), validation: 20 volumes (contrast-enhanced) and 6 volumes (non-contrast), and testing: 20 volumes (contrast-enhanced) and 6 volumes (non-contrast)). For training the coarse segmentation network RAP-Net, we downsample all training volumes to a resolution of $2 \times 2 \times 6$ with the dimension of $168 \times 168 \times 64$. The low-resolution volumes are leveraged to train a low-resolution segmentation model with Adam optimizer using a batch size of 1 and learning rate of $1e-4$. We then use the coarse segmentation output to guide and extract organ-specific patches with dimension of $128 \times 128 \times 48$. The patch-wise segmentation refinement

model is trained with Adam optimizer using a batch size of 2 and learning rate of $1e-4$. For contrastive learning, we perform patients-level sampling and extract 30 2D query patches of each anatomical target in each axial slices of a subject scan. Such sampling strategy ensures that all patches are fully covered the organ-specific ROIs with significant variation of anatomical morphology. More than 400k patches with dimensions 128×128 are used and shuffled to train with stochastic gradient descent (SGD) optimizer for 5 epochs with a batch size of 4 and learning rate of 5×10^{-4}. We have evaluated the variation of the temperature parameter towards the segmentation performance and $T = 0.1$ achieved the best performances across all other temperature values. For segmentation task, the encoder's weight is frozen and only the decoder with ASPP module is trained for 10 epochs with Adam optimizer using a batch size of 4 and a learning rate of 10^{-4}. We used the validation set to choose the model with the highest mean Dice score for all semantic targets segmentation and perform inference as the quantitative representation on the testing set.

Experimental Setup: We evaluate the segmentation performance with Dice similarity coefficient on current state-of-the-art approach in contrastive learning and segmentation task for medical imaging domain, including the testing phase of the BTCV dataset, testing cohort of the non-contrast clinical cohort and the random sampled cohort from FLARE dataset. We further perform different pre-training strategies with the multi-class image-level label using different scenarios. Apart from learning image-wise embeddings with self-supervised setting, inspired by Khosla et al. (108), we introduce patch-wise multi-label (modality & organ) classification as the pretext task via the canonical cross-entropy (CE) loss in a fully-supervised setting. The learned representations are classified into label-corresponding clusters in AGCL. The CE loss is defined as following:

$$\mathscr{L}_{ce} = \sum_{i=1}^{C} y_i \log(p_i) \tag{7.3}$$

where y_i is the ground-truth multi-class label and p_i is the Softmax probability for the $i^{th}, i \in 1,\ldots,C$ classes. Apart from different pretext task strategies, we also perform ablation studies with the variation of hyperparameters such as temperature and the number of label used for pretraining, to investigate the optimal effect on fine-tuning segmentation task. For the encoder network, we evaluated with two common backbone architectures for segmentation in medical imaging domain: DeeplabV3+ with ResNet-50 and ResNet-101 encoder. The normalized activation of the final pooling layer with $D_E = 2048$ are used as the distinctive feature representation vector. More details of both network training and data preprocessing are demonstrated in the supplementary material.



(a) (b) (c)

Figure 7.6: **a)** The segmentation performance gradually improved with the additional quantities of image-level labels for AGCL. **b)** Ablation studies of temperature scaling the distance between positive/negative pairs demonstrates that the segmentation performance is best optimized when $T = 0.1$. **c)** Performance trade-off is demonstrated between non-contrast and contrast-enhanced CT with multi-modal training and the segmentation performance significantly improves in contrast-enhanced CT samples.

Table 7.2: Ablation studies of segmentation performance in various network backbones of the BTCV testing cohort.

Encoder	Pretrain	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A	Mean
ResNet50	×	0.932	0.877	0.887	0.860	0.761	0.962	0.941	0.832	0.815	0.735	0.833	0.587	0.840
ResNet50	SSCL	0.953	0.922	0.930	0.842	0.822	0.972	0.907	0.899	0.874	0.800	0.854	0.625	0.868
ResNet50	CE	0.959	0.948	0.957	0.890	0.868	0.978	0.956	0.935	0.919	0.884	0.903	0.725	0.905
ResNet50	AGCL	**0.971**	**0.955**	**0.963**	**0.910**	**0.886**	**0.984**	**0.965**	**0.941**	**0.932**	**0.893**	**0.917**	**0.769**	**0.923**
ResNet101	×	0.939	0.870	0.880	0.859	0.745	0.960	0.915	0.840	0.800	0.736	0.825	0.567	0.834
ResNet101	SSCL	0.950	0.928	0.935	0.805	0.792	0.969	0.900	0.905	0.877	0.800	0.846	0.602	0.868
ResNet101	CE	0.960	0.933	0.945	0.887	0.822	0.975	0.952	0.920	0.901	0.834	0.877	0.670	0.891
ResNet101	AGCL	**0.965**	**0.948**	**0.954**	**0.901**	**0.875**	**0.981**	**0.962**	**0.930**	**0.917**	**0.876**	**0.902**	**0.748**	**0.917**

7.5.1 Segmentation Performance

We first compare the proposed AGCL with a series of state-of-the-art approaches including 1) fully supervised approaches (training on ground-truth labeled data only), 2) a partially-supervised approach (training on one contrast phase dataset, and another with partial labels), and 3) contrastive learning approach for segmentation tasks. As shown in Table 7.1, the contrastive learning approach demonstrates significant improvement followed by the partial-supervision and full-supervision approaches. Chaitanya et al. integrates the SSCL across global to local scale and demonstrates significant improvement across organs. Khosla et al. provides an additional single class label to address the correspondence on embeddings, which outperforms all current approaches in supervised and self-supervised contrastive learning settings. By further adding multi-class labels as conditional constraints, AGCL achieves the best performance among all state-of-the-arts with a mean Dice score of 0.926. The additional gains demonstrate that our use of supplemental imaging information allows for recognition of more positive pairs with additional label constraints. To further evaluate the generalizability of our approach, we performed external evaluations on another public multi-organ labeled datasets

Table 7.3: Ablation studies of segmentation performance with adapting different constraints in contrastive loss.

Label Constraints	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A	Mean
Modality	0.965	0.948	0.958	0.890	0.865	0.975	0.950	0.930	0.920	0.879	0.895	0.732	0.910
Organ	0.968	0.947	0.959	0.893	0.872	0.979	0.957	0.935	**0.925**	0.886	0.903	0.744	0.914
Modality + Organ	**0.975**	**0.958**	**0.960**	**0.914**	**0.890**	**0.985**	**0.970**	**0.937**	0.920	**0.901**	**0.925**	**0.772**	**0.926**

FLARE for multi-organ segmentation. In Table 7.4, AGCL demonstrates substantial improvement on all organs segmentation when compared against the current all contrastive state-of-the-arts.

7.5.2 Ablation Study for AGCL

Comparing with first stage training approaches: To investigate the effect of using multi-class label for contrastive learning, we perform evaluation of different pretraining approaches with/out multi-class label: 1) training with self-supervised contrastive loss (SSCL), 2) training with cross-entropy (CE) loss as classification tasks, and 3) random initialization (RI) without any contrastive learning on both ResNet-50 and ResNet-101 encoder backbone. As shown in Table 7.2 and Figure 7.4, SSCL improves the segmentation performance over RI by 3.12%, which is expected because RI considers no constraint in the lower-dimensional space and relies on the decoder ability for downstream tasks. With the supervision of modality and anatomical information, the supervised image classification strategies significantly boost the segmentation performance by 5.93%. Pretraining with CE is to classify representations into corresponding embeddings related to the label given and representations in the same class is moved towards each other. Therefore, such improvement demonstrates that a good definition of the latent space in encoder can help address the corresponding representation for each semantic target and starts to achieve more favorable with the segmentation task. Eventually, AGCL surpasses CE by 1.32% in mean Dice and demonstrates the best performance across all pretraining strategies. Instead of constraining same class representations to move near only, our contrastive loss allows to push the representations out if they are not in the same class and provide a better definition on separating embeddings than pretraining with CE.

To further evaluate the segmentation using different contrastive learning approaches, the qualitative representation of the segmentation prediction with each training method is demonstrated on Figure 7.5 comparing with the ground truth label. With SSCL, the boundary of the segmentation is significantly smoother than that of with RI. However, we found that additional segmentation is performed near the neighboring structures. The similar intensity range and morphological appearance may lead to the instability of representation extraction from SSCL. Pretext task with CE demonstrates a significant improvement in label quality, while

Table 7.4: Comparison of the current contrastive state-of-the-art methods on FLARE dataset.

Method	Spleen	Kidney	Liver	Pancreas	Average Dice
Lee et al.(125)	0.956	0.903	0.954	0.730	0.885
Chai. et al.(23)	0.961	0.923	0.956	0.787	0.908
Wang et al.(220)	0.966	0.918	0.964	0.800	0.912
Khosla et al.(108)	0.963	0.918	0.966	0.830	0.919
Wang et al.(219)	0.968	0.940	0.964	0.811	0.922
Ours (SSCL)	0.960	0.910	0.960	0.756	0.896
Ours (AGCL)	**0.975**	**0.952**	**0.971**	**0.835**	**0.933**

the boundaries on particular organs (e.g. gall bladder) is not well preserved. With the additional constraints by AGCL, the boundary information between neighboring organs are clearly defined and the segmentation quality is comparable to the ground truth label.

Comparing with different constraints scenario in contrastive loss: We further perform evaluation in our proposed contrastive loss with class-wise label constraint: 1) modality-only label constraint, 2) organ-only label constraint, and 3) modality plus organ constraints with ResNet-50 encoder as the network backbone. As shown in Table 7.3, the overall superior performance is achieved when applying both modality and organ constraints.

Comparing with reduced label for AGCL: In Figure 7.6(a), we performed AGCL with the variation of label quantity and compare the segmentation performance by leveraging the amount of label information. We observed that model has the best performance with fully labeled input. A significant improvement is shown with 20% labels for AGCL comparing to that with 10% labels, while an improvement to a small extent is demonstrated by using 50% label for AGCL.

Comparing with temperature variability: We experimented with the variation of temperature to investigate the optimal effect towards the segmentation performance. Figure 7.6(b) demonstrates the effect of temperature on the multi-organ segmentation across all subjects in the BTCV testing dataset. We found that low temperature achieves better performance than high temperature, as the radius of the hypersphere defined in the latent space is inversely proportional to the temperature scaling, which increases the difficulty of finding positive samples with the decrease of radius.

Comparing with single/multiple modal contrastive learning The segmentation performance is evaluated with single modality and with multi-modality contrastive learning respectively. From Figure 7.6(c), a better segmentation performance for contrast-enhanced dataset is achieved by contrastive learning with multi-modality images. Interestingly, we observe that the segmentation performance of non-contrast imaging is improved to a small extent with non-contrast modal pre-training only.

7.5.3 Discussion & Limitations

In this work, we present a co-training framework that leverages organ attention into contrastive learning and defines representations into conditional embeddings with image-level labels only. We hypothesize that the conditional embeddings defined are beneficial to the segmentation task. By using organ attention as additional input channel, we can extract meaningful representation within the organ-specific regions, instead of randomly extracting representations that may affect by the neighboring organs.

It also allows to learn and define the organ-specific context into corresponding semantic categories. Apart from using organ attention, we further leverage the multi-class labels to constrain pairwise representations into sub-class embeddings. Instead of constraining contrastive loss in pixel-wise setting, we demonstrate that constraining the latent space with multiple image-level labels is also beneficial to enhance the segmentation performance for each organ of interest. From Table 7.1, we have shown that our proposed learning scheme outperforms the current contrastive learning state-of-the-art for multi-organ segmentation. Furthermore, Table 7.2 has shown the comparison of different pre-training strategies with/out multi-class image-level labels. It provides a better understanding about the contribution of our proposed contrastive loss in defining semantic-aware latent space for segmentation task.

Although AGCL tackles current challenges of integrating contrastive learning into multi-object segmentation, limitations still exist in the process of AGCL. One limitation is the dependency of the coarse segmentation quality. As 2D patches are extracted with the attention information in each slice, patches without corresponding organ regions may also be possible to extract due to inaccurate coarse segmentation. Incorrect label definition inputs may bring into contrastive learning process. Another limitation is performing contrastive learning in object-centric setting. We aim to innovate contrastive learning strategy with complete volume inputs for multi-object segmentation in our future work.

7.6 Conclusion

Performing robust multi-object semantic segmentation using deep learning remains a persistent challenge. In this work, we propose a novel semantic-aware contrastive framework that extends self-supervised contrastive loss and integrates attention guidance from coarse segmentation. Our proposed method leads to a significant gain in segmentation performance on two public contrast-enhanced CT datasets and one in-house non-contrast CT dataset.

CHAPTER 8

Adaptive Contrastive Learning with Dynamic Correlation for Multi-Phase Organ Segmentation

8.1 Overview

[1]Recent studies have demonstrated the superior performance of introducing "scan-wise" contrast labels into contrastive learning for multi-organ segmentation on multi-phase computed tomography (CT). However, such scan-wise labels are limited: (1) a coarse classification, which could not capture the fine-grained "organ-wise" contrast variations across all organs; (2) the label (i.e., contrast phase) is typically manually provided, which is error-prone and may introduce manual biases of defining phases. In this paper, we propose a novel data-driven contrastive loss function that adapts the similar/dissimilar contrast relationship between samples in each minibatch at organ-level. Specifically, as variable levels of contrast exist between organs, we hypothesize that the contrast differences in the organ-level can bring additional context for defining representations in the latent space. An organ-wise contrast correlation matrix is computed with mean intensity differences under one-hot attention maps. We further multiply the correlation matrix to scale the cosine distance between feature pairs, which adapt the contrast shift as variable levels of feature separability in the latent space. We evaluate our proposed approach on multi-organ segmentation with both non-contrast CT (NCCT) datasets and the MICCAI 2015 BTCV Challenge contrast-enhance CT (CECT) datasets. Compared to the state-of-the-art approaches, our proposed contrastive loss yields a substantial and significant improvement of 1.41% (from 0.923 to 0.936, p-value$<$0.01) and 2.02% (from 0.891 to 0.910, p-value$<$0.01) on mean Dice scores across all organs with respect to NCCT and CECT cohorts. We further assess the trained model performance with the MICCAI 2021 FLARE Challenge CECT datasets and achieve a substantial improvement of mean Dice score from 0.927 to 0.934 (p-value$<$0.01). The code is available at: https://github.com/MASILab/DCC_CL

8.2 Introduction

Multi-phase contrast CT delineates different anatomical structures with significant variation in contrast intensity (199; 80). As contrast agents progress through blood vessels, each organ of interest has its distinctive contrast uptake patterns, which lead to a wide range of intensity distribution across organs. It is challenging to generalize deep learning models and achieves consistent segmentation performance with significant contrast variations. The most naive way to adapt multi-phase representations is to train phase-specific networks and fuse the output features for organs segmentation (176). Phase/Modality-dependent normalization is further

[1]In submission at: Lee, Ho Hin, et al., "Adaptive Contrastive Learning with Dynamic Correlation for Multi-Phase Organ Segmentation.", Medical Image Analysis, 2023 (129)

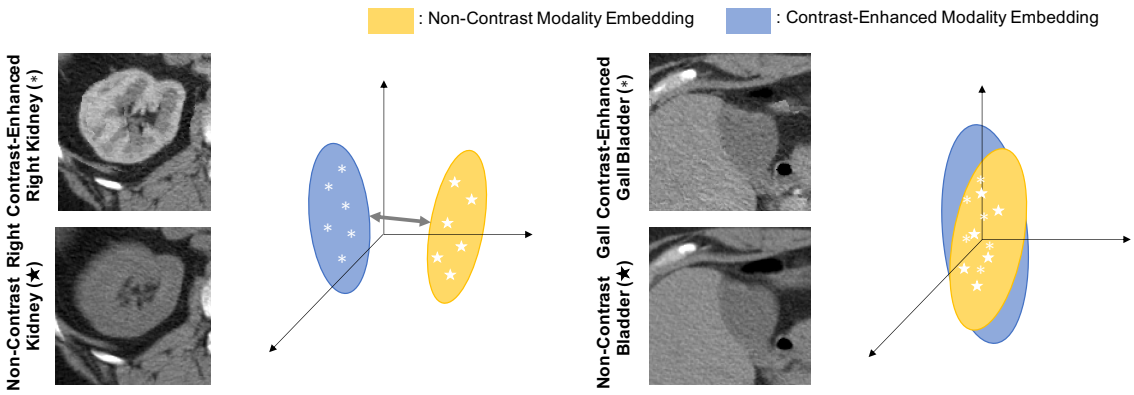


Figure 8.1: Between NCCT and CECT, some organs have significant contrast variation (such as the kidneys, liver, and spleen). We expect the corresponding embedding to be separable by contrast phases. However, some organs of interest such as the gall bladder, pancreas, and adrenal glands have similar contrast appearance across both modalities. As such, we expect the embedding of these organs to be aligned across phases, instead of separating into independent clusters using contrast label supervision. Such variation exposes a key limitation of current contrastive learning state-of-the-art approaches.

proposed to adapt independent normalization layers for each phase and mitigates the discrepancy between phases (54). However, limited studies have been proposed to study the multi-phase correspondence of abdominal organs in the feature level. Contrastive learning, a variant of self-supervised learning, has been shown to learn class-aware representations in a "scan-wise" setting and achieves significant improvements in the downstream tasks (33; 205; 23; 220). The theory of contrastive learning consists of two main operations: 1) pulling the representation of target image (anchor) and the matching sample together as a "positive pair", and 2) pushing the anchor representation from the remaining non-matching samples apart as "negative pairs". The goal of contrastive learning is to define class-aware embeddings without additional guidance. Furthermore, supervised contrastive learning is introduced to further enhance the ability of defining class-wise embeddings with the scan-wise label given (108; 88). However, scan-wise contrast information (e.g., phases) is limited to represent different scales of contrast variation across organs of interest. For example, we observed that kidney organ demonstrates a distinctive comparison between the appearance in contrast-enhanced and non-contrast phase, while only subtle variance is demonstrated for the gall bladder organ (in Figure 8.1). Therefore, it is challenging to leverage single hard label to represent such dynamic changes across organ intensities. As such, we can further ask: **Can we leverage the appearance difference to control and define organ-wise representations with contrastive learning?**

In this work, we propose a data-driven contrastive loss that leverages contrast correspondences between organs to adaptively model the representations into organ-specific embeddings. Specifically, we extract mean intensity difference under the organ-specific regions and scale the pairwise feature similarity/dissimilarity

with the contrast prior knowledge across all samples in each minibatch. The goal of using contrast correlation is to dynamically regularize the contrastive loss based on the wide range of contrast enhancement. Our proposed contrastive loss is evaluated with two public contrast-enhanced CT (CECT) cohorts and one research non-contrast CT (NCCT) cohort. The experimental results demonstrate consistent improvement in multi-organ segmentation using deeplabv3+ architectures with a ResNet-50 encoder backbone (79; 32). Our main contributions are summarized as follows:

- We propose an adaptive contrastive learning framework to improve and generalize multi-organ segmentation performance in multi-phase contrast CT.
- We propose a data-driven contrastive loss function that adapts the organ-specific contrast correlation as an adaptive weighting constraint to control the distance between multi-phase representations on the organ-level.
- We demonstrate that the proposed contrastive loss captures the wide range of multi-phase contrast differences without trading off performance from one of the phases and achieves significant improvement for downstream segmentation tasks.

8.3 Related Works

Medical Image Segmentation: Most modern approaches to perform medical image segmentation typically train a deep neural network directly in supervised setting with post-processing techniques (221). However, the model performance is greatly dependent on both the quality of ground-truth labels and the resolution of volumes (62). Patch-wise approaches and hierarchical approaches are proposed to adapt coarse-to-fine features and leverage the mutli-scale capabilities to generate refine segmentation (182; 259; 200). However, multiple models are typically needed to train for multiple semantic targets segmentation. Single coarse-to-refine network is proposed to adapt multi-organ segmentation by integrating binary organ-corresponding attentions as additional input channel (125). To further enhance the segmentation performance with stability, significant efforts are put into exploring the possibility of adapting unlabeled data in both semi-supervised or self-supervised setting. Generating quality assurance score for the intermediate predicted mask are proposed as an alternative supervision using unlabeled data (130). Different pretext tasks including colorization, deformation and image rotation, have been leveraged as pretraining strategies to provide a better initialization for segmentation networks (258). Furthermore, learning spatial context by predicting the degree of rotation and relative patch position are proposed to be beneficial for finetuning segmentation network (8; 260).

Image-Level Contrastive Learning: Significant efforts have been put into self-supervised learning to extract meaningful representations from unlabeled data. Previous works have demonstrated to learn represen-

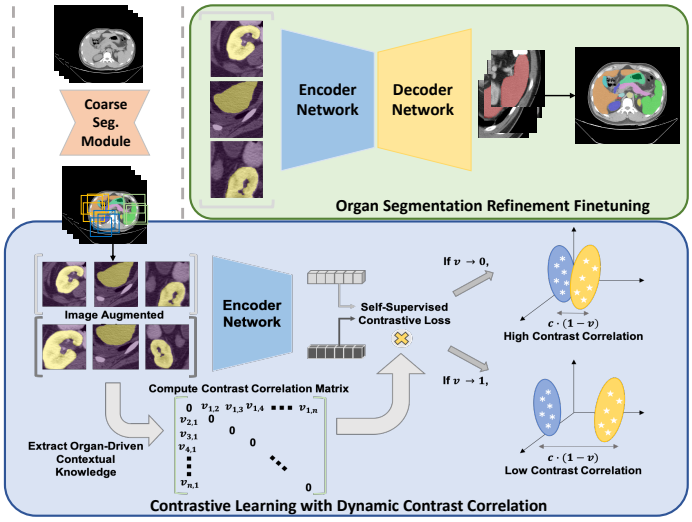


Figure 8.2: The complete contrastive framework can be divided into three hierarchical steps: 1) We first compute coarse segmentation and extract organ-corresponding patches for contrastive learning. The organ-specific attention masks are leveraged as additional channels to guide the representations extracted within specific regions. 2) For contrastive learning, we compute a contrast correlation matrix between samples in each minibatch to control the weighting of the contrastive loss dynamically across all pairs. We first compute the mean intensity of each organ of interest under each one-hot attention bounded region. **v** is the mean intensity difference between the augmented samples in each minibatch. **n** is the batch size of each minibatch. If the value of **v** is low, it corresponds to a high contrast correlation to scale the cosine distance between representations **c** (approaching to 1). 3) We finally finetune the well-pretrained encoder followed with a decoder network to generate refine multi-organ segmentation.

tations in the latent space using image reconstruction (243). The idea of contrastive learning further extends to learn and classify representations into embeddings by evaluating the pairwise similarity in the latent space. By leveraging data augmentations, the augmented pairs from the same image are defined as the positive pair and pull their representations closer together, while pushing the remaining representations apart as negative pairs (33). Instead of evaluating the feature similarity, maximizing the mutual information between representations have also been proposed to correlate the similar representations in the latent space (7). As the feature similarity is evaluated within a batch, increasing the batch size is another alternative to enhance the efficiency of contrastive learning. Using the memory bank and the momentum encoder for contrastive learning have been proposed to compare the query representations with more dissimilar representations within a minibatch (154; 77). Another perspective of contrastive learning is the definition of positive pair. Representations from the same class may also sample in the minibatch, while the contrastive loss can only select a specific positive pair. Supervised contrastive learning is introduced and leverages the image-level label to define arbitrary number of positive pairs, providing a better definition of the latent space for fintuning classification tasks (108). In the medical domain, continuous proxy meta-data is leveraged as the image-level additional guidance for contrastive learning with brain MRI (58). Furthermore, leveraging relative position of the extracted

patch for contrastive learning is demonstrated to be beneficial for fintuning detection task (134). For medical image segmentation, image-wise labels such as positional information, are additionally used to constrain representations in the latent space (241). To tackle the multi-contrast phase organ segmentation, the contrastive loss has extended to adapt multi-class "scan-wise" labels (e.g. contrast phases and organs) and defines representations into sub-classes embeddings to benefit segmentation performance. However, such hard labels are difficult to represent the dynamic contrast level across organ of interests. Current contrastive approaches are limited to provide flexibility of controlling the feature separability between representations adaptively.

Pixel-Level Contrastive Learning: While prior works have demonstrated leveraging contrastive learning to enhance "image-wise" downstream task performance, several approaches have been extended the theory basis of contrastive learning to pixel-wise setting. Instead of using linear projection, dense projection is used to compute dense mapping and evaluates the pixel-level similarity in self-supervised setting (220). After that, pixel-wise contrastive loss is proposed to evaluate the feature similarity with ground-truth label guidance (248). Apart from leveraging contrastive learning as pretraining strategies, single-stage framework is proposed to cotrain segmentation task with contrastive loss and enforces the representations in the same semantic class to be more similar as independent pixel-wise embeddings (219; 90; 1). In the medical perspective, feature representations in both the local and global views are computed, and evaluate the structural similarity across views with limited samples (23). Furthermore, limited number of ground truth-labels or pseudo predicted labels are further adapted with the pixel-wise contrastive loss and enhance the downstream segmentation performances (92; 147). However, such pretraining/co-training strategies still require pixel-wise label guidance.

8.4 Methods

Inspired by the visualization of variable organ contrast with different imaging protocols from Figure 8.1, our proposed method tackles the limitation of adapting fix contrast labels in AGCL (88), which computes a contrast-correlated prior to control the pairwise feature separation and generates a contrast-guided latent space for multi-phase organ segmentation. The complete hierarchical pipeline is presented in Figure 8.2. We can divide our hierarchical pipeline into three specific stages: 1) extracting organ-specific attention from coarse segmentation, 2) contrastive learning with contrast correlation, and 3) fine-tuning for organ segmentation refinement.

8.4.1 Organ-Specific Attention from Coarse Segmentation

Given a set of multi-contrast 3D image volumes $V_i = \{X_i, Y_i\}_{i=1,\ldots,L}$, where L is the number of all imaging samples, X is the volumetric image and Y is the corresponding multi-organ label. Inspired by (88), we first

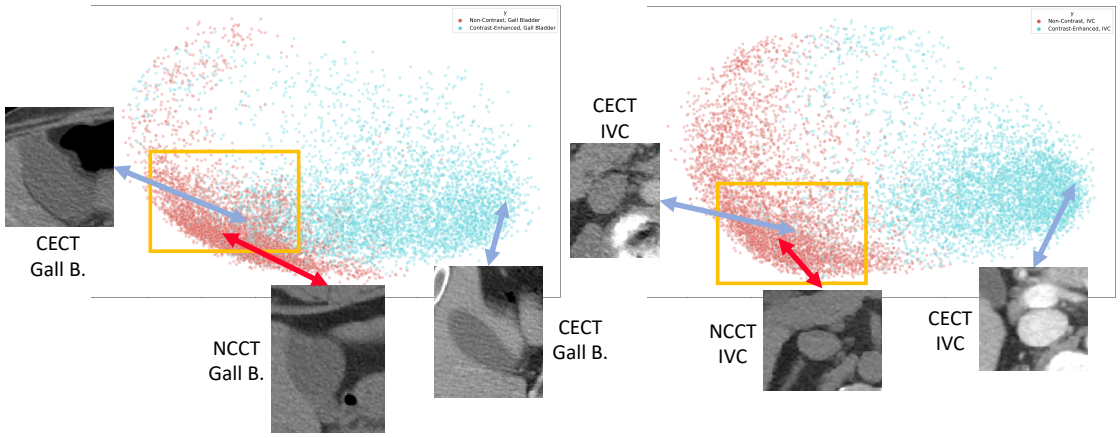


Figure 8.3: Dimensionality reduction with PCA is performed to visualize the distribution of learned representations. Both left and right plots are corresponding to the gall bladder feature and IVC feature respectively. By leveraging the contrast correlation as dynamic weighting, we found that the representations is well defined according to the contrast level, although they are in different "scan-wise" defined contrast phases.

generate coarse segmentation masks $A_i = Coarse(X_i)$ using RAP-Net, a two-stage network $Coarse(\cdot)$ that adapts organ-specific prior with image patches to perform segmentation refinement within the prior-bounded region. We define $A \in R^{H \times W \times C}$, where H and W denote as the axial dimension of the image, and C denotes as the total number of organ label classes. $Coarse(\cdot)$ is implemented with a hierarchical 3D U-Net architecture and is trained by the complete multi-contrast volumes with 5-fold cross-validationd (125). We then randomly sample organ-specific patches with the predicted masks' guidance for both the contrastive pretraining and the downstream fine-tuning. Briefly, we first convert all 3D volumes into 2D slices and extract 2-dimensional organ-specific patches $p_i = \{x_{C,i}, y_{C,i}, s_{C,i}\}_{i=1,\ldots,N}$ by randomly sampling pixel index of each organ class C as the center point to crop the regions of interest (ROIs), N denotes as the total number of query patches, $x_{C,i}$ is the image patch. Furthermore, we extract $y_{C,i}$ and $s_{C,i}$ by converting the organ ROIs in both ground-truth label patch and the coarse segmentation patch in binary setting, where $s_{C,i} \in A_i$. The coarse mapping is utilized as the organ attention prior for spatial restrictions of learning pixel-wise semantic representations. Specifically, inspired by (33), a data augmentation module $Aug(\cdot)$ is used to further extract correlated representations within the organ ROIs, including random cropping, rotation (-30 to 30 degrees), scaling (width: 0.3, height: 0.7) and generate $2n$ pairwise copies $\widetilde{x}_{C,i}, \widetilde{s}_{C,i} = Aug(x_{C,i}, s_{C,i})$. Each $\widetilde{s}_{C,i}$ is concatenated with $\widetilde{x}_{C,i}$ as a multi-channel input m_i for training encoder network $E(\cdot)$.

8.4.2 Contrastive Learning with Contrast Correlation

8.4.2.1 Phase-Driven Contrast Correlation

From Figure 8.1, we observe that the contrast uptake patterns between organs of interest are varied significantly, while subtle variations may also exist in some of the organs, even when they are in different modalities. However, a fix contrast label is limited to provide an adaptive contrast meaning that guides the separation between organ-specific features, especially with subtle contrast variation in different modalities scenario (88). Therefore, we hypothesize that the dynamic intensity difference in organ level is beneficial to adaptively define representations in the latent space. Here, we extract the mean intensity value d_i within the corresponding $\widetilde{s}_{C,i}$ and dynamically compute the contrast difference pairwisely across all image patches of each minibatch. The absolute difference of mean intensity across pairwise samples $v_{i,j}$ is computed as the contrast correlation context. We hypothesize that $v_{i,j}$ tends to 0 if the "organ-driven" mean intensity is similar. The correlation matrix is defined as following:

$$d_i = \frac{1}{|\phi|} \sum_{(x,y,c)\in\phi} \{\widetilde{x}_{C,i} \cdot \widetilde{s}_{C,i}\}(x,y,c) \tag{8.1}$$

$$v_{i,j} = |d_i - d_j|\,, \tag{8.2}$$

where x, y and c are corresponding to the x-y coordinates and number of channels within the attention bounded region. $\phi \in R$ is the number of nonzero pixels in the bounded region. i and j are the respective batch indices across all augmented samples.

8.4.2.2 Dynamic Contrast Correlation Contrastive Loss

As the current contrastive loss functions lack of flexibility in adjusting the level of separation dynamically between the multi-phase/modal representations with a fix label, we propose the dynamic contrast correlation (DCC) matrix as an adaptive weighting to multiply the cosine similarity computed in the feature-level. Instead of only computing feature-level contrastive loss in self-supervised setting, the dynamic constraints account the image-level variability as soft supervision to enhance the flexibility of standard contrastive loss, which help to control the cosine distance between the pairwise representations in each minibatch. The DCC matrix dynamically varies according to the shuffled samples in each minibatch.Our proposed contrastive loss as $\mathscr{L}_{dcc}$ is defined as follows:

$$\widetilde{z}_k, \widetilde{z}_{p(k)} = P(E(m_k, m_{p(k)}) \tag{8.3}$$

$$\mathscr{L}_{dcc} = -\sum_{k=1}^{2n} \log \frac{\exp(\widetilde{z}_k \cdot \widetilde{z}_{p(k)} \cdot (1 - v_{k,p(k)})/\mathscr{T})}{\sum_{j\in J(k)} \exp(\widetilde{z}_k \cdot \widetilde{z}_j \cdot (1 - v_{k,j})/\mathscr{T})}\,, \tag{8.4}$$

where $\widetilde{z}_k$ and $\widetilde{z}_{p(k)}$ are the pairwise feature representation vectors. The index k represents the sample of anchor and index $p(k)$ represents the corresponding positive. The hyperparameter $\mathscr{T}$ is the radius of the hypersphere that maps the representation inside as a point. $P(\cdot)$ is a linear projection network using multi-layer perceptron (MLP). Instead of constraining into label-class representation, the DCC-weighted contrastive loss preserves the data-driven knowledge for each organ and allows for similar organ-wise representations even if they are from different phases. The distance between the pairwise representations is controlled with the contrast correlation $(1 - v_{k,j})$, thus to enhance the flexibility of defining latent space.

Table 8.1: Comparison of current state-of-the-art methods on the BTCV challenge leaderboard. ($\star$: fully-supervised approach, $\triangle$: partially supervised approach, DACA: Data Augmentation with Contrast Adjustment, *: $p < 0.01$, with Wilcoxon signed-rank test.)

Method	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	Average Dice
(39)$\star$	0.906	0.857	0.899	0.644	0.684	0.937	0.790	0.886	0.808	0.643	0.724	0.632	0.784
(182)$\star$	0.935	0.887	0.944	0.780	0.712	0.953	0.823	0.880	0.804	0.680	0.741	0.653	0.816
(84)	0.920	0.894	0.915	0.604	0.692	0.948	0.810	0.857	0.828	0.640	0.735	0.619	0.790
(170)$\star$	0.939	0.895	0.915	0.711	0.743	0.962	0.827	0.891	0.826	0.700	0.739	0.639	0.815
(259)$\star$	0.935	0.886	0.944	0.764	0.714	0.942	0.835	0.879	0.803	0.651	0.752	0.669	0.814
(125)$\star$	0.959	0.920	0.945	0.768	0.783	0.962	0.867	0.910	0.847	0.712	0.767	0.680	0.842
(99)$\star$	0.956	0.923	0.940	0.760	0.764	0.965	0.870	0.905	0.850	0.703	0.771	0.670	0.839
(74)$\star$	0.959	0.912	0.940	0.724	0.746	0.968	0.873	0.905	0.840	0.767	0.786	0.701	0.836
(255)$\triangle$	0.968	0.920	0.953	0.729	0.790	0.974	0.891	0.925	0.847	0.780	0.801	0.734	0.850
(33)	0.953	0.922	0.930	0.842	0.822	0.972	0.907	0.899	0.874	0.800	0.854	0.625	0.868
(33) + DACA	0.957	0.935	0.941	0.867	0.840	0.977	0.914	0.909	0.891	0.812	0.873	0.655	0.881
(23)	0.956	0.935	0.946	0.920	0.854	0.970	0.880	0.915	0.893	0.786	0.843	0.631	0.874
(220)	0.963	0.939	0.900	0.815	0.838	0.976	0.924	0.922	0.907	0.840	0.867	0.675	0.882
(108)	0.965	0.948	0.958	0.890	0.865	0.975	0.950	0.930	0.920	0.879	0.895	0.732	0.910
(108) + DACA	0.967	0.951	0.963	0.901	0.878	0.980	0.955	0.934	0.923	0.885	0.902	0.747	0.915
(1)	0.954	0.933	0.932	0.903	0.858	0.973	0.921	0.918	0.904	0.853	0.871	0.723	0.890
(219)	0.966	0.942	0.955	0.886	0.860	0.975	0.943	0.930	0.908	0.881	0.892	0.745	0.913
(88)	0.971	0.955	**0.963**	0.910	0.886	0.984	0.965	0.941	0.932	0.893	0.917	0.769	0.923
Ours (DCC-CL)	**0.974**	**0.956**	0.960	**0.928**	**0.905**	**0.986**	**0.972**	**0.953**	**0.942**	**0.905**	**0.926**	**0.790**	**0.936**

8.4.3 Fine-tuning for Organ Segmentation Refinement

The final goal of our proposed approach is to learn the contrast-correlated semantic representations for segmentation refinement. After the encoder network is pretrained with a projection network, we withdraw the projection network and employ DeepLabV3+ network as the model backbone for segmentation refinement in a fully-supervised setting (30). Such backbone can share the same encoder structure with the contrastive pretrained encoder and adapts atrous spatial pyramid pooling (ASPP) as the decoder for multi-scale refinement. Here, we use the Dice loss to evaluate the regional similarity between the predicted segmentation and the ground-truth in binary setting for end-to-end optimization. Finally, we employ majority voting to fuse all the binary outputs from the same slice and concatenate all fused slices to generate the volumetric multi-organ mask as our refined output.

Table 8.2: Comparison of the current state-of-the-art methods on the non-contrast testing dataset. ($\star$: fully-supervised approach, $\triangle$: partially supervised approach, *: $p < 0.01$, with Wilcoxon signed-rank test.)

Method	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	Average Dice
(39)$\star$	0.937	0.856	0.912	0.690	0.631	0.920	0.768	0.880	0.769	0.596	0.723	0.465	0.762
(182)$\star$	0.940	0.890	0.923	0.701	0.724	0.948	0.747	0.878	0.770	0.574	0.724	0.470	0.771
(84)	0.910	0.865	0.889	0.624	0.656	0.930	0.743	0.860	0.759	0.556	0.720	0.458	0.748
(259)$\star$	0.948	0.880	0.920	0.710	0.734	0.950	0.791	0.879	0.803	0.603	0.745	0.510	0.790
(125)$\star$	0.954	0.874	0.928	0.701	0.753	0.958	0.814	0.897	0.794	0.611	0.764	0.525	0.798
(99)$\star$	0.963	0.930	0.946	0.876	0.792	0.970	0.836	0.919	0.843	0.637	0.786	0.544	0.836
(74)$\star$	0.971	0.923	0.947	0.857	0.789	0.965	0.805	0.908	0.821	0.597	0.757	0.509	0.820
(255)$\triangle$	0.960	0.900	0.943	0.739	0.810	0.965	0.878	0.920	0.856	0.651	0.798	0.563	0.833
(33)	0.964	0.938	0.946	0.800	0.801	0.969	0.946	0.901	0.869	0.739	0.804	0.386	0.848
(33) + DACA	0.973	0.953	0.950	0.843	0.825	0.974	0.953	0.913	0.881	0.754	0.822	0.439	0.857
(23)	0.969	0.940	0.955	0.910	0.834	0.970	0.943	0.911	0.867	0.748	0.812	0.420	0.854
(1)	0.965	0.945	0.957	0.914	0.837	0.974	0.954	0.918	0.910	0.770	0.845	0.473	0.873
(220)	0.979	0.961	0.964	0.941	0.865	0.979	0.960	0.937	0.923	0.775	0.837	0.534	0.887
(108)	0.975	0.952	0.962	0.943	0.857	0.976	0.958	0.925	0.915	0.769	0.831	0.484	0.879
(108) + DACA	0.980	0.959	0.967	0.948	0.869	0.980	0.962	0.933	0.921	0.779	0.849	0.523	0.889
(219)	0.970	0.960	0.949	0.940	0.832	0.974	0.948	0.922	0.920	0.762	0.836	0.461	0.872
(88)	0.982	0.962	0.965	0.834	0.879	0.982	0.967	0.945	0.929	0.790	0.850	0.560	0.892
Ours (DCC-CL)	**0.984**	**0.966**	**0.970**	**0.957**	**0.890**	**0.982**	**0.971**	**0.951**	**0.939**	**0.803**	**0.893**	**0.584**	**0.910**

8.5 Experimental Setup

8.5.1 Datasets

[I] MICCAI 2015 BTCV Challenge is comprised of 100 de-identified 3D contrast-enhanced CT scans with 7968 axial slices. 20 available scans are publicly available for testing phase. This dataset includes 12 well-annotated organs, including the spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava (IVC), portal splenic vein (PSV), pancreas and right adrenal gland. Each volume consists of $47 \sim 133$ slices of 512×512 pixels at a resolution of with resolution of $([0.54 \sim 0.98] \times [0.54 \sim 0.98] \times [2.5 \sim 7.0])mm^3$.

[II] Research Non-Contrast CT cohort is retrieved and consists of 56 volumetric CT scans with 3687 axial slices and expert refined annotations for the same 12 organs in BTCV dataset. Each volume consists of $49 \sim 174$ slices of 512×512 pixels, with resolution of $([0.64 \sim 0.98] \times [0.64 \sim 0.98] \times [1.5 \sim 5.0])mm^3$.

[III] MICCAI 2021 FLARE Challenge (151) adapts large scale of abdominal contrast-enhanced CT with 511 cases from 11 medical centers. In total, 361 CT consists of well-annotated labels for spleen, kidney, liver and pancreas organs. Each volume consists of $43 \sim 384$ slices of 512×512 pixels, with resolution of $([0.64 \sim 0.98] \times [0.64 \sim 0.98] \times [1.0 \sim 5.0])mm^3$.

8.5.2 Data Preprocessing

Before we input the complete volume into the network for coarse multi-organ segmentation, we apply hierarchical steps for data preprocessing. The first step is to perform soft tissue windowing between the range

Table 8.3: Ablation studies of segmentation performance with different pretraining scenarios of the BTCV testing cohort.

Encoder	Pretrain	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A	Mean
ResNet50	×	0.932	0.877	0.887	0.860	0.761	0.962	0.941	0.832	0.815	0.735	0.833	0.587	0.840
ResNet50	SSCL	0.953	0.922	0.930	0.842	0.822	0.972	0.907	0.899	0.874	0.800	0.854	0.625	0.868
ResNet50	CE	0.959	0.948	0.957	0.890	0.868	0.978	0.956	0.935	0.919	0.884	0.903	0.725	0.905
ResNet50	AGCL	0.971	0.955	**0.963**	0.910	0.886	0.984	0.965	0.941	0.932	0.893	0.917	0.769	0.923
ResNet50	DCC-CL	**0.974**	**0.956**	0.960	**0.928**	**0.905**	**0.986**	**0.968**	**0.950**	**0.939**	**0.901**	**0.923**	**0.776**	**0.936**

Table 8.4: Ablation studies of segmentation performance in various network backbones of the in-house non-contrast testing cohort.

Encoder	Pretrain	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A	Mean
ResNet50	×	0.960	0.918	0.921	0.754	0.783	0.964	0.950	0.840	0.839	0.691	0.796	0.372	0.816
ResNet50	SimCLR	0.964	0.938	0.946	0.800	0.801	0.969	0.946	0.901	0.869	0.739	0.804	0.386	0.848
ResNet50	CE	0.972	0.952	0.961	0.812	0.859	0.974	0.959	0.934	0.914	0.768	0.838	0.551	0.875
ResNet50	AGCL	0.982	0.962	0.965	0.834	0.879	0.982	0.967	0.945	0.929	0.790	0.850	0.560	0.892
ResNet50	DCC-CL	**0.984**	**0.966**	**0.970**	**0.957**	**0.890**	**0.982**	**0.971**	**0.951**	**0.939**	**0.803**	**0.893**	**0.584**	**0.910**

of -175 and 250 Hu. Intensity normalization is then performed for each volume and uses min-max normalization: $(X - X_1)/(X_{99} - X_1)$ to normalize the intensity value between 0 and 1, where X_p denotes as the p^{th} intensity percentile in X. Furthermore, due to various domain shifts and the variation of imaging protocols, the images collected usually present with different contrast and Fields Of Views (FOVs). Here, we leverage Body Part Regression (BPR) network to minimize the difference of FOVs and only extract the abdominal region of interest from each image (198). BPR is a self-supervised method to predict a continuous score for each axial slice of CT volumes as the normalized body standardize value without any labels. The BPR network predicts score in the range of -12 to +12 for each slice, which correspond well to the approximate anatomical location of each body part region (e.g., -12: upper chest, -5: diaphragm/upper liver, 4: lower retroperitoneum, 6: pelvis). To extract the abdominal-to-retroperitoneal regions of interest, we limit the predicted scores within -4 to 5 and crop the volumes that are out of ranges.

8.5.3 Model & Training Details

For contrastive pretraining step, the complete backbone of the network consists of a ResNet-50 image encoder and a projection network with two linear layers. For organ segmentation finetuning step, DeepLabV3+ network is employed by withdrawing the projection network and follows with a decoder using ASPP modules. Both the pretraining and finetuning steps are optimized with an Adam optimizer independently (weight decay: $1 \times 10^{-}4$; momentum: 0.9; batch size: 4). We pretrain the model with our contrastive approach for 10 epochs using a learning rate of 0.0003, while we finetune the model with 5 epochs with learning 0.0001. On a NVIDIA-Quadro RTX 5000, 1 epoch of pretraining takes about 12 hours to finish with batch size of 4.

Table 8.5: Ablation Studies on the effectiveness of Data Augmentation with Contrast Adjustment (DACA) in Finetuning

Portal Venous CT (BTCV)														
Pretraining	Finetuning w. DACA	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	Average Dice
×	×	0.932	0.877	0.887	0.860	0.761	0.962	0.941	0.832	0.815	0.735	0.833	0.587	0.840
×	✓	0.942	0.896	901	0.877	0.789	0.967	0.952	0.851	0.834	0.761	0.840	0.634	0.853
SimCLR	×	0.953	0.922	0.930	0.842	0.822	0.972	0.907	0.899	0.874	0.800	0.854	0.625	0.868
SimCLR	✓	0.968	0.943	0.959	0.877	0.861	0.978	0.932	0.910	0.889	0.854	0.878	0.683	0.894
AGCL	×	0.971	0.955	0.963	0.910	0.886	0.984	0.965	0.941	0.932	0.893	0.917	0.769	0.923
AGCL	✓	0.973	0.960	0.965	0.921	0.901	0.986	0.972	0.950	0.940	0.900	0.927	0.785	0.930
DCC-CL	×	0.974	0.956	0.960	0.928	0.905	0.986	0.972	0.953	0.942	0.905	0.926	0.790	0.936
DCC-CL	✓	0.977	0.965	0.968	0.934	0.911	0.988	0.975	0.958	0.949	0.910	0.934	0.798	0.940
Non-Contrast CT (Research Non-Contrast In-house)														
Pretraining	Finetuning w. DACA	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	Average Dice
×	×	0.960	0.918	0.921	0.754	0.783	0.964	0.950	0.840	0.839	0.691	0.796	0.372	0.816
×	✓	0.963	0.925	0.924	0.778	0.791	0.968	0.957	0.843	0.844	0.711	0.797	0.423	0.828
SimCLR	×	0.964	0.938	0.946	0.800	0.801	0.969	0.946	0.901	0.869	0.739	0.804	0.386	0.848
SimCLR	✓	0.972	0.949	0.953	0.821	0.865	0.974	0.951	0.919	0.891	0.764	0.823	0.489	0.864
AGCL	×	0.982	0.962	0.965	0.834	0.879	0.982	0.967	0.945	0.929	0.790	0.850	0.560	0.892
AGCL	✓	0.984	0.964	0.969	0.875	0.893	0.983	0.971	0.952	0.938	0.819	0.863	0.623	0.902
DCC-CL	×	0.984	0.966	0.970	0.957	0.890	0.982	0.974	0.951	0.939	0.828	0.862	0.611	0.910
DCC-CL	✓	0.987	0.970	0.973	0.958	0.901	0.984	0.975	0.956	0.940	0.835	0.870	0.623	0.914

8.5.4 Implementation Details

For internal evaluation, we evaluate the model performance with a downstream multi-organ segmentation task and compare with current state-of-the-art of fully-supervised and contrastive learning approaches for both BTCV and the in-house non-contrast CT datasets. For external validation, we randomly select 100 samples from FLARE dataset and perform inference with the model trained with internal cohorts. We use Dice similarity coefficient as an evaluation metric to compare the overlapping regions between the predictions and the ground-truth label. Furthermore, we performed ablation studies with hyperparameter variations and different strategies to pretrain the encoder network.

8.6 Experimental Results

8.6.1 Comparison with Fully/Partially Supervised Approaches

As the preliminary basis of our approach is to extract organ-specific patches with a coarse segmentation network, we evaluate and validate current supervised state-of-the-arts as our coarse segmentation backbone. All supervised approaches are presented in Table 8.1 and 8.2 for both internal cohorts, including 3D CNN-based network (39), hierarchical networks (182; 259; 125) and transformer-based network (74). Significant levels of improvement (from mean Dice 0.784 to 0.836, $p > 0.01$) are demonstrated starting from (39) to (74) for both contrast-enhanced BTCV and in-house non-contrast cohorts. (125) achieves the best segmentation performance in BTCV, while (74) demonstrates the best segmentation performance in non-contrast dataset. As the model is trained coherently in multi-contrast setting, the trained model may bias to one of the contrast

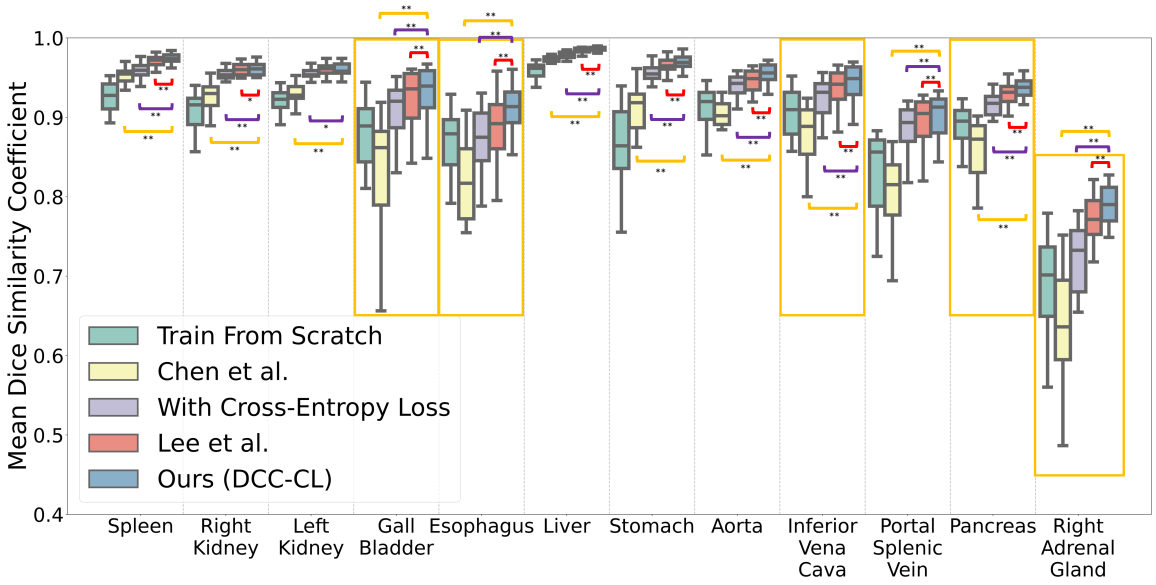


Figure 8.4: By adapting the contrast prior information into different pretraining strategies, such ablation studies demonstrate that DCC-CL outperforms the current state-of-the-art contrastive learning methods with *Chen et al*. and *Lee et al*.. The yellow box demonstrates the significant improvement in organs with subtle contrast variation between modalities. (*: p<0.05; **: p<0.01)

modality and thus leads to a level of degradation on certain organ segmentation performance with specific contrast. Furthermore, (255) demonstrates to leverage partial labeled samples to enhance the performance for multi-organ segmentation. We convert the partial supervised training scenarios of (255) as using one contrast samples for downstream multi-organ segmentation, while the another contrast samples are used in partially labeled setting. Interestingly, (255) demonstrates a stable improvement on both BTCV (mean Dice score from 0.842 to 0.850) and in-house non-contrast (mean Dice score from 0.815 to 0.833) cohorts. Such improvement in performance may due to the organ-specific prior distribution generated from (255) is beneficial to adapt the contrast variation in multi-organ segmentation. By leveraging the psuedo output from the supervised-trained model as organ attentional guidance for contrastive learning, the downstream segmentation performance is significantly enhanced across the contrastive learning state-of-the-arts.

8.6.2 Comparison with Contrastive Learning Approaches

After looking into the supervised approaches, we compare our proposed DCC-CL with a series of contrastive learning approaches for both internal and external testing cohorts, including SimCLR ((33)), local and global contrastive learning ((23)), DenseCon ((220)), SupCon ((108)), PixelCon (219) and multi-label contrastive learning ((88)), in Table 8.1 & 8.2. By adapting self-supervised approaches with the organ-attention guidance,

Table 8.6: Ablation studies of segmentation performance with Single/Multiple Phase Contrastive Pretraining (PV: Portal Venous CT; NC: Non-Contrast CT)

Portal Venous CT (BTCV)													
Contrast Phases	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A	Mean
PV Only	0.975	0.962	0.966	0.933	0.914	0.985	0.970	0.955	0.943	0.908	0.929	0.792	0.938
PV + NC	0.974	0.956	0.960	0.928	0.905	0.986	0.968	0.950	0.939	0.901	0.923	0.776	0.936
Non-Contrast CT (Research Non-Contrast In-house)													
Contrast Phases	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A	Mean
NC Only	0.979	0.962	0.966	0.949	0.881	0.980	0.969	0.940	0.928	0.812	0.857	0.597	0.902
PV + NC	0.984	0.966	0.970	0.957	0.890	0.982	0.974	0.951	0.939	0.828	0.862	0.611	0.910

Table 8.7: Ablation Studies on the effectiveness of correlation matrix in different scenarios

Portal Venous CT (BTCV)													
Correlation Matrix	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	Average Dice
×	0.932	0.877	0.887	0.860	0.761	0.962	0.941	0.832	0.815	0.735	0.833	0.587	0.840
Random Guassian	0.954	0.923	0.934	0.878	0.810	0.971	0.956	0.901	0.899	0.810	0.877	0.659	0.881
Dynamic Contrast	0.974	0.956	0.960	0.928	0.905	0.986	0.972	0.953	0.942	0.905	0.926	0.790	0.936
Non-Contrast CT (Research Non-Contrast In-house)													
Correlation Matrix	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	Average Dice
×	0.960	0.918	0.921	0.754	0.783	0.964	0.950	0.840	0.839	0.691	0.796	0.372	0.816
Random Guassian	0.969	0.934	0.945	0.847	0.973	0.971	0.962	0.901	0.878	0.754	0.810	0.483	0.869
Dynamic Contrast	0.984	0.966	0.970	0.957	0.890	0.982	0.974	0.951	0.939	0.828	0.862	0.611	0.910

we adapt the intrinsic structure of SimCLR to define the organ-attentional representation in the latent space and demonstrate significant improvements of 2.49% and 6.27% in Dice score for BTCV and non-contrast cohorts respectively. By further extracting representations in both local and global field of view, a slight increase is demonstrated from Dice score 0.863 to 0.874 for BTCV and 0.848 to 0.854 for non-contrast cohort. Apart from using linear projection to extract global-wise features, DenseCon proposes to adapt dense projection and evaluates the similarity between each feature vector in the dense mappings. Interestingly, a significant enhancement with 3.86% Dice score ($p > 0.01$) is demonstrated for non-contrast cohort, while only a slight increase with 0.915% Dice score is shown for BTCV.

(108) and (88) further extend the self-supervised contrastive loss and leverage single/multi-class labels to define conditional positive pairs for learning independent embedding of each semantic class. The integration of image-level labels demonstrates significant improvement on finetuning downstream segmentation task. (219) extract the semantic context from the segmentation and define the class-wise positive pairs from the psuedo predictions. Interestingly, both (108) and (219) demonstrate significant improvement of Dice score 0.907 and 0.913 for BTCV respectively, while slightly decreases of performance are shown in the non-contrast cohort. The reason for such degradation may correspond to the contrast variation and the learned representations only defined as class-wise embeddings without contrast information. (88) target the limitation

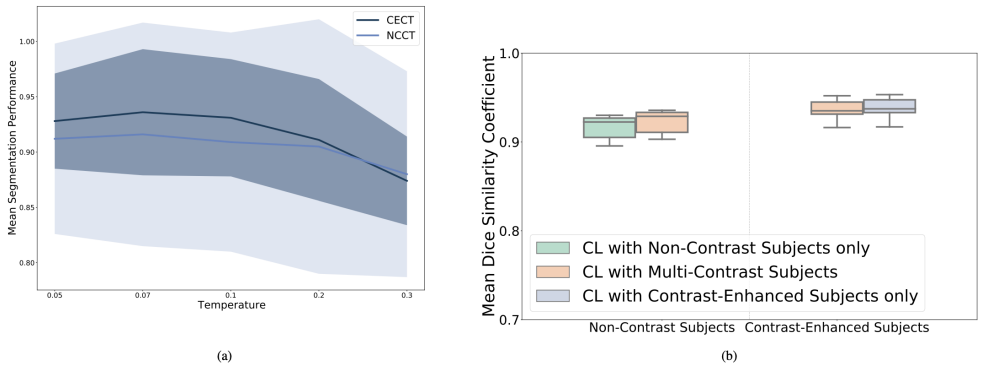


Figure 8.5: **a)** Variability of temperature scaling demonstrates that the segmentation performance is best optimized when $T = 0.07$. The color bands represents the standard deviation of the computed mean Dices score from all testing samples. **b)** Segmentation performance of NCCT is significantly improved without trading off CECT performance.

of using contrast information and adapt both contrast and organ label to define representations in the latent space. (88) outperforms the previous contrastive learning state-of-the-arts for both internal cohorts with Dice score of 0.923 and 0.892 respectively. By further adapting the contrast correlation as dynamic weighting constraints, DCC-CL achieves the best performance among all state-of-the-arts with a mean Dice score of 0.936 and 0.910 for both BTCV and non-contrast cohorts respectively. The additional gains demonstrate that the use of data-driven information provides flexibility to control and define multi-phase latent spaces.

Furthermore, we perform external evaluations on FLARE dataset with all contrastive state-of-the-art approaches, as shown in Table 8.8. As we leverage RAP-Net ((125)) as the coarse segmentation backbone, it demonstrates a stable performance on external dataset with a mean dice of 0.885. A stable enhancement of performance is demonstrated across all state-of-the-arts from Dice score of 0.885 to 0.927. Leveraging the dynamic weighting with contrast correlations further benefit the generalizability of the model and shows that DCC-CL outperforms all contrastive learning state-of-the-arts with the mean Dice of 0.927 to 0.934 .

8.6.3 Ablation Studies

Comparing with Contrastive Pretraining Strategies. To investigate the effect of using image-level contrast correlation as dynamic weight for contrastive learning, we perform multiple pretraining scenarios to compare the downstream tasks performance: 1) training from scratch, 2) pretraining with self-supervised contrastive loss (SSCL, (33)), 3) pretraining with cross-entorpy (CE) loss as classification tasks, 4) pretraining with multi-label contrastive loss (AGCL, (88)) in ResNet-50 encoder backbone. We start with the scenario of "training from scratch" and leverage the self-supervised contrastive loss for pretraining as the basis of comparison. After that, instead of using contrastive loss to constrain the representations, we further implemented the

Table 8.8: Comparison of the current state-of-the-art methods on FLARE dataset.

Method	Spleen	Kidney	Liver	Pancreas	Average Dice
(125)	0.956	0.903	0.954	0.730	0.885
(33)	0.960	0.910	0.960	0.756	0.896
(23)	0.961	0.923	0.956	0.787	0.908
(220)	0.966	0.918	0.964	0.800	0.912
(108)	0.963	0.918	0.966	0.830	0.919
(219)	0.968	0.940	0.964	0.811	0.922
(88)	0.971	**0.948**	0.968	0.823	0.927
Ours (DCC-CL)	**0.975**	0.944	**0.971**	**0.847**	**0.934**

classification scenario by classifying the representations into organ and contrast via the canonical CE loss. We finally compare the scenario of leveraging hard multi-class label with our proposed scenario. All pretraining scenarios are to define representations into corresponding contrast and organ defined embeddings.

As shown in Figure 8.4 and Table 8.3 & 8.4, pretraining with SSCL demonstrates a significant improvement of 3.12% Dice than the scenario of without pretraining. Such improvement is expected that pretraining with SSCL provides an initial definition of the learned representations in the latent space and hypothesizes that such learned embeddings are beneficial to the downstream tasks. The scenario of without pretraining is only relied on the ability of the decoder and defines representations with the downstream tasks guidance. By adapting the hard image-level label for pretraining, the segmentation performance significantly boosts from Dice score 0.868 to 0.905 and further enhances to 0.923 by leveraging the multi-label into the contrastive loss. Using CE loss as pretraining is to classify representations into class-wise embeddings with the labels given and the learned representations in the same class are moved towards each other. By adapting the multi-label context into contrastive loss, AGCL demonstrates the flexibility of moving the same class representation near and pushing the unrelated classes representations away in the latent space, improving the segmentation performance with a better definition of latent space. Eventually, DCC-CL surpasses AGCL by 1.41% in mean Dice. Instead of providing hard labels for searching similar representations, the computed contrast correlation matrix can leverage as a soft weighting to control the cosine similarity between pairwise representations and adapts the contrast variation more effectively across organ of interests.

Finetuning with Contrast Adjustment. One of the naive solutions to tackle the contrast variation challenges is to adjust the contrast level as one kind of data augmentations during training. Here, we investigate the effectiveness of the contrast adjustment in finetuning step. Table 8.5 demonstrates the quantitative performance of each pretraining scenario with/out contrast adjustment in finetuning. We observe that significant

improvements in performance are demonstrated in both the "train from scratch" and SimCLR scenarios, while subtle enhancement are demonstrated in both AGCL and DCC-CL scenario. With the additional constrains in pretraining step (AGCL: contrast label; DCC-CL: contrast correlation matrix), the benefits of contrast adjustment may become saturated if the latent space is already defined with contrast prior guidance.

Temperature Variability. We further vary temperature to investigate the effect of temperature on fine-tuning the segmentation task across scales. Figure 8.5(a) demonstrates the effect of temperature across all subjects in BTCV testing dataset. We find that the optimal temperature is roughly around 0.07 and the segmentation performance starts to degrade with the increasing temperature. Low temperature tends to penalize more to the highly similar representations and enhances the difficulties of searching positive pairs. Therefore, computing the contrastive loss with low temperature may achieve better segmentation performance than that using high temperature in our scenario.

Single/Multiple Phase Contrastive Learning. In the training scenario, we also want to investigate the effectiveness of applying DCC-CL with single phase CT only and with multi-phases CT for downstream segmentation. In Figure 8.4(b) & Table 8.6, we observe that the segmentation performance with non-contrast cohort is significantly improved by training in multi-contrast scenario, while the segmentation performance with BTCV is comparatively similar to each other. The contrast correlation matrix aims to generalize different level of contrast variation. The segmentation performance with both modalities samples are well preserved without trading off one of the modalities performances.

Correlation Matrix Variability As the contrast correlation matrix is leveraged to scale the cosine distance between the learned features, we further verify the effectiveness of the correlation matrix in Table 8.7 by replacing it with a matrix of random Gaussian-distributed variables. Interestingly, we found that significant improvements of segmentation performance are demonstrated in both portal venous CT and non-contrast CT by adapting a random Gaussian matrix to control the feature separation in the latent space. Although it provides a random constrain to control the feature separation, such constrain is limited to introduce the contrast correlation prior to adapt the multi-contrast characteristics across organs. By using our proposed matrix, it demonstrates its effectiveness in both multi-contrast CTs with the best performance.

8.6.4 Discussion & Limitations

In this study, we present an adaptive contrastive pre-training framework that leverages the variable level of contrast characteristics in each organ of interests to guide the separability of semantic embeddings. We hypothesize that such dynamically defined latent space is more beneficial to the downstream segmentation task, compared to the hard label defined latent space. With the organ attention guidance, the model allows to extract meaningful representations within the organ attention region. Also, we can further extract additional

context with significant variability from the image itself (e.g. contrast). Instead of using one-hot hard labels to define the "scan-wise" contrast level, the similarity between the mean intensity demonstrates to be a dynamic soft constrain and define the representations adaptively. From Figure 8.3, the PCA plot demonstrates the separability of the corresponding organ-wise representations. The orange bounding box localizes the organ representations with similar contrast, while the images are sampled from different contrast phases. The corresponding image visualization further demonstrates the contrast correspondence in the organ and the effectiveness of leveraging the contrast information for pretraining. In both Table 8.1 and 8.2, we have shown that our proposed pretraining strategy outperforms the current contrastive state-of-the-art approaches for multi-contrast organ segmentation. Meanwhile, we preserve the organ segmentation performance by training in the multi-contrast setting, instead of trading off the performance from one of the modalities. Such dynamic weighting further provides a better understanding on leveraging soft labels to define latent space for downstream segmentation task.

Our proposed pretraining strategy DCC-CL leverages a data-driven dynamic constrain with contrast intensity correlation to control the cosine distance separation between the learned features, while the previous approach AGCL only adapts fix image-level labels to constrain feature into semantic clusters. From Figure 8.3, we observe that the learned features for organs with similar contrast is pulled near, while organs with. The features from organs with different contrast are pushed away. However, for AGCL, the learned features of organs with similar contrast are still pushed away when they are in different modalities. With the domain shift of contrast appearance in multi-contrast CT, such intensity correlation constrains define the feature separation for adapting the domain shift in latent space and provides the most benefit to enhance the generalizability for downstream segmentation. While the correlation matrix is data-driven, the mean intensity difference may not be the optimal criteria to represent the domain shift of variable contrast. Generating a learnable matrix to control the cosine distance between the learned features may be a potential direction as future work.

One limitation of DCC-CL is its dependence on coarse pseudo labels. As the 2D patches are extracted with the pseudo label guidance, inaccurate organ patches may also be extracted and be unable to provide accurate organ-wise intensity information. Another limitation is that the representations are learned in an organ-centric setting. We aim to adapt the contrast correlation between the neighboring organs to an end-to-end framework in our future work.

8.7 Conclusion

Adapting the multi-phase contrast imaging with deep learning models remains a persistent challenge for performing robust semantic segmentation with wide range of contrast variation. In this work, we propose a novel contrastive loss function that control the distance between representations with dynamic contrast correlation

guidance. Our proposed contrastive loss contributes a significant gain of the segmentation performance across multi-phase contrast CT.

CHAPTER 9

Region-based Contrastive Pretraining for Medical Image Retrieval with Anatomic Query

9.1 Overview

[1] We introduce a novel Region-based contrastive pretraining for Medical Image Retrieval (RegionMIR) that demonstrates the feasibility of medical image retrieval with similar anatomical regions. RegionMIR addresses two major challenges for medical image retrieval i) standardization of clinically relevant searching criteria (e.g., anatomical, pathology-based), and ii) localization of anatomical area of interests that are semantically meaningful. Our approach utilizes a Region-Of-Interest (ROI) based image search that works at scale, enabling clinicians to search and retrieve selected ROIs that correspond to the same anatomy and/or similar pathological conditions. Previous approaches match similar images as a whole, without capturing the fine-grained details of specific anatomical regions. In this work, we propose an ROI image retrieval image network that retrieves images with similar anatomy by extracting anatomical features (via bounding boxes) and evaluate similarity between pairwise anatomy-categorized features between the query and the database of images using contrastive learning. ROI queries are encoded using a contrastive-pretrained encoder that was fine-tuned for anatomy classification, which generates an anatomical-specific latent space for region-correlated image retrieval. During retrieval, we compare the anatomically encoded query to find similar features within a feature database generated from training samples, and retrieve images with similar regions from training samples. We evaluate our approach on, both: anatomy classification and image retrieval tasks using the Chest ImaGenome Dataset. Our proposed strategy yields an improvement over state-of-the-art pretraining and co-training strategies, from 92.24 to 94.12 (2.03%) classification accuracy in anatomies. We qualitatively evaluate the image retrieval performance demonstrating generalizability across multiple anatomies with different morphology.

9.2 Introduction

Using recent advances in imaging and AI technology from clinics, hospitals and other medical sites have generated large digital stores of medical data (109), leading to a number of challenges for robust clinical workflows that provide data security, accessibility, interoperability and retention. However, this trend has also brought with it unprecedented opportunities for exploration of new technical capabilities, such as Content Based Image Retrieval (CBIR) (140; 34). Despite the advances in CIBR for natural images, and increased

[1] Published at: Lee, Ho Hin, et al., "Region-based Contrastive Pretraining for Medical Image Retrieval with Anatomic Query.", ArXiv, 2023 (124)

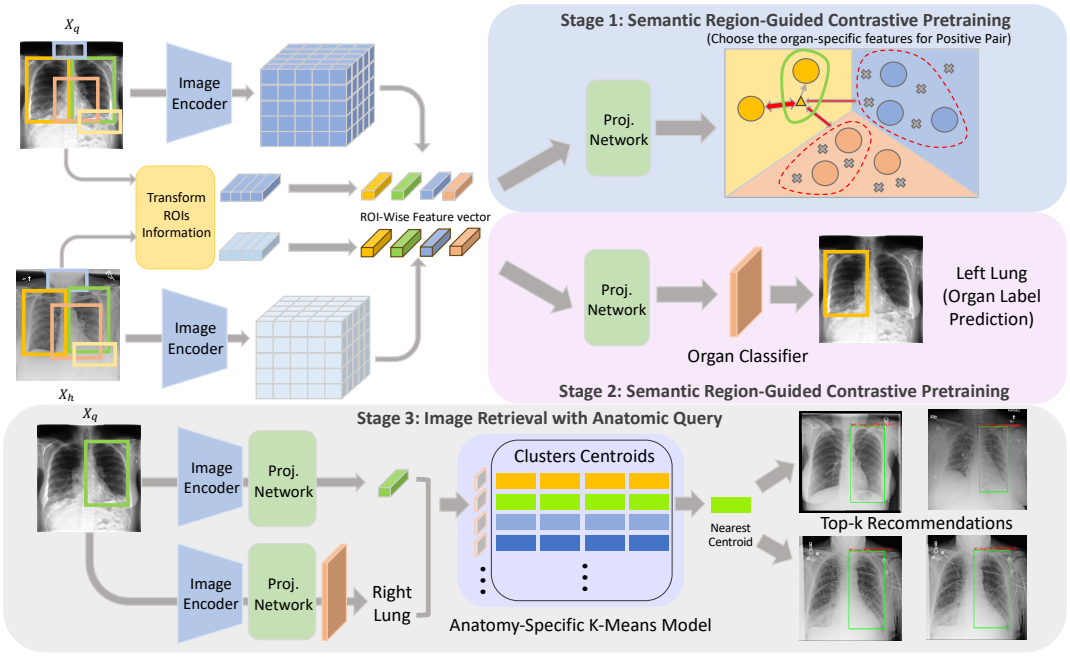


Figure 9.1: Overview of RegionMIR. RegionMIR consists of three main steps: 1) regional pooling and constrain the projected vectors into anatomical-specific clusters, 2) fine-tuning both the encoder and the projection network with anatomy classification task to stabilize the anatomical-specific latent space, and 3) training anatomical-specific K-means models to search the most similar centroid for image recommendations.

accessibility of large scale storage of medical images, only limited development for medical image retrieval algorithms has occurred(174; 250; 168; 37).

A number of challenges obstruct widespread progress in CBIR for medical imaging including: i) standardization of a clinical-relevant retrieval criteria (e.g., anatomy, pathology) and ii) localization of anatomical area of interests that are semantically meaningful. Though recent works have demonstrated feasibility of medical CBIR through regional context with multiple pathologies (250; 168; 37; 235; 106), these works have been limited to 2D and complete image content. The benefits of robust medical CBIR algorithms to diagnosis and treatment of patients are far reaching. For example, faster diagnosis and improved treatment plans can be developed through identification of pathologies by searching for similar characteristics in other patients. In addition, discovery and recognition of rare diseases through image retrieval are based on specific ROIs within large clinical databases. To enhance the efficiency of image retrieval, deep learning frameworks are employed to extract meaningful features and evaluate similarity (197; 89; 37). We observed that the image retrieval performance depends on both the quality and our defined semantic meanings of the representations learned. Contrastive learning offers semantically rich representations by grouping samples into self-supervised semantic-specific clusters, which leads to a significant improvement in downstream tasks (33; 88; 129; 214). We hypothesize that defining clusters with semantic meanings can enhance the robustness

of image retrieval. Currently, most of the contrastive learning strategies involve the use of 'whole' images, while limited works are proposed to incorporate both the local and contextual information within region of interests (ROIs). Furthermore, multiple semantic meanings may exist within a subject image or a ROI. It is challenging to define clusters with multiple meanings using current contrastive learning scenarios. As such, we ask: *can we adapt a contrastive learning framework to define regional representations with single semantic meaning and perform robust image retrieval with region queries?*

In this work, we propose a complete hierarchical framework for enhancing medical image retrieval by adapting to anatomy-specific representations in a regional setting. Our framework, named RegionMIR, extends the CBIR task by incorporating regional contrastive learning to generate an anatomy-defined latent space and evaluate the query feature similarity for image retrieval. We employ a methodolgy built upon BioViL(17) and RegionCLIP (251), and further develop a hierarchical search strategy based on the similarity from unsupervised clustering centroids, which differentiates the learned representations into fine-grained conditions (e.g. patholgies) and improves computational efficiency for image retrieval. Our proposed framework is evaluated with one public chest x-ray dataset using 5-fold cross-validation. The experimental results demonstrate a consistent improvement in anatomy classification with a ResNet-50 encoder backbone. Our main contributions are summarized as follows: 1) We propose a hierarchical framework RegionMIR to adapt anatomy-specific representations for image retrieval with regional query; 2) We propose to leverage an unsupervised centroid-based hierarchical search to retrieve top-5 images with reduced time complexity; 3) We demonstrate that RegionMIR learns the region-wise representation with their corresponding semantic meanings, achieving consistent improvements for downstream anatomy classification task and accurate retrieval in different anatomical regions.

9.3 Methods

We introduce the overview of our complete framework RegionMIR in Figure 9.1. Our primary goal is to learn region-guided representations that are able to differentiate different anatomies within medical images. RegionMIR consists of three hierarchical stages: i) semantic region-guided contrastive pretraining, ii) fine-tuning with anatomy classification, and iii) image retrieval with region query.

9.3.1 Semantic Region-Guided Contrastive Pretraining

Given a set $X = \{x_i, A_i\}_{i=1,\ldots,n}$, where n is the total number of images, $A = \{b_i, y_i\}_{i=1,\ldots,c}$ is the corresponding bounding box and image-wise labels for all anatomies. The index c represents the total number of semantic regional classes. As a single slice image usually contains rich semantics of multiple anatomies, RegionMIR leverages anatomy-specific bounding boxes to generate large pool of semantic embeddings and learns the

regional concepts, regardless of individual full images. First, we randomly sample images from X as query $Q = \{x_q, A_q\}_{q=i,\ldots,i+m-1}$ and anchor $H = \{x_h, A_h\}_{h=j,\ldots,j+m-1}$, where i,j are the randomly sampled index and m is the batch size for training. High-dimensional feature mappings of both query and anchor samples are extracted by the image encoder $E(\cdot)$. Next, we pool features with bounding box label b then linearly project the result into an 1-D embedding space using a multi-layer perceptron (MLP) $P(\cdot)$ as follows:

$$\begin{aligned} \{r_{q,i}, r_{h,i}\}_{i=1,\ldots,c} &= \mathit{ROIPool}(\{E(x_q), b_q\}, \{E(x_h), b_h\}) \\ \{z_{q,i}, z_{h,i}\}_{i=1,\ldots,c} &= P(\{r_{q,i}, r_{h,i}\}_{i=1,\ldots,c}) \end{aligned} \tag{9.1}$$

where $r_{q,i}$ and $r_{h,i}$ are the region-pooled anatomy-specific representations, while $z_{q,i}$ and $z_{h,i}$ are the corresponding projected vectors. Instead of evaluating the image-wise feature similarity between augmented pairs (33; 108), we compute feature similarity region-to-region and define the anatomical-specific features from pairwise samples (query and anchor) as the positive pair. We extend the image-wise contrastive loss to region-wise setting as follows:

$$\mathscr{L}_{region} = -\frac{1}{c}\sum_{i=1}^{c}\sum_{q=1}^{n}\sum_{h=1}^{n}\log\frac{\exp(\widetilde{z}_{q,i}\cdot\widetilde{z}_{h,i}/\tau)}{\sum_{j=0}^{2n-1}\sum_{k=1}^{c}\exp(\widetilde{z}_{q,i}\cdot\widetilde{z}_{j,k}/\tau)}, \tag{9.2}$$

where $z_{j,k}$ are the remaining features within the minibatch. Our proposed regional contrastive loss $\mathscr{L}_{region}$ constrains the same features (possibly with different morphologies) into anatomy-specific embeddings, forcing the model to have a coarse anatomy-to-region correspondence with semantic meanings.

9.3.2 Finetuning with Anatomy Classification

Our ultimate goal is to generate an anatomy-defined latent space and leverages regional features to perform image retrieval of images that contain similar regions. After the encoder network is contrastive-pretrained with a MLP, instead of only fine-tuning the encoder itself due to the image retrieval strategy (discussed in greater details in section 2.3), we fine-tune both the encoder and the MLP followed by a linear layer $L(\cdot)$ for anatomy classification. Finally, we use the multi-class cross-entropy loss to classify each region-pooled feature into specific anatomical embeddings as follows:

$$\{\tilde{y}_i\}_{i=1,\ldots,c} = L(\{z_{q,i}\}_{i=1,\ldots,c}) \tag{9.3}$$

$$\mathscr{L}_{finetune} = -\frac{1}{|n|}\sum_{k=0}^{n}\sum_{i=0}^{c} y_{k,i}\log(\tilde{y}_{k,i}), \tag{9.4}$$

by classifying each region-pooled feature with the anatomy label given, the regional representations are further refined and constrained into the anatomical-specific clusters, as shown in Figure 9.1. With the enriched

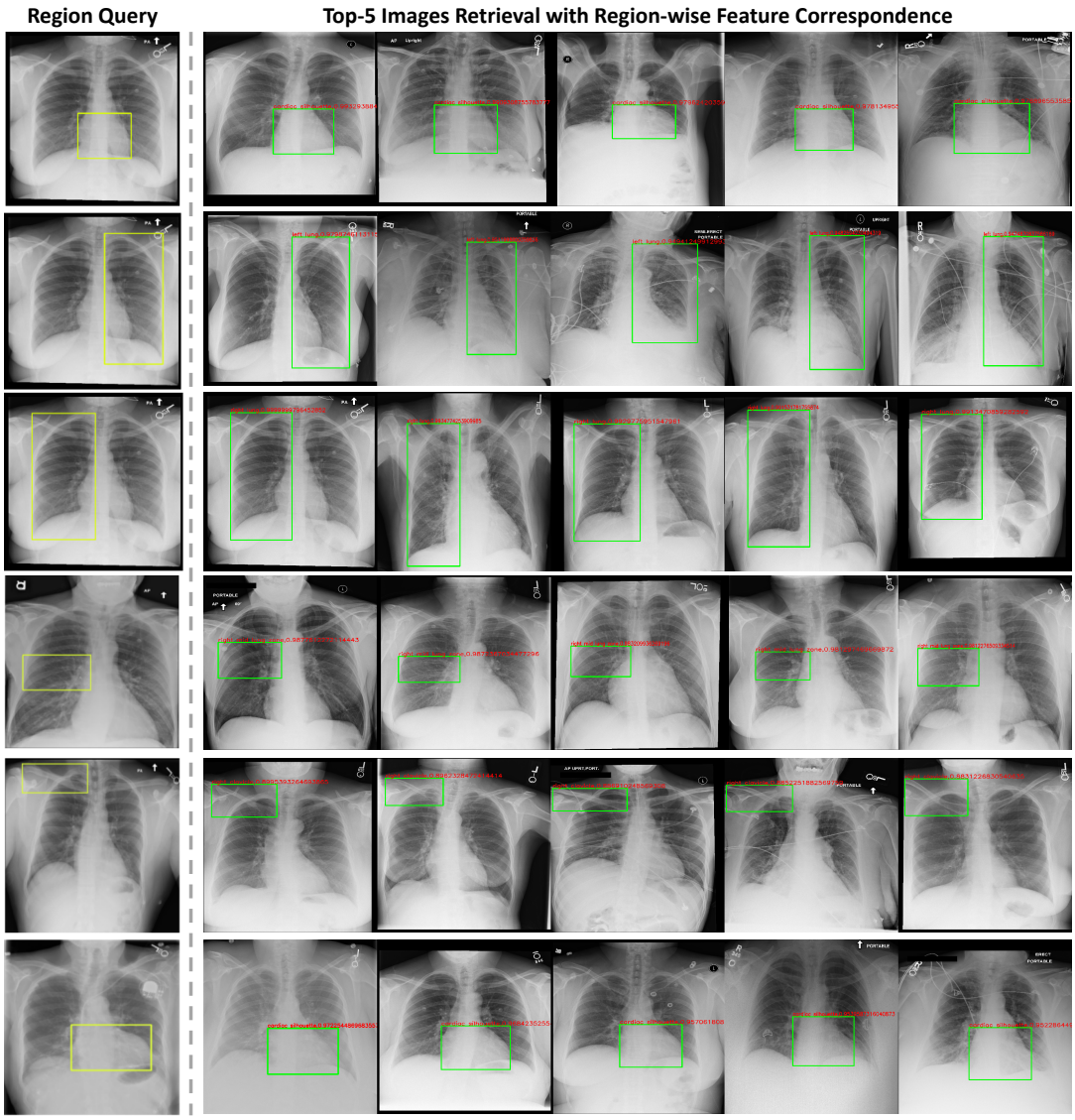


Figure 9.2: Top-5 image retrieval examples of different anatomical regions. We observe that images with same regions can be retrieved with the region query, while the organ morphology across all recommendation subjects significantly vary.

anatomical meanings in regional representations, such network enhance the feasibility of extending image retrieval with regional similarity and search images with corresponding regional semantics.

9.3.3 Image Retrieval with Anatomic Query

After fine-tuning the learned representations into anatomy-corresponding embeddings, we leverage $P(\cdot)$ project all features from the encoder network into an L^2-normalized space and evaluate the feature-wise similarity for image retrieval. Here, we use the cosine similarity to evaluate the representations and retrieve images with the similar region-pooled representations. In our searching scenario, we pre-compute represen-

tations from training samples as the retrieval database and use the testing samples as input queries to retrieve images from the database. The most naive approach is to compute similarity between the query representations and all representations in the database for ranking. However, with a time complexity of $O(n)$, this brute force ranking strategy is not computational efficient. Instead, we leverage the anatomy classification predictions and reduce the searching space for each query. Specifically, we train anatomy-specific K-means models to classify the pre-computed representations into different clusters' centroids. We further leverage these anatomy pseudo labels to choose the corresponding K-means model and evaluate the similarity across the centroids. It further enhance the computation efficiency and reduce the time complexity of image retrieval significantly from $O(n)$ to $O(4(centroids) * 26(anatomies))$. We finally retrieve the top-k images that are within the highest similarity cluster, where k is the number of searched images.

9.4 Implementation Setups

Dataset Details. In this study, we focus on the Chest ImaGenome dataset (226), which was constructed from the MIMIC-CXR dataset with 2-dimensional chest x-rays (106). This dataset consists of two standards: 1) gold standard and 2) silver standard for dataset curation. Here, we use the gold standard dataset with ground-truth quality label from four independent clinicians to perform training and evaluation. In total, 500 random patients' scans were sampled and annotated with bonding boxes consisting of 26 classes of anatomies (provided in appendix A1) as well as pathology label and clinical attributes.

Model Details. The image encoder is with a ResNet-50 backbone and follows with a projection network, consisting of two consecutive linear layers with ReLU activation in-between. The implementation of training anatomy-specific K-means models is based on scikit-learn and we set the number for unsupervised clusters in each model as 4. Note that the batch size of the anchor input has to be larger than the number of unsupervised clusters, as insufficient samples are limited to demonstrate the distribution of clusters through K-means.

Training Details. Both contrastive pre-training and anatomy classification fine-tuning are trained with an Adam optimizer (weight decay: 10^{-4}; momentum: 0.9; batch size: 8). For contrastive pre-training, we pre-train the model for 50 epochs with a learning rate of 3×10^{-4} and resize all input samples into 512×512 without cropping background. For anatomy classification finetuning, we finetune both the encoder and the projection network following with an additional linear layer for an additional 50 epochs a a lower learning rate of 10^{-4}. During this phase we use the same set of resizing parameters as the pre-training. We perform both training and inference on one Tesla-V100-32GB GPU. Fifty training epochs take about 10 hours with the batch size of 8.

Performance Comparisons. We evaluate the quantitative performance in anatomy classification task and the quantitative performance in image retrieval tasks. RegionMIR is compared with different pretraining

Table 9.1: Quantitative Performance on Anatomy Classification

Pretraining	Finetuning	Classification Accuracy (%)
×	×	92.24
BioVIL	✓	93.01
RegionMIR (Co-train)	×	87.94
RegionMIR	✓	94.12

framework, including BioVIL, and different training scenarios such as co-training downstream task. For fair comparison, we report the performance of these methods under 50 epochs finetuning on anatomy classification.

9.5 Results

Anatomy Classification: From Table 9.1, we demonstrate the performance quantitatively across different scenarios. By training from scratch, the classification accuracy reaches 92.21 without contrastive learning strategies. By using transfer learning from BioViL pretrained weights, the classification accuracy significantly improves from 92.21 to 93.13. Furthermore, we perform the semantic region-guided contrastive learning in both pretraining and co-training scenarios. Interestingly, the performance further improved from 94.14 using pretraining scenario, while the performance significantly degraded from 87.94 using co-training strategies.

Image Retrieval: Figure 9.2, we demonstrate the qualitative representation of image retrieval with multiple anatomic queries. We observe that RegionMIR can successfully retrieve the images with similar anatomical regions. Interestingly, significant morphological differences are demonstrated across the images retrieved. Furthermore, we compute two plots to demonstrates the semantic meanings of the learned representations with our framework (right) and evaluate the similarity between the query representations and the anchor database (left). We found that the learned representations are well separable into clusters with anatomical meanings. For instance, the representations for both left (pink) and right lungs (blue) demonstrate separable clusters with the similarity measures above 0.9, while the clusters with small overlapping regions demonstrate a degradation in similarity measures.

9.6 Discussions

In figure 1, we have compared the finetuning classification performance by adapting different contrastive pretraining strategies, such as BioVIL. BioVIL adapts multimodal context (e.g. image & text) for contrastive learning and aligns representations with different semantics, including anatomy and pathologies. We leverage BioVIL as one of our baseline scenario and demonstrate a consistent improvement in classification accuracy

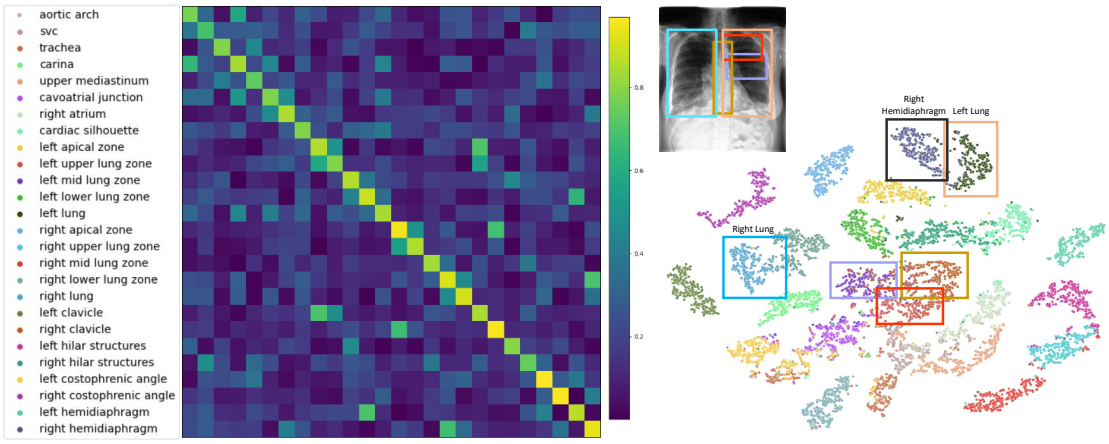


Figure 9.3: Left: similarity matrix for anatomical regions of interest, right: TSNE plot with RegionMIR-defined latent space. The order of labels for both the similarity matrix and TSNE plot are aligned. Separable clusters (e.g. left & right lung) demonstrate high cosine similarity (above 0.9), while clusters with small overlapping region demonstrate a lower cosine similarity value.

(92.24 to 93.01). However, the feature generated from BioVIL is from complete image and the limitation of adapting regional representation still exists in BioVIL. With RegionMIR, the classification performance further improve to 94.12. RegionMIR leverage the bounding box label to specifically pool the local representations and learned the representations with anatomical-meanings alignment. As we define pairwise representations in same anatomy as positive pair, we hypothesize that the latent space is initially defined with anatomy-specific clusters. We further classify representations into clusters with anatomy classification task as model finetuning. Each cluster is well separable with its corresponding anatomical meanings, as shown in Figure 9.3. For image retrieval scenario, we generate large scale features from training samples as our anchor database, while testing samples are used as the anatomic query. We propose to train K-mean models with anatomy-specific features and search the nearest unsupervised centroid for image recommendations (in Figure 9.2), which reduce the image retrieval time complexity significantly. The main goal of performing hierarchical search with K-means is to further differentiate the pathology characteristics in the chosen ROI and retrieve images with similar conditions within the ROI. Hierarchical adapting multiple semantic conditions into the contrastive learning framework for region-based image retrieval will be one of our future direction.

Although RegionMIR have demonstrated the feasibility of adapting regional information for image retrieval, several limitations still exist for the image retrieval task. The first limitation is the dependency on the pseudo prediction of anatomy classification. During the image retrieval, a mismatch K-means model may be chosen due to the inaccurate anatomy prediction, leading to the inaccurate searching criteria and retrieve images with different anatomical focus. Another limitations is the conditional constrains in both contrastive pretraining and the anatomy classification tasks. From Figure 9.2, we observe that the retrieved

image has the corresponding ROI with different morphologies. Such observation may correspond to the definition of positive pair in the contrastive pretraining step. As we randomly sample two images and define their corresponding anatomical feature as the only positive pair, the random sampled image may have significant morphological difference in certain anatomies. Additional constrains for contrastive pretraining will be another potential direction to include multiple constrains for region-based image retrieval.

9.7 Conclusion

We presented RegionMIR, to the best of our knowledge this is the first region-based contrastive learning framework for image retrieval with region query. We extend the image-wise contrastive loss into semantic region-based setting and generate an latent space with anatomy-semantic meanings. Furthermore, we leverage such latent space to retrieve images with similar region query and demonstrate the feasibility of performing region-based image retrieval with significantly reduced time complexity in medical domain. RegionMIR demonstrates a significant improvement on anatomy classification comparing to different pretraining and contrastive learning framework. The latent space of RegionMIR is demonstrated to successfully retrieve images with similar anatomical regions with different morphologies.

CHAPTER 10

Multi-contrast computed tomography healthy kidney atlas

10.1 Overview

[1]The construction of three-dimensional multi-modal tissue maps provides an opportunity to spur interdisciplinary innovations across temporal and spatial scales through information integration. While the preponderance of effort is allocated to the cellular level and explore the changes in cell interactions and organizations, contextualizing findings within organs and systems is essential to visualize and interpret higher resolution linkage across scales. There is a substantial normal variation of kidney morphometry and appearance across body size, sex, and imaging protocols in abdominal computed tomography (CT). A volumetric atlas framework is needed to integrate and visualize the variability across scales. However, there is no abdominal and retroperitoneal organs atlas framework for multi-contrast CT. Hence, we proposed a high-resolution CT retroperitoneal atlas specifically optimized for the kidney organ across non-contrast CT and early arterial, late arterial, venous and delayed contrast-enhanced CT. We introduce a deep learning-based volume interest extraction method by localizing the 2D slices with a representative score and crop within the range of the abdominal interest. An automated two-stage hierarchical registration pipeline is then performed to register abdominal volumes to a high-resolution CT atlas template with DEEDS affine and non-rigid registration. To generate and evaluate the atlas framework, multi-contrast modality CT scans of 500 subjects (without reported history of renal disease, age: 15-50 years, 250 males & 250 females) were processed. PDD-Net with affine registration achieved the best overall mean DICE for portal venous phase multi-organs label transfer with the registration pipeline (0.540 ± 0.275, p¡0.0001 Wilcoxon signed-rank test) comparing to the other registration tools. It also demonstrated the best performance with the median DICE over 0.8 in transferring the kidney information to the atlas space. DEEDS perform constantly with stable transferring performance in all phases average mapping including significant clear boundary of kidneys with contrastive characteristics, while PDD-Net only demonstrates a stable kidney registration in the average mapping of early and late arterial, and portal venous phase. The variance mappings demonstrate the low intensity variance in the kidney regions with DEEDS across all contrast phases and with PDD-Net across late arterial and portal venous phase. We demonstrate a stable generalizability of the atlas template for integrating the normal kidney variation from small to large, across contrast modalities and populations with great variability of demographics. The linkage of atlas and demographics provided a better understanding of the variation of kidney anatomy

[1]Published at: Lee, Ho Hin, et al., "Multi-contrast computed tomography healthy kidney atlas.", Computers in Biology and Medicine 146 (2022): 105555. (132)

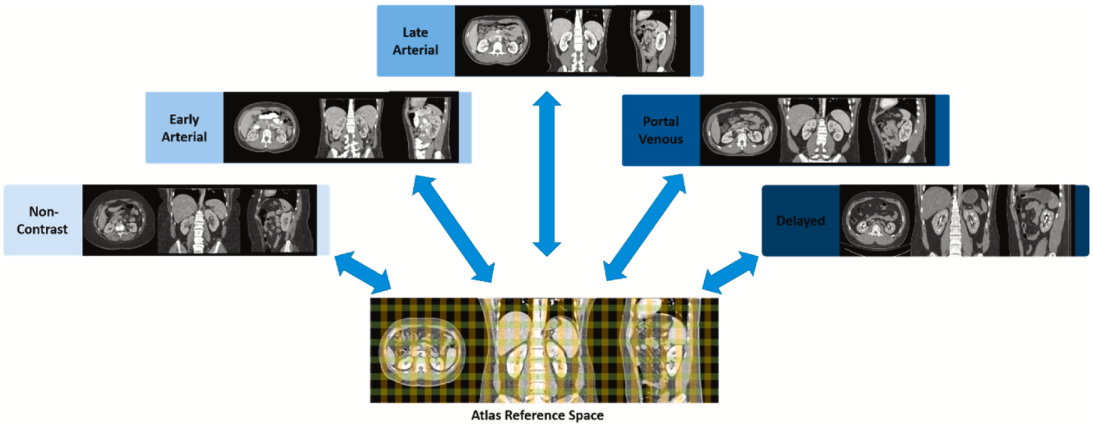


Figure 10.1: Illustration of multi-contrast phase CT atlas. The color grid in the three-dimensional atlas space represents the defined spatial reference for the abdominal-to- retroperitoneal volume of interest and localizes abdominal and retroperitoneal organs with each contrast phase characteristics. Blue arrows represent the bi-directional transformation across the atlas target defined spatial reference and the original source image space.

across populations.

10.2 Introduction

Physiological and metabolic processes are performed in parallel in the human body. Complicated relationships between cells are required for proper organ function but are challenging to analyze. Extensive studies mapping the organization and molecular profiles of cells within tissues or organs are needed across the human body. While the majority of efforts are distributed to the cellular and molecular perspectives (183), generalizing information from cell to organ level is essential to provide a better understanding of the functionality and linkage across scales (119). The use of computed tomography provides an opportunity to contextualize the anatomical characteristics of organs and systems in the human body. In addition, by creating a generalizable framework with integration of micro-scale information and system-scale information, this will provide clinicians and researchers the ability to visualize the complex organization of tissues from a cellular to tissue level. Abdominal CT provides information about abdominal organs at a system scale. Contrast enhancement demonstrates additional anatomical and structural details of organs and neighboring vessels by injecting a contrast agent before imaging procedures. Five different contrast phases are typically generated corresponding to the timing of the contrast agent in the imaging cycle: 1) non-contrast, 2) early arterial, 3) late arterial, 4)

portal venous and 5) delayed. The intensity range of organs fluctuates across the contrast enhanced imaging cycle and the variation of intensity helps to capture and specify contextual features of each specific organ. The kidneys, which are located retroperitoneally, also have challenges in imaging. From the anatomical information provided from the contrast enhanced CT of large clinical cohorts, healthy kidney morphometry and appearance may vary. An atlas reference framework is needed to generalize the anatomical and contextual features across the variations in sex, body size, and imaging protocols. However, due to the large variability in anatomy and morphology of various abdominal and retroperitoneal organs, generating a standard reference template for each of these organs is still challenging and no atlas framework for abdominal or retroperitoneal organs is currently publicly available. Creating an atlas for particular anatomical regions has widely been used with magnetic resonance imaging (MRI). Extensive efforts are allocated in multiple perspectives of brain atlas with brain MRI. Kuklisova-Murgasova et al. generated multiple atlases for early developing babies with age ranging from 29 to 44 weeks using affine registration (113), while Shi et al. proposed an infant brain atlas using unbiased group-wise registration with three varying scanning time points of brain MRI from 56 males and 39 females normal infants (190). Unbiased spatiotemporal 4-dimensional MRI atlas and time-variable longitudinal MRI atlas for infant brain were also proposed by Ali et al (64). with diffeomorphic deformable registration with kernel regression in age, and by Yuyao et al. using patch-based registration in spatial-temporal wavelet domain respectively (244). Apart from generating atlas framework for normal brain, Rajashekar et al. generated two high-resolution normative disease-related brain atlases in FLAIR MRI and non-contrast CT modalities, to investigate lesion-related diseases such as stroke and multiple sclerosis in elderly population (175). Meanwhile, limited studies have proposed creating a standard reference framework for abdominal and retroperitoneal organs. Development of abdominal and retroperitoneal organ atlases is challenging across multi-modality images (CT and MRI) due to the limited robustness of the registration methods and significant morphology variations in multiple organs associated with patients' demographics (232). To generate an atlas framework with stable context transfer, previous works have sought to improve the registration performance and the conventional frameworks perform spatial alignments between one image to another following with a global affine transformation and deformable transformation (3; 5; 10). Significant effort has been invested in brain imaging by optimizing a regularized deformation field to match the moving image to a single fixed target with classical approach such as b-spline based deformation (184), discrete optimization (43; 67), Demons (212), and Symmetric normalization (5). Apart from the traditional approaches, learning-based methods provide an opportunity to reduce the time of inference and extract meaningful feature representation for deformation prediction. Voxelmorph provides an opportunity to learn a generalizable function to compute deformation field with unsupervised setting and localize the deformation with few millimeters (11; 42). For abdominal imaging, significant variation in body size and shape of organs are demonstrated with

the several centimeters and large deformation for organs interest is needed to transfer the anatomical context to match the morphology of the fixed image. Zhao et al. proposed a recursive cascaded network to extract the region of interest (ROI) of particular organs and progressively refine the intermediate registered image to the fixed image space with multiple cascaded models (246). To further extend the abdominal registration with complete abdominal volumetric scans, Heinrich et al. proposed a probabilistic dense displacement network to adapt the large anatomical difference with organ label supervision (249). This network aims to improve the robustness of the registration performance in the abdominal regions with limited label guidance and capture the deformation with lightweight feature representation.

In this work, we present a contrast-preserving CT retroperitoneal atlas framework, optimized for healthy kidney organs with contrast-characterized variability and the generalizable features across a large population of clinical cohorts as shown in Figure 10.1. Specifically, as to reduce the failure rate of transferring the morphological and contrastive characteristics of kidney organs, we initially extracted the abdominal-to-retroperitoneal volume of interest with a similar field of view to the atlas target image, using a deep neural network called body part regression (BPR) (198). 2D slices of the CT volume assessed with BPR model and generate a value ranging from -12 to +12, corresponding to the upper lung region and the pelvis region respectively in the body. By limiting the range of values for both abdominal and retroperitoneal regions, each CT volume is cropped and excludes other regions apart from the abdomen and retroperitoneum, such as the lung and pelvis. A two-stage hierarchical registration pipeline is then performed, registering the extracted volume interest to the high-resolution atlas target with traditional metric-based (83; 84) and deep learning based (11; 85) method respectively across all contrast phases. To ensure the stability and the variation localized in the atlas template, average and variance mappings across the multi-contrast registered output is computed to demonstrate a better understanding of anatomical details of kidney organs across different contrasts. Overall, our main contributions are summarized as:

- We constructed the first multi-contrast CT healthy kidney atlas framework for the public usage domain.
- We proposed metric-based and deep learning-based framework optimizing for kidney organs with corresponding contrast phase, and generalized the anatomical context of kidneys with significant variation of morphological and contrastive characteristics across demographics and imaging protocols.
- We evaluate the generalizability of the atlas template by transferring the atlas target label to the 13 organs well-annotated CT space with inverse transformation. An unlabeled multi-contrast phase CT cohort is used to compute average and variance mapping to demonstrate the effectiveness and stability of the proposed atlas framework. Our proposed atlas framework demonstrates a stable transfer ability in both left and right kidneys with median Dice above 0.8.

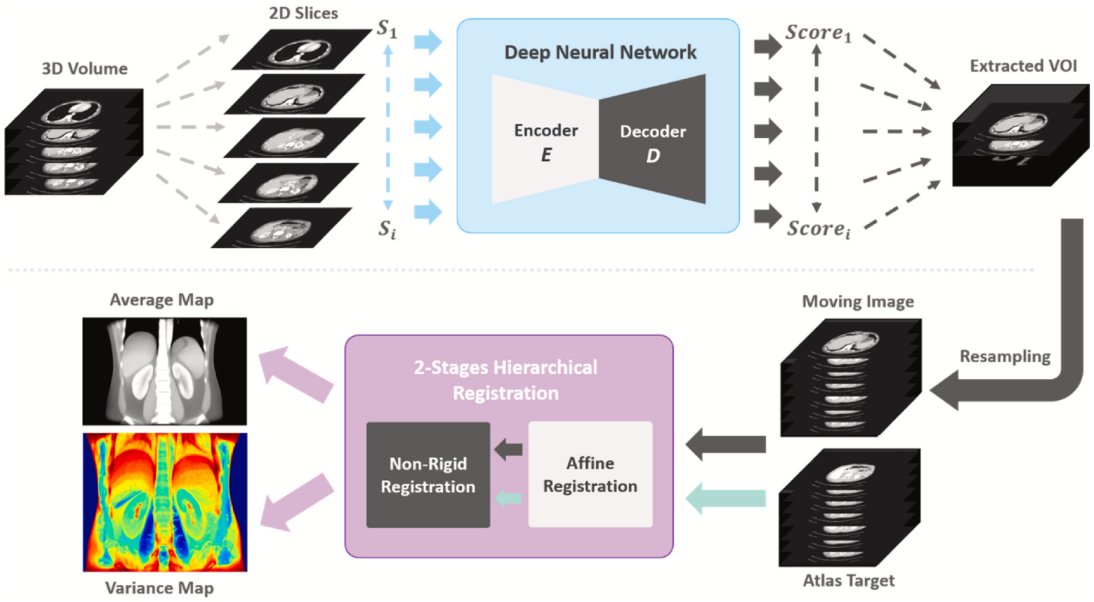


Figure 10.2: The overview of the complete pipeline to create kidney atlas template is illustrated. The input volume is initially cropped to a similar field of view with the atlas target. The extracted volumes of interest are resampled to the same resolution and dimension of the atlas template and performed 2-stages hierarchical registration. The successfully registered scans are finally used to compute the average template and the variance maps.

- • The average template generated, and the associated 13 organs labels will be public for usage through HuBMAP.

10.3 Materials and Methods

Figure 10.2 presents an overview of the complete pipeline for generating the kidney atlas framework. The volume of interest is first extracted with a deep learning-based BPR algorithm to obtain a similar field of view with the atlas target image, increasing the stability of registration in the abdominal body and kidneys. Here, we define the stability as the atlas not changing with randomized subjects and measure the stability with the mean or variance mapping of the atlas template. The integration of deep learning-based volume of interest extraction and classical registration provided an opportunity to reduce the subjectivity of choosing the field of view between source and target image, and increase the robustness of the image registration across the clinical cohorts.

10.3.1 Datasets of Studies

We evaluated the stability of our kidney atlas with a large cohort of multi-contrast unlabeled CT and a public portal venous contrast phase multi-organ labeled dataset. All CT data are unpaired and collected from dif-

ferent cohorts. With the use of both labeled and unlabeled datasets, we conducted comprehensive qualitative and quantitative evaluations for generalizing cross-modality information of kidney organs in CT. **Clinical Multi-Contrast Abdominal CT Cohort:** A large clinical cohort of multi-contrast CT was employed for abdominal and retroperitoneal organs registration. In total, 2000 patients' de-identified CT data were initially retrieved in de-identified form from ImageVU with the approval of Institutional Review Board (IRB). In these 2000 patients, since some had renal disease, criteria in ICD-9 codes and age range from 18-50 years old were set and applied to extract scans with healthy kidneys from all subjects. 720 subjects out of 2000 were identified after quality assessment and extract the corresponding contrast phase abdominal CT scans, which included 290 unlabeled CT volumes in total with: 1) non-contrast: 50 volumes, 2) early arterial: 30, 3) late arterial: 80 volumes, 4) portal venous: 100 volumes, 5) delayed: 30 volumes. All CT volumes are initially reoriented to standard orientation before further processing (101). BPR was performed to each modality volumes and obtain the similar field of view with the atlas target. They were then resampled to the same resolution and dimensions with the atlas target for performing registration pipeline. We aim to adapt a generalized atlas framework for localizing the anatomical and contextual characteristics of kidney organs across multi-contrast. **Multi-Organ Labeled Portal Venous Abdomional CT Cohort:** We used a separate healthy clinical cohort with 100 portal venous contrast phase abdominal CT volumes and 20 of the volumetric scans are the testing scans in the 2015 MICCAI Multi-Atlas Abdomen Labeling challenge. The ground truth labels of 13 multiple organs are provided including: 1) spleen, 2) right kidney, 3) left kidney, 4) gall bladder, 5) esophagus, 6) liver, 7) stomach, 8) aorta, 9) inferior vena cava (IVC), 10) portal splenic vein (PSV), 11) pancreas, 12) right adrenal gland (RAD), 13) left adrenal gland (LAD), with which we conduct label transfer on this dataset to evaluate the generalizability and stability of the atlas template. In order to reduce the number of failed registrations to the atlas target, BPR is performed on abdominal and retroperitoneal regions only with soft-tissue window. To evaluate the atlas framework, the inverse transformation was applied to the multi-organ atlas label, and labels were transferred back to the spatial space of each portal venous phase CT. **High Resolution Single Subject Atlas Template:** We choose the single subject atlas template with several conditions: 1) high-resolution characteristics, 2) significant appearance in kidney morphology and 3) clear kidney boundary with contrast. The atlas template is provided by Human Biomolecular Atlas Program (HuBMAP) with high resolution setting of $0.8 \times 0.8 \times 0.8$ and healthy condition. The dimension of the atlas subject is $512 \times 512 \times 434$ with 13 Organs annotated by joint label fusion from the registered subjects. The volumetric measure of both left and right kidneys are 256 cc.

10.3.2 Deep Body Part Regression Network

The use of deep learning in the medical imaging domain contributed to a great increase in automatic models for classification and segmentation. Due to the shift of various domains and the variation of imaging protocols, medical images usually present with different visual appearances and fields of view. The goal of the body part regression network (BPR) is to narrow the difference of field of view between the source images and the atlas target image for reducing the failure rate of registration. Formally, given an unlabeled dataset $\{x_i^m\}_{i=1}^N$ from the moving image domain, and a labeled dataset $\{x_i^a, y_i^a\}_{i=1}^1$ from the atlas target domain, we aim to crop the original volume of interest x^m to an approximate field of view with the atlas target x^a. The obtained volume of interest only consists of abdomen regions and is resampled to the same voxel resolution as the atlas target. Tang et al. proposed a self-supervised method to predict a continuous score for each axial slice of CT volumes as the normalized body coordinate value without any labels (234; 198). The self-supervised model predicts scores in the range of -12 to +12 and each body part region corresponds well to an approximate score (e.g., -12: upper chest, -5: diaphragm / upper liver, 4: lower retroperitoneum, 6: pelvis). Linear regression is performed to correct the discontinuity of the predicted score, and we use the regressed output as the self-supervised label to train a new refined model. Both the atlas target image and the unlabeled dataset are input into the well-trained model and compute scores for each slice of the volume. To extract the abdominal-to-retroperitoneal regions only for each dataset, we limit the slices with a range of scores from -5 to 5 and crop the slices that are out of this range. All unlabeled datasets x^m are then enforced to have a closer field of view to the atlas target image x^a.

10.3.3 Two-Stage Hierarchical Metric-Based Registration Pipeline

The metric-based registration pipeline is composed of 2 hierarchal stages: 1) affine registration and 2) non-rigid registration. Dense displacement sampling (DEEDS) is a 3D medical registration tool with a discretized sampling space that has been shown to yield a great performance in abdominal and retroperitoneal organs registration, is used for both affine non-rigid registration in this pipeline (232; 83; 84). The DEEDS affine registration is first performed to initially align both moving images and the atlas target to preserve 12 degrees of freedom of transformation and provide a prior definition of the spatial context and each affine component. An affine transformation matrix is generated as the output and become the second stage non-rigid registrations' input. The DEEDS non-rigid registration is refined with the spatial context as the local voxel-wise correspondence with its specific similarity metric, which will be illustrated below. Five different scale levels are used with grid spacing ranging from 8 to 4 voxels to extract patches and displacement search radii from six to two steps between 5 and 1 voxels (232; 83; 84; 82). Deformed scans with the displacement data from selecting control points is generated and transfer the source image space voxel information to the atlas target

space after deformation. To ensure the stability of the atlas generated, all successfully deformed scans are averaged and variance mapping is used to visualize the intensity fluctuation and variation around the abdominal body and kidney organs. The similarity metric defined in DEEDS registration tool is self-similarity context with the patches extracted from moving images. Such a similarity metric aims to find a similar context around neighboring voxels in patches (83). The self-similarity metric S is optimizing a distance function D between the image patches extracted from the moving image M. A function q^2 is computed to estimate both the noise in local and global perspectives. Meanwhile, a certain number of neighborhood relations N is also defined as to determine the kinds of self-similarities in the neighborhood. As extracting an image patch with a center at x, the measurement calculation of the self-similarity can be demonstrated as follows:

$$D(M,x,y) = \exp(\frac{S(x,y)}{q^2}),\ x,y \in N \tag{10.1}$$

where y is defined as the center of another patch from one of the neighborhood N. This similarity metric helps to avoid the negative influence of image artifacts or random noise from the central patch extracted and prevent a direct adverse effect in calculation. Twelve distances between pairwise patches are calculated within six neighborhoods and concentrate in extracting the contextual neighboring information, instead of the direct shape representation.

10.3.4 Experimental Settings

In the preprocessing stage, the BPR model with U-Net like architecture is pretrained with a total of 230,625 2D slices from a large-scale cohorts of 1030 whole body CT scans from the public domain (198). The portal venous multi-organ labeled cohorts are used only for external validation. The pretrained model is trained from scratch and optimized with Adam using learning rate of 0.0001. The batch size is 4 for end-to-end training.

For the atlas generation process, we investigate multiple registration methods including traditional registration tools and deep learning registration methods to compare the robustness of registration across all contrast phases scans in qualitative and quantitative perspective. We performed comprehensive analysis which included ANTS (232; 5), NIFTYREG (NIFTYR) (232; 155) and DEEDS (232; 83; 84; 82) as the traditional medical image registration tools, while VoxelMorph (11) and PDD-Net (81) are used as deep learning-based registration framework on our portal venous dataset. For metric-based registration, the preprocessed moving scans are upsampled to the same resolution with the preprocessed atlas target and directly performed hierarchical affine and deformable registration in high resolution setting. For deep learning-based method, we downsampled the moving scans and the atlas target to a resolution of $1.5 \times 1.5 \times 1.5$ with dimension

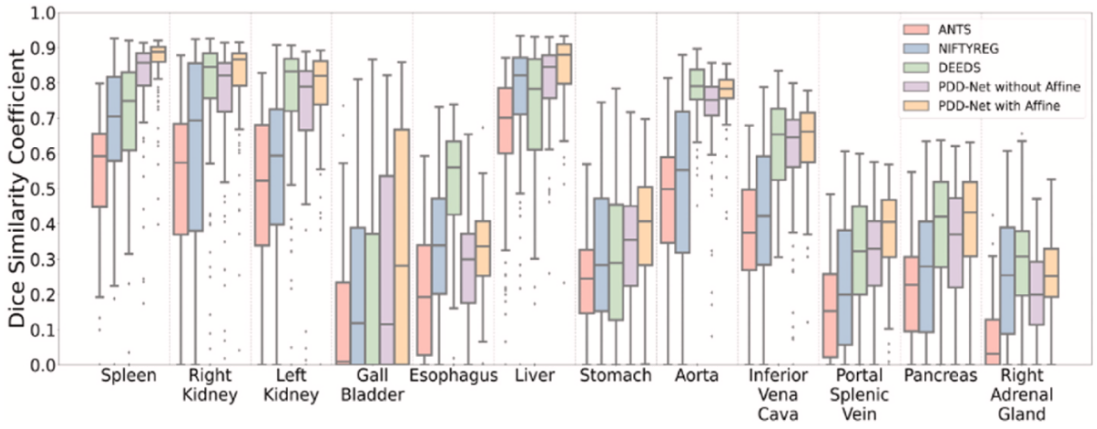


Figure 10.3: The quantitative representaiton of label transfer with multi-organ portal venous dataset are demonstrated that PDD-Net with affine registration outperforms the other four traditional methods in an organ-wise manner. Significant increase of Dice is also demonstrated with medical Dice over 0.8 in the transferring result of both left and right kidneys using PDD-Net with improving outliers.

$192 \times 160 \times 256$. Affine registration is performed with NiftyReg to coarse align the moving scans and the atlas fixed target before non-linear mapping prediction with deep network pipelines. VoxelMorph is trained with the multi-organ labeled portal venous scans as the moving scans and the atlas target as the fixed scans in unsupervised setting with 100 epochs and learning rate of 0.0001. We use the pretrained model of PDD-Net from (81) and directly predict a 4D displacement field for non-linear transformation. The pretrained model of PDD-Net is trained with 3-fold cross validations on 10 contrast-enhanced CT scans of the VISCERAL3 training data with learning rate of 0.01 for 1500 iterations (105). The training process of both VoxelMorph and PDD-Net are optimized end-to-end with Adam and batch size of 1. The predicted deformation fields from deep learning based framework are finally upsampled to the atlas template resolution and compute deformation warp with non-linear transformation in high resolution setting.

For ablation studies, we further investigate the effectiveness on the BPR preprocessing and located the affect of the significant difference of field of view for registration with variance mapping. As there is significant intensity variation across each organ interests in each phase, we investigate the effect of domain shift for the robustness of the registration pipeline and the label transfer of non-contrast phase dataset is further evaluated to demonstrate the registration stability with significant domain shift across metric-based to deep learning-based framework.

Table 10.1: Evaluation metric on 100 portal venous registration on 13 Organs (Mean±STD), note that p<0.0001 with Wilcoxon signed-rank test ⋆, A: affine registration only.

Methods	Dice Score	MSD (mm)	HD (mm)
ANTS (A)	0.246 ± 0.224	12.8 ± 10.2	60.2 ± 47.5
NIFTYR (A)	0.270 ± 0.222	13.1 ± 11.1	62.1 ± 50.1
DEEDS (A)	0.200 ± 0.199	19.6 ± 18.2	80.9 ± 59.3
ANTS	0.319 ± 0.252	10.1 ± 9.00	51.6 ± 45.3
NIFTYR	0.406 ± 0.279	10.9 ± 13.8	55.0 ± 51.8
DEEDS	0.496 ± 0.284	8.52 ± 17.1	41.6 ± 51.2
VoxelMorph without A	0.335 ± 0.275	12.5 ± 10.9	34.9 ± 21.7
VoxelMorph with A	0.435 ± 0.283	9.27 ± 9.47	30.8 ± 23.5
PDD-Net without A	0.486 ± 0.286	8.47 ± 9.87	31.0 ± 25.0
PDD-Net with A ⋆	0.540 ± 0.275	6.92 ± 8.73	28.5 ± 25.4

10.3.5 Evaluation Metric

We employed three commonly used metrics to evaluate the similarity between the prediction label from automatic models and the original ground truth label: 1) Dice score, 2) mean surface distance (MSD), and 3) Hausdorff distance (HD). The definition of Dice is to measure the overlapping of volume between the segmentation label prediction and the ground truth segmentation label:

$$DICE(P,G) = \frac{2|P \cap G|}{|P| + |G|} \tag{10.2}$$

where P is the predicted label from models and G is the original ground truth segmentation label, while $\|$ is the L1 norm operation.

The rendered surface is another perspective which used to evaluate result. The 3-dimensional coordinates of vertices were initially extracted from both the prediction label and the ground truth label. The average distance and the Hausdorff distance between the sets of vertices coordinate were calculated as follows:

$$MSD(V_p, V_g) = avg\,inf\,Dist(V_p, V_g)$$

$$HD(V_p, V_g) = sup\,inf\,Dist(V_p, V_g)$$

where V_p and V_g represents the vertices coordinates of prediction label and ground truth label respectively, while avg refers to average, and sup and inf refers to the greatest lower bound and least upper bound of the distance function measure $Dist$.

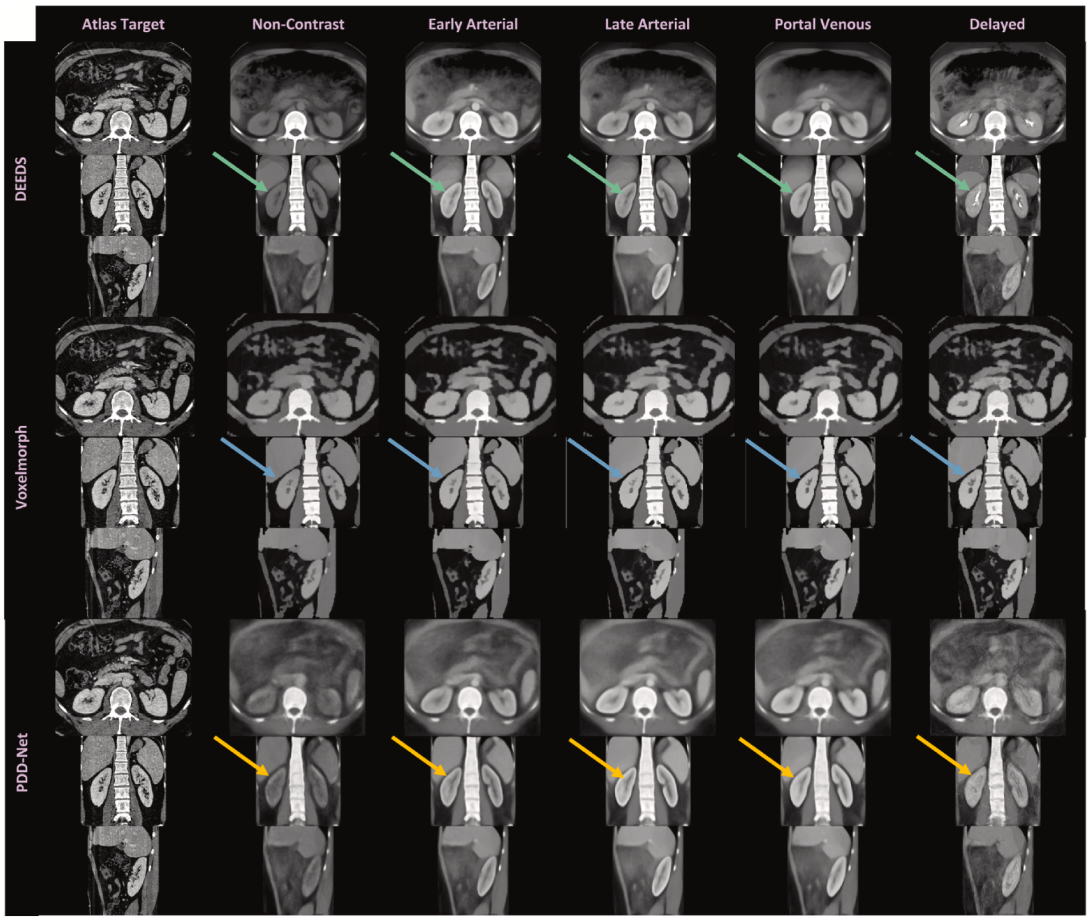


Figure 10.4: We investigate the registration stability across all contrast phase with average mapping. The metric-based registration representative DEEDS demonstrates a stable transfer of the kidney contextual findings across all phases, while the deep learning representative PDD-Net illustrates sharp appearance in the kidney sub-structural context in contrast-enhanced phase such as late arterial and portal venous, but with unstable registration appearances in non-contrast and delayed phase. We additionally compare the average mapping from the VoxelMorph. It is limited to transfer the sub-structural contrast characterisitics and preserve the boundary of kidney organs well compared with DEEDs and PDD-Net. (See arrows).

10.4 Results

Registrations were performed to the atlas target image with the pre-processed cropped volume of interest. We performed five registration methods with portal venous cohorts to ensure the highest accuracy for the localization of both left and right kidneys. The quantitative representation of each organ in the registered output was then demonstrated in terms of Dice score, MSD and HD, and illustrated the distribution of the performance across the multi-organ labeled portal venous cohort with Figure 10.3 and Table 10.1. As shown in Table 10.1, the deep registration pipeline PDD-Net with affine registration from NiftyReg achieved the best overall mean Dice across all 13 organs. From the demonstration of Figure 10.3, NiftyReg demonstrated a better performance in registering liver organ comparing to DEEDS, while ANTs performed registration with

inferiority across all organs comparing to the other two methods. Apart from comparing with metric-based methods, we also performed deep learning registration baseline with VoxelMorph and characterize the efficiency of the registration with deep networks. The performance of PDD-Net outperforms VoxelMorph with 24.2% Dice across 13 organs by training from scratch with both with or without initial affine registration scenario. In terms of optimizing for kidney organs with the atlas template, the Dice score of both left and right kidneys are separately computed to obtain the ability of kidneys localization with the atlas template. Both PDD-Net and DEEDS have a comparable performance in the kidney regions and outperforms the registration performance of kidney with NiftyReg and ANTs. PDD-Net demonstrate significant improvement in right kidneys registrations, while DEEDS perform better with a small extent in transferring the left kidney regions. NiftyReg and ANTs illustrated the lack of generalizability of transferring kidney organs and computed significant variance of Dice across all registrations. The reduction of variance and population of outliers are shown from boxplots with PDD-Net, leading to a significant improvement of Dice score, MSD, and HD for both left and right kidneys comparing to the other two methods. The Wilcoxon signed-rank test showed that PDD-Net was significantly better (p ¡ 0.0001) than all other methods in Dice (178). The median Dice of both transferred left and right kidney using DEEDS and PDD-Net are above 0.8, while it is a significant boost comparing with the other registration pipelines.

Apart from the quantitative result, we compare the qualitative representations of DEEDS and PDD-Net registrations across multiple contrast phases and shown in terms of average template and variance mapping in Figure 10.4 and Figure 10.5. The average mapping of each contrast phase was then computed with all registered contrast-corresponding volumes. With DEEDS, the contrast and anatomical context of kidney regions in each phase are stably transferred to the atlas space, while other organs' regions such as liver and spleen, are demonstrated with blurry appearance. With PDD-Net, the contextual representations of the kidney regions are demonstrated with sharper appearance comparing with the DEEDS average template in early arterial, late arterial and portal venous phase. The anatomy of kidney sub-structure and renal-related vessels can appear with sharpness in the average template of early arterial and late arterial phase. However, the non-contrast and delayed phase template demonstrate an unclear structure of the kidney anatomy with PDD-Net. To further ensure the stability of transferring the kidneys' anatomical information, variance maps of each contrast phase template with corresponding registration methods are also computed to demonstrate the voxel variability of each organ across the clinical cohort. With DEEDS registration, the small variation in the kidney is illustrated with a color range from yellow to green, while significant variation in voxels is shown near the diaphragm region and the color range from orange to red indicated the highly deformed variability across the registered outputs. For PDD-Net, the boundary around the kidney regions demonstrated significant variation across each contrast phase, while the intensity variance is small in the renal cortex regions

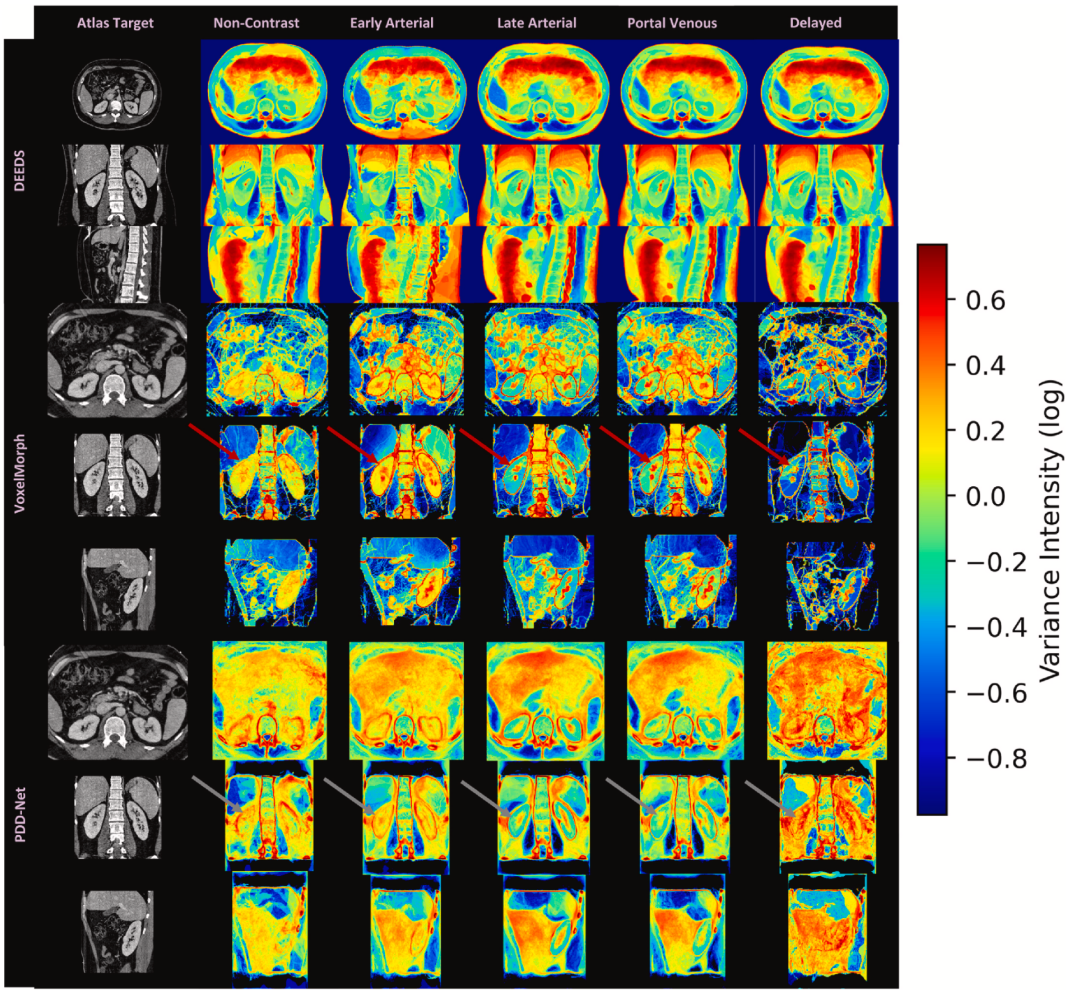


Figure 10.5: We further evaluate the intensity variance across the registration outputs with the average template in each contrast phase. The variance mapping of DEEDS demonstrates the kidney context transferal with stability and the variance value near the kidney region is 0–0.15, while significant variance are localized in the boundary of the kidney region with the variance mapping of PDD-Net and VoxelMorph. For late arterial and portal venous phase, PDD-Net well preserved the core context of the renal cortex region. However, unstable registrations are demonstrated with the high variance value in kidney with non-contrast and delayed phase mapping (see arrows), which match the blurry appearance of kidney regions in the average mapping.

in late arterial and portal venous phase. The variance map of non-contrast and delayed phase demonstrate the significant variability across the contrast intensity and the kidney anatomy, which correlated the instability of the registration of both phase subject scans. Overall, contrast phase context can be stability preserved with DEEDS in the kidney regions and PDD-Net demonstrated a higher robustness of transferring the kidney findings with sharp appearance in late arterial and portal venous phase.

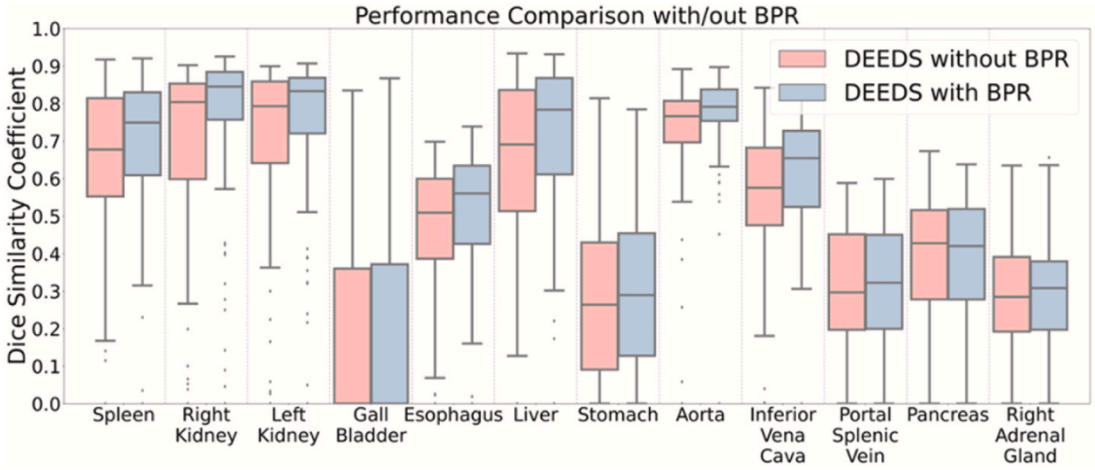


Figure 10.6: We evaluate the failure rate of the registration with/out using BPR. The use of BPR reduce the number of outliers in some of organs, especially for right and left kidneys.

10.5 Discussion

In this study, we constructed a healthy kidney atlas for five different contrast phases CT and generalized the kidney anatomical context across population demographics and the variation of contrast characteristics. High variance score is located near the diaphragm regions and the boundary of the abdominal body (see Fig. 5). Such variability is contributed to the large deformation of the lower and upper boundary of the volume interest and transfer specific organs' contextual information to other organs' anatomical locations. The low variance score in both the left and right kidney region indicated the stable registration of structural information, as the kidney organs are localized in the middle region of the volume interest with contrast. The surface rendering of the kidney across the morphological sizes visualizes the generalizability of the atlas template across shape variability (see Figure 10.5). We adapt a 2D color-space checkerboard to visualize the deformation on the surface. The color of each box in the checkerboard pattern changes both horizontally and vertically. The color-boxes in atlas space are equivalent to that in the inverse original space. The smoothnesss of the registration is qualitatively defined by the movement of the checkerboard pattern. If there is significant movement of the colored pattern, there is significant deformation on that particular regions and the smoothness of deformation field is low. From Figure 10.7, stable deformation is demonstrated across small to large kidneys. The high-resolution characteristics of the atlas template preserve highly detailed voxel-wise information across all organs. DEEDS provided an overall performance with mean Dice of 0.50 (232). However, the DEEDS

Table 10.2: Time consumption of preprocessing and sampling data samples

Methods	Extraction Time per Scan (Sec)
Apply BPR	30.237
Downsampling for Deep Learning Registration	13.111
Upsampling to Atlas Resolution	124.071

performance cannot provide accurate measures towards the organ of interests due to high sensitivity of field of view, leading to significant deformation. With the use of PDD-Net (81), it yields the best performance with mean Dice of 0.54 across all organs. The sharpness of the organ interests' structure such as liver and spleen become more appealing and the substructure context of kidneys are also demonstrated with stability across the contrast-enhanced phase such as late arterial and portal venous. In addition to the ablation study, we evaluate the effectiveness of the BPR preprocessing for abdominal registration with DEEDS. As abdominal registration is sensitive to the similarity of the field of view between the fixed scans and the moving scans, the BPR algorithm provides an opportunity to extract approximate ranges within the abdominal region and allows a certain extent of deformation, while limited field of view is generated by cropping the ROI of the organs with the segmentation masks and may lead to over-deformation. We aim to increase the successful rate of registration with the BPR algorithm with a proportional increase of label transfer performance. Significant improvement is demonstrated with the use of BPR in Table 10.2 & 10.3 for both volumetric segmentation using 3D U-Net as network architecture (39) and image registration respectively. The cropped abdominal interest reduces the over-deformation of organs towards the lung or pelvis region. The variance in registration performance is also reduced with the decrease of outliers in Figure 10.6 and demonstrate the effect reduction by the significant field of view variability. Among the average template of each contrast phase in Figure 10.4, the portal venous phase average template provides a sharper and smoother abdominal body comparing to that of the other four phases. However, the portal venous phase average template is not suitable to be an unbiased template for registration. The small tissue voxels are difficult to represent and cannot provide sufficient accurate information for registration. Large population of clinical cohorts may need to be used to obtain a higher confidence level of voxel representation that is as sharp as the atlas target image.

Initially, one of the bottlenecks of providing a stable anatomical information transfer for organs is the registration method. The registration performance of kidneys may be affected by the secondary targets such as liver and spleen. Further optimization of the registration pipeline is needed for reducing the possibility of significant deformation. Instead of relying on similarity metric (mutual information, cross correlation, Hamming distances of the self-similarity context) as the loss function (212; 209), a learning-based method may be another promising direction to learn a registration function and predict the registration field for the

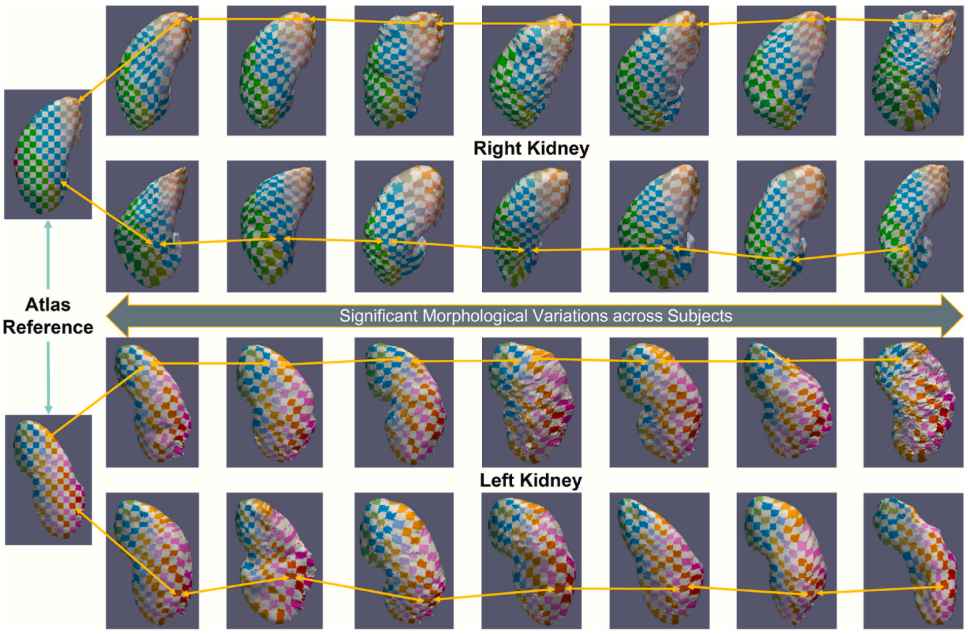


Figure 10.7: The surface rendering of the registered kidney with significant morphological variation are also illustrated. The 2D checkerboard pattern with arrows demonstrates the correspondence of the deformation from atlas space to the moving image space. A stable deformation across the change in volumetric morphology of kidney (100 cc∼308 cc) are demonstrated ewith the deformed checkerboard.

moving images (81; 156). However, most of the proposed learning-based pipeline is focused on the brain (11; 44). PDD-Net provide an opportunity to reduce the gap for robust abdominal region registration with deep neural networks to adapt large deformation field. However, the significant improvement of the registration is limited with contrast-enhanced dataset such as late arterial phase and portal venous phase scans as shown in Figure 10.4. We further perform the evaluation in registration performance with non-contrast dataset as Table 10.4 and a significant decrease of robustness in registration is illustrated with PDD-Net. As the pretrained model of PDD-Net is trained with limited portal venous dataset, the significant of domain shift contribute to the adverse performance of the registration. Multi-Modality registration with deep learning approaches can be the next step to contribute a robust generation of atlas with abdominal organs.

The kidney atlas template also provides contributions in the segmentation of other abdominal and retroperitoneal organs. The high-quality atlas multi-organ label can be transferred with the inverse transformation and use to provide accurate measures for abdominal and retroperitoneal organs. Also, high quality labels can help perform training strategies to innovate automatic learning-based model. Huo et al. proposed a whole brain segmentation using spatially localized network tiles with the atlas-transferred label (98). Dong et al. proposed left ventricle segmentation network, which integrate the ventricle atlas at echocardiogram into learning framework and provides consistency constraints with atlas label to perform accurate segmentations (51). Bai

Table 10.3: Time consumption of metric-based & deep learning-based registration methods

Methods	Extraction Time per Scan (Sec)
DEEDS (A)	170.105
DEEDS (D)	591.349
PDD-Net	**7.43**

Table 10.4: Evaluation metric on 50 Non-Contrast registration on 13 organs (Mean±STD), note that $p<0.0001$ with Wilcoxon signed-ranked test $\star$, A: affine registration.

Methods	Dice Score	MSD (mm)	HD (mm)
DEEDS $\star$	**0.485±0.275**	**9.45±18.2**	**43.4±48.5**
PDD-Net with A	0.278 ± 0.223	12.4 ± 15.6	59.4 ± 45.1

et al. presented a population study of relating the phenome-wide association to the function of cardiac and aortic structures using machine-learning-based segmentation pipeline (9). Clinical validation and phenotypic analysis can be performed and reveal biomarkers of specific organs in certain conditions such as disease pathogenesis, with high-quality segmentation labels. Further exploration can be investigated in the abdominal and retroperitoneal domain with the use of the high-quality atlas label.

10.6 Conclusion

This manuscript presents a healthy kidney atlas to generalize the contrastive and morphological characteristics across patients with significant variability in demographics and imaging protocols. Specifically, the healthy kidney atlas provides a stable reference standard for both left and right kidney organs in 3-dimensional space to transfer kidney information using an adapted registration pipeline. Significant variance on the field of view and the organ shape can be focused as the optimization parameters to reduce the possibility of failure registration. Potential future exploration with the use of the atlas template can be further investigated in both engineering and clinical perspectives, to provide better understandings and measures towards the kidneys.

CHAPTER 11

Supervised Deep Generation of High-Resolution Arterial Phase Computed Tomography Kidney Substructure Atlas

11.1 Overview

[1]The Human BioMolecular Atlas Program (HuBMAP) provides an opportunity to contextualize findings across cellular to organ systems levels. Constructing an atlas target is the primary endpoint for generalizing anatomical information across scales and populations. An initial target of HuBMAP is the kidney organ and arterial phase contrast-enhanced computed tomography (CT) provides distinctive appearance and anatomical context on the internal substructure of kidney organs such as renal context, medulla, and pelvicalyceal system. With the confounding effects of demographics and morphological characteristics of the kidney across large-scale imaging surveys, substantial variation is demonstrated with the internal substructure morphometry and the intensity contrast due to the variance of imaging protocols. Such variability increases the level of difficulty to localize the anatomical features of the kidney substructure in a well-defined spatial reference for clinical analysis. In order to stabilize the localization of kidney substructures in the context of this variability, we propose a high-resolution CT kidney substructure atlas template. Briefly, we introduce a deep learning preprocessing technique to extract the volumetric interest of the abdominal regions and further perform a deep supervised registration pipeline to stably adapt the anatomical context of the kidney internal substructure. To generate and evaluate the atlas template, arterial phase CT scans of 500 control subjects are de-identified and registered to the atlas template with a complete end-to-end pipeline. With stable registration to the abdominal wall and kidney organs, the internal substructure of both left and right kidneys are substantially localized in the high-resolution atlas space. The atlas average template successfully demonstrated the contextual details of the internal structure and was applicable to generalize the morphological variation of internal substructure across patients.

11.2 Introduction

Considerable effort has been made by the Human BioMolecular Atlas Program (HuBMAP) to relate molecular findings in organ anatomy across cellular to organ systems levels (183). With the previous efforts of mapping the organization and molecular to profile cells across different tissues and organs (119), it is vitally important to contextualize the anatomical characteristics of organs in well-defined reference templates to act

[1]Published at: Lee, Ho Hin, et al., "Supervised Deep Generation of High-Resolution Arterial Phase Computed Tomography Kidney Substructure Atlas.", Medical Imaging 2022: Image Processing. Vol. 12032. SPIE, 2022. (126)

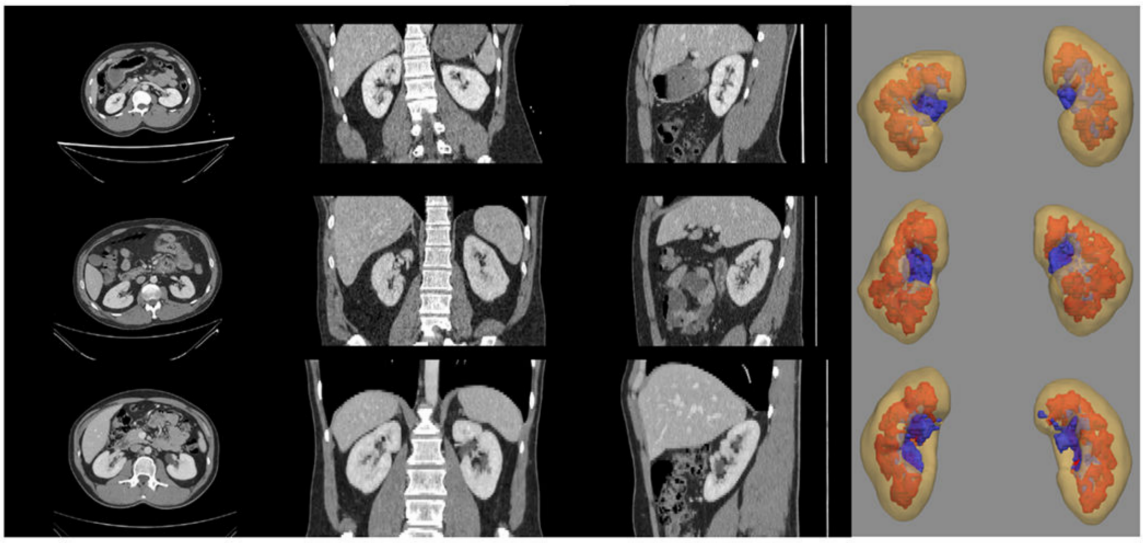


Figure 11.1: Significant variations in contrast intensity and mophology are demonstrated in both the external and internal structure of the kidney (1) yellow: reanl cortex, 2) red: medulla and 3) pelvicalyceal system). The asymmetric property in the kidney appearance and the anatomical (such as size and position) variation is shown and it is challenging to adapt a well-defined anatomical reference for kidney organs.

as a common framework for contextualizing multi-modal molecular information across scales. Computed tomography (CT) provides an imaging platform to visualize anatomical context at organ system level. Contrast enhancement, is used to emphasize the structural and anatomical context between neighboring organs with the injection of contrast agent and guide to extract information from regions of interest (ROI) (199). The kidney is the initial target of HuBMAP to analyze the structural anatomy across scales. Arterial phase CT provides distinctive representation for further analysis of the kidney substructure perspective (201). Figure 11.1 illustrates how the kidneys are localized with significant variability of size and anatomical regions. Additionally, the surface rendering of each internal structure demonstrates the large volumetric difference across patients from different demographics. It is challenging to adapt such variable morphological characteristics in a single well-defined reference for population analysis.

Here, we aim to adapt the conventional information of each kidney substructure on a single anatomical atlas template. Previous works have been demonstrated in building an atlas platform with neuroimaging (64; 165). Multiple brain atlases are built to reveal the population characteristics of brains in both adults and infants (100; 175; 113). Apart from looking into the anatomical characteristics, atlas reference is used as a platform to perform segmentation with unsupervised settings (216). Multiple atlas references are randomly picked and perform registration between the subject moving scans to the multiple atlases' platform (184).

Segmentation predictions are computed with joint label fusion using the guidance of multiple registered outputs. However, there has been more limited work in generating atlas frameworks for the specific organs in abdominal regions (43; 67). With the variability of demographics, several challenges are raised with 1) the difference of abdominal body shape and 2) the performance of registrations for adapting large deformation. Currently, registration pipelines are specifically designed for the abdominal regions to increase the robustness of adapting the abdominal organ-corresponding information such as DEEDS (135). With the DEEDS registration pipeline, over-deformation is demonstrated and leads to great variance in registering the liver and spleen regions due to the field of view varies across patients with different imaging protocols (131; 132). Therefore, a stable atlas construction pipeline is needed to generalize the organ context in the abdomen across demographics and construct a well-defined atlas target with a high successful rate of transferring ROI context needed for anatomical evaluation.

In this study, we construct a contrast-preserving kidney substructure atlas with a deep unsupervised preprocessing and registration pipeline to increase the robustness of adapting organ context across patients in high resolution. With a total of 500 arterial subject scans with healthy kidneys, we generated the average mapping of healthy kidney substructures across this population. Atlas target and moving subject scans are initially downsampled to input for the deep registration pipeline. The predicted displacement field is upsampled back to the atlas resolution and warped with the upsampled subject images. Registration performances are evaluated in both quantitative and qualitative perspectives and the kidney substructure context is stably transferred with contrastive and morphological characteristics to the atlas space.

11.3 Methods

11.3.1 Preprocessing

With demographic variability across patients, the field of view varies significantly in 3D abdominal CT scans across populations and may contribute to the possibility of registration failure. Here, we introduce a deep learning preprocessing pipeline body part regression to generate anatomical information for cropping the kidney-corresponding area of interest to perform stable registration (198; 130). Specifically, 2D axial slices are initially extracted from each scan and a prediction score is generated to identify the approximate anatomical location in the human body for each slice. The output value for each slice is in the range of -12 to + 12 (arbitrary units), which specify regions from the heart to the pelvis. We defined the value between -4 to 3 as the kidney regions of interest and crop the volumetric scans to ensure a similar field of view between atlas target and subject scans.

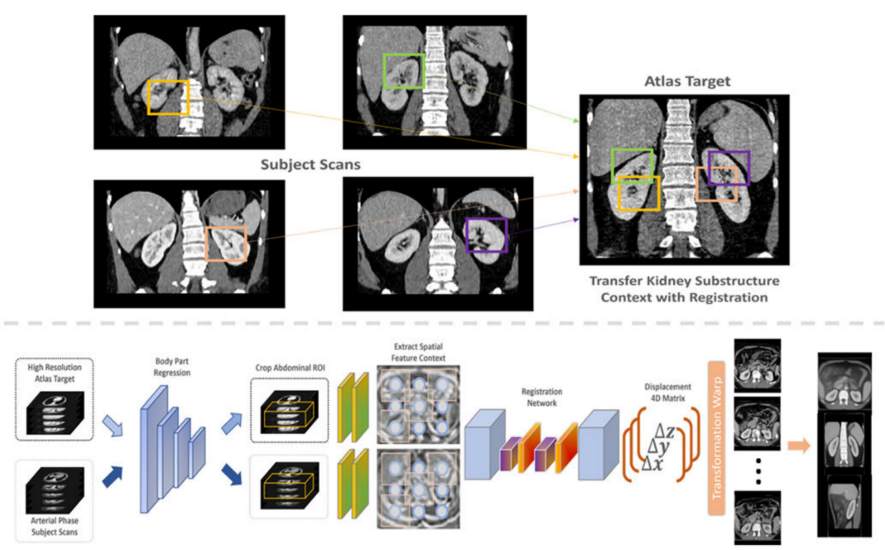


Figure 11.2: We aim to transfer the significant variability of contrast and morphology to a healthy-defined anatomical atlas target with linear and non-linear transformations for generalizing anatomical context across scales (top panel). The complete pipeline (lower panel) can be divided into two steps: 1) body part regression preprocessing and 2) deep supervised registration. We initially crop the abdominal are of interest for both atlas target and subject scans with the guidance of body part regression network. We downsample both volumes and input into a deep registration network to predict the voxel displacement across tri-planar perspective. We finally warp the predicted transformations to each subject scan and compute average map for analysis.

11.3.2 Registration Pipeline

To adapt to the significant variation of the abdominal regions across patients, a deep network registration pipeline is introduced with probability dense displacement networks (PDD-Net), which aims to align a 3D moving image I_m to the fixed image I_f space with an optimal spatial transformation learned on the extracted deep feature context (81; 85). The complete overview of the registration pipeline is illustrated in Figure 11.2. Both preprocessed subject scans and atlas target are input into a small network to learn a meaningful non-linear mapping to align from input intensity to a dense feature context. The Obelisk approach is employed for the small network f to effectively capture the spatial context with significant deformation and a normal $5 \times 5 \times 5$ convolution kernel is added to learn the edge feature for body-aligned registration (85). After the extraction of the spatial feature context, our goal is to predict an optimized displacement field $\delta(k) \rightarrow v$ that apply a vector v to every set of control points $k \in \mathscr{R}^3$ on a grid for non-linear transformation and additionally achieve the best similarity between the organ labels. With the use of conventional discrete registration (83) and the correlation layer in (53), we sample the context from discrete grids and compute a 6D tensor D representing the dissimilarities with the feature dimension z. Here, we use the negated mean square error

across the feature dimension to generate the 6D tensor:

$$D(k,v) = -\frac{1}{|z|}\sum_{z}(f_z(I_f)_k - f_z(I_c)_{k+v})^2 \tag{11.1}$$

As abdominal structures vary significantly across demographics, it causes over-deformation and the registered images to become ill-posed with non-linear registration. PDD-Net provides an opportunity to model the regularization constraints with diffusion regularization and use fast mean-field inference with two iterations only for discrete optimization (111). It consists of two steps: 1) transformation with label compatible that assigns on spatial control points and 2) the additional average pooling layers with stride 1 for filtering message. Previous work demonstrated that the diffusion regularization can be generated using min-convolutions with a lower envelope of parabolas rooted at the offset of 3D displacement with heights equivalent to the sum of dissimilarity terms and the previous iteration of mean-field inference (61). We approximately compute the diffusion regularization for dense displacement with a min-pooling layer to extract the local minima in the cost tensor. Two average pooling layers are additionally used for smoothing the context extracted. The integration of min-pooling and average pooling aims to perform the regularization approach in multiple dimensions: 1) 3 displacement dimensions (min-convolution) and 2) 3 spatial dimensions (mean-field inference) for end-to-end optimization.

11.3.3 Deep Label Supervision for Registration

Apart from adapting the dissimilarities in intensity level, we aim to adapt the anatomical context of a specific region of interest (ROI) with high stability. Here, we use a supervised label loss term to preserve the morphological information of ROI instead of over-deformation. We use a SoftMax activation to compute a probability mapping with the regularized output over the displacement context. One-hot representations from the moving subject segmentations are wrapped in the corresponding spatial location with predicted displacements and compute the dissimilarity between wrapped subject labels and fixed target labels with mean square error (MSE). The predicted displacement is a 4D multi-channel output and the number of channels represented the displacement field on each direction (Δx,Δy,Δz). The final prediction is resampled to the original input resolution with tri-linear interpolation.

11.4 Data and Experiments

11.4.1 Data and Platform

To evaluate the atlas template, abdomen CT volumetric scans from 1971 patients were retrieved in de-identified form from ImageVU with the approval of the Institutional Review Board (IRB) (IRB number:

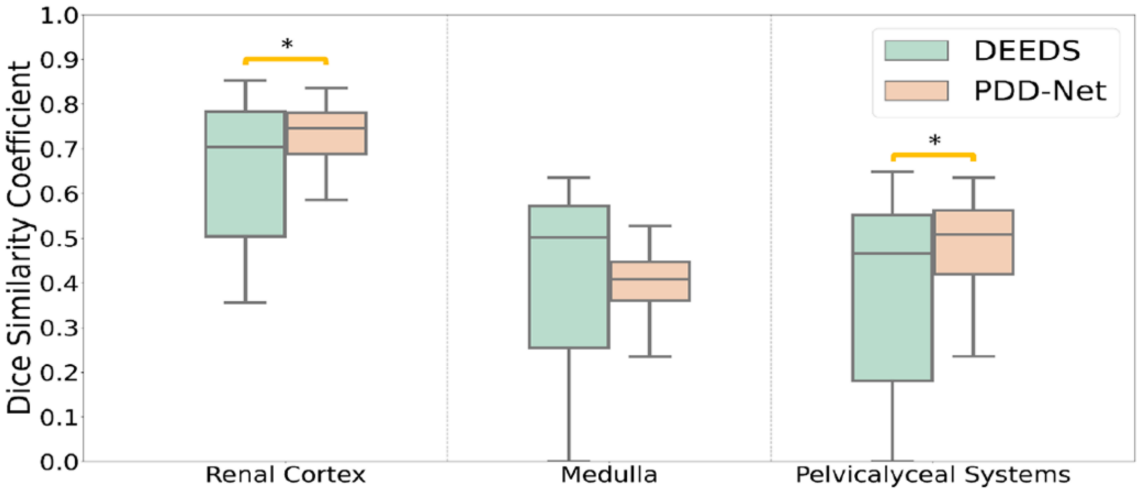


Figure 11.3: This quantitative representation demonstrates the reigstration performance from subject space to the atlas target with subject label transfer. Supervised registration performs with a substantial improvement in the renal cortex and pelvicalyceal system segmentation, while the performance of the medulla is limited with the decrease of variance.

160764). Exclusion criteria are set based on ICD-9 codes to include subjects with healthy kidney organs only. Out of 1971, 500 subjects are retrieved with the assurance of ICD-9 codes and within the age range of 18 to 50 years old. We limited our studies to the subjects with arterial phase CT only, as only the arterial phase CT provides adequate data on the distinctive anatomy of kidney substructures. Therefore, in a total of 500 subjects 3D volumetric CT are used to generate and evaluate the atlas template. For the atlas target image, a single subject volume with high resolution ($0.8 \times 0.8 \times 0.8$) is used with a dimension of $512 \times 512 \times 434$. The criteria to choose data as the atlas template are based on the contrastive and morphological characteristics of the kidneys . With the use of body part regression preprocessing, both subject scans and atlas target are cropped to the abdominal interest only. We downsampled the subject scans and atlas target to an isotropic voxel resolution of $1.5mm \times 1.5mm \times 1.5mm$ and a dimension of $192 \times 160 \times 256$. The downsampled scans are then input into the deep registration pipeline to predict the displacement field for transferring the kidney substructure context to the atlas space.

11.4.2 Experiments

11.4.2.1 Registration Comparison

We performed a conventional registration algorithm DEnsE Displacement Sampling (DEEDS) as our baseline method and previous works have demonstrated that DEEDS outperformed other traditional registration tools (ANTS, NiftyReg) in the inter-patient 3D abdominal CT registration study. DEEDS calculate the image similarity between a number of random sampling voxels of each control point and model the diffusion regularization through a pair-wise term of the displacement field. Hierarchical steps are performed following with 1) DEEDS affine registration and 2) DEEDS deformable registration. A displacement matrix is an output to provide the transformation information corresponding to each control point. We apply the displacement matrix to the subject ground truth labels and compare the transformed label with the atlas target label for evaluating the registration performance. The ground truth label of kidney substructure corresponds to the renal cortex (label: 1), medulla (label: 2), and pelvicalyceal system (label: 3) (201). Dice Coefficient is used to measure the overlapping regions between the prediction label and ground truth labels.

$$Dice(P,G) = \frac{2|P \cap G|}{|P| + |G|} \tag{11.2}$$

11.4.2.2 Atlas Construction

Both subject scans and atlas target are input corresponding to the moving image and fixed image respectively. As the image inputs have to be downsampled to fit into the deep registration pipeline without memory outage, we aim to minimize the loss of the high-resolution context from the atlas target. After we output the predicted displacement field, we further resample the displacement field to the original resolution of the atlas target with tri-linear interpolation and warp the displacement field on upsampled subject scans. All warped upsampled scans are summed and generate the average map to evaluate the transition of the anatomical context across kidney substructures qualitatively.

11.4.2.3 Deep Registration Model

We initially use the pre-trained model from (81), which trained on 10 contrast-enhanced 3D CT scans of the VISERAL3 training dataset. Each scan in VISERAL3 training dataset consists of nine anatomical structures well-annotated including 1) liver, 2) spleen, 3) pancreas, 4) gallbladder, 5) urinary bladder, 6) right kidney, 7) left kidney, 8) right psoas major muscle and 9) left psoas major muscle (105). All images are downsampled to an isotropic voxel of $2.0mm \times 2.0mm \times 2.0mm$ with dimensions of $192 \times 160 \times 256$ for training input and no initial affine alignment is performed between the subject scans and fixed target scan.

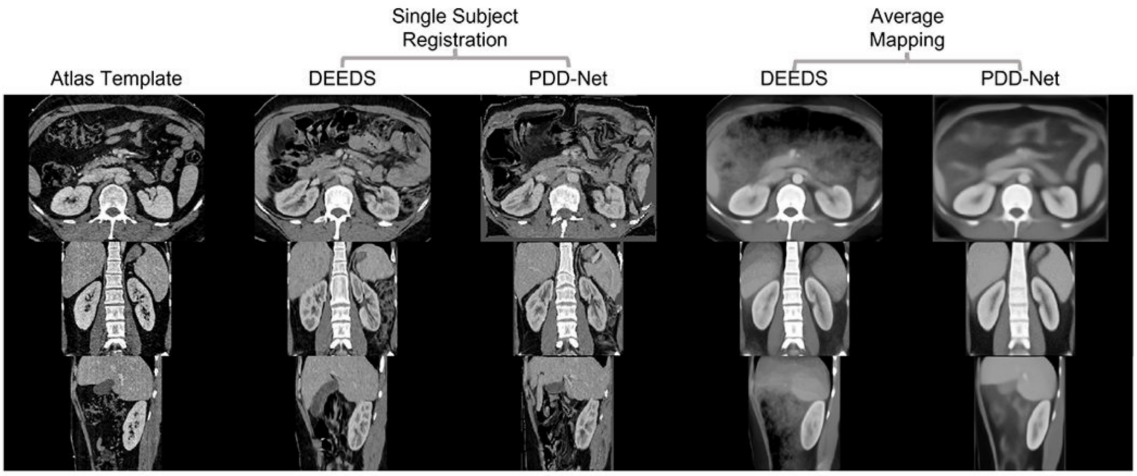


Figure 11.4: This qualitative representation demonstrates the comparison of transferring kidney substructure information with DEEDS and PDD-Net. The single subject is arbitrary picked and shows that the renal structure is comparatively over-deformed with DEEDS. From the average template, the contrastive and morphological context of the substructure is illustrated more appealing both in the boundary and the cortex anatomy with PDD-Net.

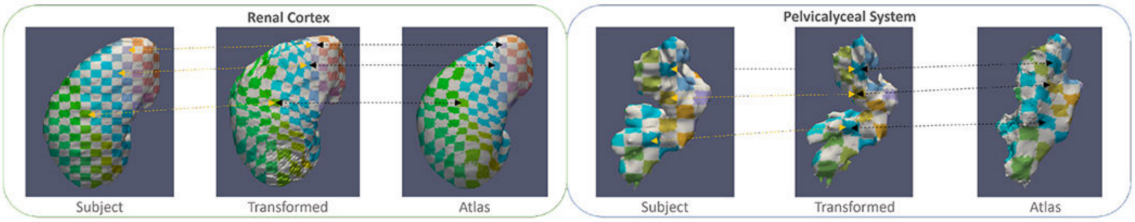


Figure 11.5: We illustrate the correspondence of the anatomical information in between the subjects' space and atlas space after performing deep registration. Yellow arrows show the corresponding information of both organs before and after registration, while the black arrows correspond to the location of the registered context in both transformed images and the atlas target. The atlas template demonstrates stable adaptation of renal cortex context and significant deformation of pelvicalyceal system morphology across subject space.

11.5 Results

To evaluate the generalizability of the atlas template on the localizing contextual information of kidney substructures, we first compare the single subject registration qualitatively, and manually visualize the quality of the registered kidney substructures. As shown in Figure 11.4, the registered output from DEEDS demonstrates a fair registration quality of transferring both left and right kidney information. However, over-deformation is demonstrated with a small extent to the left kidney structure and the boundary information between the right kidney and liver cannot be clearly separated. With the PDD-Net registration, the context of the left and right kidney is comparable to the anatomical information of the kidneys in the atlas target. The substructure context is stably transferred to the atlas-defined space with limited over-deformation. We further evaluate the atlas target across all populations and generate the average mapping to visualize the localizing context of the kidney substructure in the atlas template. The average mapping generates with PDD-Net demonstrates a more distinctive appearance in the kidney substructures, while a certain extent of blurriness in the kidney substructure context is shown with DEEDS registration. The contrastive and morphological characteristics of the kidney substructures are transferred with stability in the atlas target across all populations. In terms of label-wise measure, we apply the predicted displacement field to the moving subject labels and compare the similarity between the warped labels and the atlas labels in Figure 11.3. Significant improvement ($p<0.001$) on the renal cortex and pelvicalyceal system segmentation is demonstrated with PDD-Net, while DEEDS demonstrates a better performance on the medulla segmentation. The medulla segmentation with PDD-Net demonstrates a significant decrease of variance in performance and shows the generalizability across the hard registration case for kidneys.

Apart from looking into the image registration performance, we evaluate the ability to adapt the significant morphology of kidney substructures by generating 2D color correspondence mapping with surface rendering. Here we compute 2D cie-lab checkerboard for each of the kidney substructure labels in the corresponding space and warp the checkerboard to the kidney substructure rendering, as shown in Figure 11.5. Each color represents the corresponding anatomical information located in the subject/atlas space. After warping with the displacement field, the checkerboard pattern follows the guidance of the displacement field and is deformed to demonstrate the adaptation of the subject information to the atlas space. The checkerboard pattern stably deformed from subject space to atlas space and well adapted the structural characteristics of the subject-wise kidney substructure. Grids in the pattern have not been overly deformed in the renal cortex labels, while a significant deformation is demonstrated to transfer the pelvicalyceal system information due to the significant variation of morphology.

11.6 Discussion and Conclusion

With the qualitative and quantitative representation above, the contrastive and morphological context of the kidney substructures is demonstrated with stability using the deep representation pipeline. The label transfer performance in the medullary regions of the kidney does not demonstrate significant improvement, although variance between registered labels is decreased in trend. As the deep registration model is trained with multi-organ labels, the registration aims to optimize the transformation of contextual information in the organ-corresponding regions, which is the complete kidney morphology, instead of the kidney substructures. With the opportunity of using label context in deep registration pipeline, we will adapt the current pre-train model to reduce the adverse effect from the morphological variability of other organs (such as the liver, spleen), and further fine-tune the model with the kidney substructure labels to optimize the kidney substructure registration. In this paper, we constructed a stable standard anatomical reference to localize the context of kidney substructure with deep network registration. The average mapping demonstrated the contrastive characteristics of each substructure across patients and the atlas target stably adapted the substructure information with the illustration of correspondence figure. We aim to create a minimal bias average template for healthy kidney substructures as our future long-term goal and analyze variability across populations.

CHAPTER 12

Multi-Contrast Computed Tomography Atlas of Healthy Pancreas

12.1 Overview

[1] With the substantial diversity in population demographics, such as differences in age and body composition, the volumetric morphology of pancreas varies greatly, resulting in distinctive variations in shape and appearance. Such variations increase the difficulty at generalizing population-wide pancreas features. A volumetric spatial reference is needed to adapt the morphological variability for organ-specific analysis. Here, we proposed a high-resolution computed tomography (CT) atlas framework specifically optimized for the pancreas organ across multi-contrast CT. We introduce a deep learning-based pre-processing technique to extract the abdominal region of interests (ROIs) and leverage a hierarchical registration pipeline to align the pancreas anatomy across populations. Briefly, DEEDs affine and non-rigid registration are performed to transfer patient abdominal volumes to a fixed high-resolution atlas template. To generate and evaluate the pancreas atlas template, multi-contrast modality CT scans of 443 subjects (without reported history of pancreatic disease, age: 15-50 years old) are processed. Comparing with different registration state-of-the-art tools, the combination of DEEDs affine and non-rigid registration achieves the best performance for the pancreas label transfer across all contrast phases (non-contrast: 0.497, arterial: 0.505, portal venous: 0.494, delayed: 0.497). We further perform external evaluation with another research cohort of 100 de-identified portal venous scans with 13 organs labeled, having the best label transfer performance of 0.504 Dice score in unsupervised setting. The qualitative representation (e.g., average mapping) of each phase creates a clear boundary of pancreas and its distinctive contrast appearance. The deformation surface renderings across scales (e.g., small to large volume) further illustrate the generalizability of the proposed atlas template.

12.2 Introduction

With the complicated relationship between physiological and metabolic process in the human body, substantial efforts are underway to map the organization and molecular profiles of cells within specific tissues (183). Adapting multi-scale context from cell to organ level is also essential to provide a better definition to correlate the biomarkers across different imaging domains (119). Computed tomography (CT) is widely used to visualize the patients' anatomy at a system scale (18). While CT provides the anatomical correspondence in a system-scale only, a standardized framework is needed to generalize information across scales and enhances

[1] Published at: Lee, Ho Hin, et al., "Supervised Deep Generation of High-Resolution Arterial Phase Computed Tomography Kidney Substructure Atlas.", Medical Imaging 2022: Image Processing. Vol. 12032. SPIE, 2022. (126)

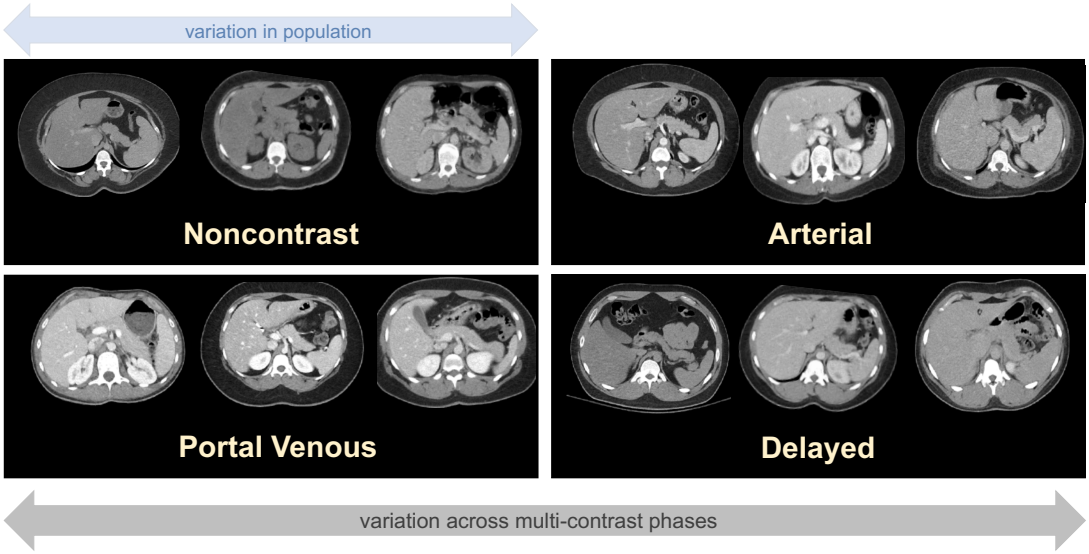


Figure 12.1: The anatomical characteristics of pancreas organs vary widely in population. Visual difference between each contrast phase is also shown a) noncontrast b) arterial c) portal venous d) delayed. A generalizable reference is needed to adapt individual morphological variability within contrast phases.

the ability of visualizing the complex organizations of tissues across populations.

During the imaging procedures, contrast enhancement is performed by injecting a contrast agent before imaging. Multi-phase contrast enhanced CT are generated with the sharpened anatomical and structural details between neighboring organs for diagnosis (185; 163; 199). Four different phases are typically generated according the retaining time of contrast agent in the imaging cycle: 1) non-contrast, 2) arterial, 3) portal venous, and 4) delayed. With the variation in contrast level between organs, the large range of intensity levels are beneficial to capture the fine-grained contextual features of each specific organ, especially for the pancreas organ. By visualizing the anatomical context of pancreas across large population cohorts, we observe that the volumetric morphology of healthy pancreas organ varies with respect to demographics (e.g., sex, body size), as shown in Fig. 1. To investigate the population-wise healthy biomarkers of pancreas organ in the systemic level, a standardized imaging atlas framework is needed to adapt the population-wise anatomical characteristic onto one single template with image registration technique. However, with the large variation in organ anatomies and body sizes across population, generating such standard reference template for pancreas organ is still challenging and no atlas framework for pancreas organ is currently available in public.

Creating an organ/tissue-specific atlas framework has been widely leveraged with magnetic resonance imaging (MRI) and extensive efforts have been applied to leverage brain MRI to investigate biomarkers for multiple perspectives (217; 148). Due to the similarity between human brain and mouse brain, Kovačević et. al proposed a 3D variational atlas with mouse brain to represent the average anatomy and the variation

among the population (110). *Wang* et.al created a population average reference framework leveraging 1675 specimens of mouse brain MRI (217). On the other hand, *Shi* et al. created an unbiased infant brain atlas with group-wise registration from three different scanning time points using MRI from 56 males and 39 females (190). *Kuklisova-Murgasova* et al. proposed multiple atlas to generalize the aging characteristics from 29 to 44 weeks infants (113). *Ali* et al. generates an unbiased spatial-temporal 4-D atlas and time-variable longitudinal atlas for infant brain (64). To further investigate into the aging characteristics in brain tissue, *Yuyao* et al. leveraged patch-based registration in spatial-temporal wavelet domain to generate longitudinal atlas (244). While previous efforts was concentrated on generating healthy brain atlas template, *Rajashekar* et al. proposed high-resolution normative atlases to visualize the population-wise representation of brain disease (e.g., brain lesion, stroke) in both FLAIR MRI and non-contrast CT modalities. For abdominal regions, pioneer studies have been demonstrated to develop a multi-contrast kidney atlas that generalize both contrast and morphological characteristic within kidney organs (131; 132). Furthermore, such kidney atlas template is further extended to generalize the substructure organ (e.g., medulla, renal context, pelvicalyceal systems) in kidney regions with arterial phase CT (126). However, limited studies have proposed to create a standard reference atlas for pancreas organs with its challenging morphology associated with patients' demographics.

In addition to putting efforts into generating tissue/organ atlases, robust image registration algorithms to transfer the anatomical context onto one single template are also vitally important. Previous works have sought to enhance the registration performance by innovating conventional frameworks with affine and deformable transformation (3; 5; 10). The spatial transformation is optimized by regularizing the deformation field to align the anatomical context from moving image to a single fixed template with traditional approach such as discrete optimization (43), b-spline deformation (184), Demons (212), and symmetric normalization (5). To further enhance the efficiency and robustness of registration algorithm, deep learning was introduced to generate large deformation field by extracting meaningful representations for deformation field predictions. VoxelMorph is a pioneering network that optimizes with a generalized function to compute deformation field in unsupervised setting (11; 42). While VoxelMorph was initially optimized with brain imaging, a large deformation field is needed for abdominal imaging due to the significant variation in body size and organ morphology across demographics. *Zhao* et al. adopt Voxelmorph framework and extended as a recursive cascaded network that leverage organ labels to crop the organ-specific region of interests (ROIs) and progressively registered the anatomical context to the fixed template (246). *Zhao* et al. introduced a deep learning framework that generates bounding boxes to initially localize multiple organ ROIs and leverage the organ-specific patches for registration (249). However, previous approaches only demonstrate the feasibility of registering organ-specific ROIs. *Heinrich* et al. adopt substantial deformation in compute volumetric scans

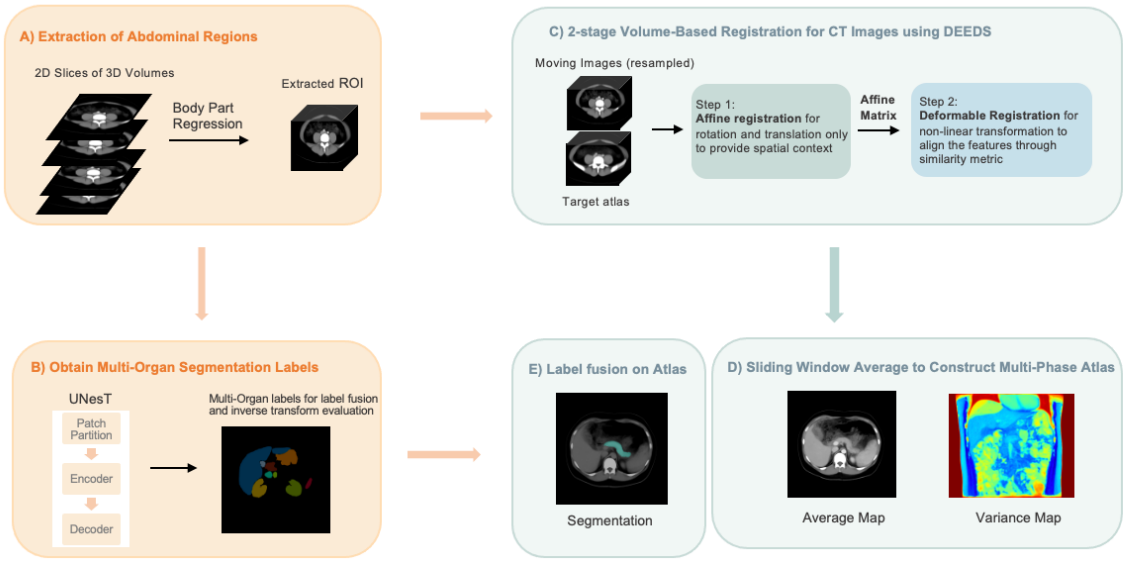


Figure 12.2: Complete overview of our propose atlas generation framework. We extract the abdominal regions from CT scans using the body part regression network. We register the cropped ROI to the reference image with a hierarchical two-stage registration and compute both average and variance mappings to evaluate the effectiveness of the atlas framework. Furthermore, we statistically fuse the pseudo predictions from the transformer-based segmentation network UNesT and perform inverse transformation back to the subject space for evaluation.

by innovating a probabilistic dense displacement network with organ label supervision (81). Yet, voxel-wise labels are needed to supervise the training process and instability in performance may exist due to the domain shift with unseen data (132).

In this work, we propose a high-resolution CT pancreas atlas framework that optimized for the healthy pancreas. Briefly, we initially crop the abdominal ROI from each each subject scan with a deep neural network called body part regression (BPR), which aims to minimize the field of view (FOV) differences between scans and reduces the failure rate of registration. Specifically, we slice the volumetric scans and input each slice into BPR network to compute a regressed score ranging from -12 to +12, referring as the upper lung region to the pelvis region in the body. We then limit the predicted score range and extract the ROI within abdominal region only. After data preprocessing, a two-stage hierarchical registration pipeline is performed to register each subject scan to the high-resolution atlas target with metric-based registration across all contrast phases (83; 84). To evaluate the quality of anatomical transfer across scans, we compute average and variance mappings to demonstrate the organ appearance across all registered outputs in each phase and further quantify the registration performance with inverse label transfer from atlas framework. Overall, our main contributions are summarized as four folds:

- We established a standardized framework to obtain a population-based pancreas atlas.
- We propose a hierarchical metric-based framework optimizing for healthy pancreas to generalize the anatomical and contrast characteristics of pancreas organ across demographics and domain shift of contrast phases.
- We evaluate the effectiveness of our proposed atlas template by inversely transforming the atlas target labels to a de-identified research cohort of 100 abdominal scans with 13 organs ground-truth labeled. A large population of unlabeled multi-contrast phase CT cohort is leveraged to compute average and variance mappings to demonstrate the generalizability of the atlas framework. Our proposed atlas framework achieves a stable transfer ability in pancreas organ with an average Dice of 0.504 in unsupervised setting.
- The average template generated, and the associated pancreas organ labels is available public for usage through HuBMAP.

12.3 Methods

12.3.1 Self-Supervised Body Part Regression Network for Preprocessing

With the substantial variation in imaging protocols, the imaging samples from a large cohort usually present with different range of field of views (FOVs). Such variability of FOV may increase the failure rate of registration when the FOV difference between the subject scan and the atlas template is large. Here, we adapt body part regression (BPR) network to extract similar FOV within the abdominal regions only, thus to enhance the registration performance. Specifically, given an unlabeled dataset $\{x_{ui}\}_{i=1}^{N}$ as the moving image domain, and the atlas image $\{x_a\}$, our goal is to crop the volumetric x_u to have an approxiamte FOV of abdominal interest only with the atlas template x_a. *Tang* et al. proposed a self-supervised BPR network to generate a continuous score for each axial slice in the volumetric scans as the normalized body localization values (234; 198). Each score is within the range of -12 to + 12, which refers to different anatomical location (e.g., -12: upper chest, -5: diaphragm/upper liver, 4: lower retroperitoneum, 6: pelvis). For abdominal region, we limit the score ranges across all axial slices within -5 to 5 and crop the slices.

12.3.2 Transformer-Based Segmentation Network

Apart from data preprocessing, pancreas organ labels are needed to further quantify the healthy biomarkers across populations and perform statistically label fusion to generate the atlas label. With the recent advance of deep neural networks in medical image segmentation (200; 125; 121), transformer-based networks have been further proposed to leverage with respect to their proficiency in capturing long-range dependency and

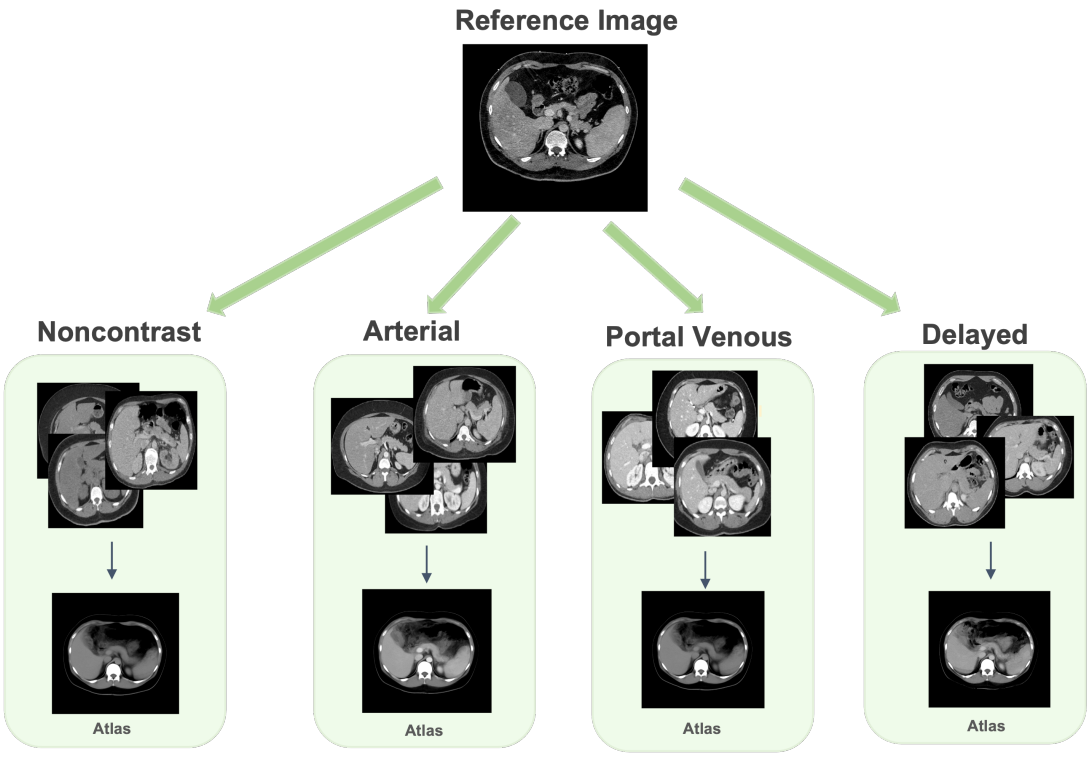


Figure 12.3: To evaluate the effectiveness of our atlas framework, inverse transformation is performed to backpropagate the anatomical label from the atlas template to the moving subject domain and quantify the similarity with the corresponding ground truth label.

the large receptive filed behavior (74; 73). Here, we adapt a transformer-based model that incorporates a hierarchical design alongside a block aggregation method, attaining state-of-the-art performance across multiple modalities and public datasets (). This model initially projects 3D volumes into a sequence of patches, and then employs a 3D block aggregation algorithm to augment communication between these patches. Within each hierarchical level, patches are transformed into blocks and introduced to the transformer layer. The output then undergoes the block aggregation algorithm and is subsequently downsampled for the next hierarchy. The model consists of three hierarchies with a configuration of 64, 32, and 1 block(s) respectively. The block aggregation algorithm leverages local attention, therefore enhancing data efficiency. In our study, the needs for segmentation labels on abdominal scans is twofold: for inverse transform evaluation and joint label fusion to visualize the pancreas region on atlases.

12.3.3 2-Stage Hierarchical Registration

Our registration framework consists of two main steps: 1) affine registration and 2) deformable registration. Inspired by *Lee* et al. (132; 126), we adapt dense displacement sampling (DEEDS) as our registration backbone for abdominal imaging. The concepts of DEEDS is to compute a large deformation filed with a

discretized sampling space to align the anatomical context across andominal organs with variable morphology (83; 84; 82). We first perform DEEDS affine registration to align the abdominal organs with 12$\circ$ degrees of freedom from moving images to the atlas template, coarsely providing prior spatial information and each affine component with the generated affine transformation matrix as our first stage output. The affine-aligned intermediates are leveraged for the DEEDS deformable registration as our second stage registration process. The DEEDS deformable registration refines the spatial relationship between randomly selected patches and compute the local voxel-wise correspondence with its specific similarity metric as follows:

$$D(x_m, p_x, p_y) = exp(\frac{S(p_x, p_y)}{q^2}, \;\; p_x, p_y \in N \tag{12.1}$$

where p_x and p_y are denoted as the center coordinate of another patch from one of the neighbourhood N, S is the self-similarity metric, which is optimized by a distance function D between the image patches from the moving samples. q^2 denotes as a function to evaluate the noise in the local and global perspectives. The similarity metric enhance the avoidance of the adverse effect from image artifacts or random noise from the central extracted patch. During the deformable registration, five different levels are leveraged in grid spaceing ranging from eight to four voxels to randomly select patches. The displacement search radii are defined from six to two steps between five and one voxels. Six neighborhoods are chosen to compute 12 distances between pairwise patches for optimization (83; 84; 82). Both deformed scans with the corresponding displacement matrix are generated as the final output and align the fine-grain anatomical details of each organ.

12.4 Experimental Setup

12.4.1 Datasets

Clinical Research Multi-Phase CT Cohort: A total of 443 multi-contrast phase CT volumes are selected for the formation of multi-phase atlases from a large cohort of abdominal CT scans of 2000 patients under the approval of Institutional Review Board (IRB #131461). The scans include retro-peritoneal and abdominal organs. Since our goal is to construct a healthy CT atlas optimized for pancreas region, we exclude patients who exhibit pancreatic lesions. 898 subjects whose age ranging from 18 to 50 years old. After quality assessment of registered subjects, 443 unlabeled CT volumes were used to create the atlas for four phases: 59 volumes for non-contrast phase, 40 volumes for arterial phase, 330 for portal venous phase, 14 for delayed phase. All volumetric scans are initially reoriented to a standard orientation (RAS) before data preprocessing (101). BPR is then performed to each scan and obtain the FOV in abdominal region only. Each scan is then resampled to the same resolution and dimension of the atlas template for registration.

Multi-Organ Labeled Portal Venous Abdominal CT Cohort: To evaluate the generalizability of our pro-

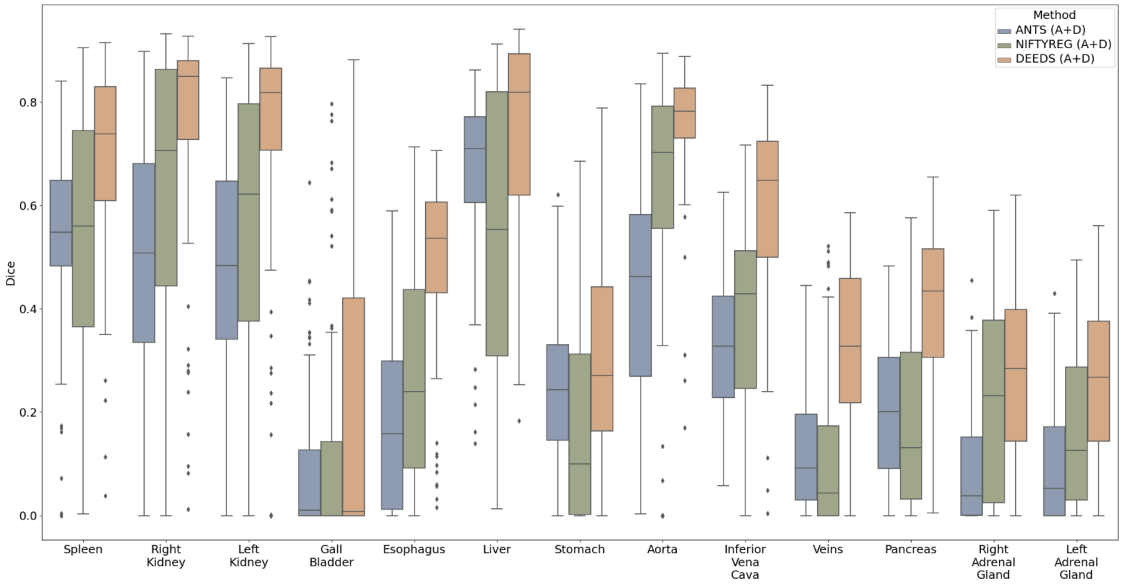


Figure 12.4: The quantitative representation of inverse label transfer with multi-organ portal venous CT dataset are demonstrated that DEEDS with affine registration outperforms the other two traditional methods.

posed atlas framework, we further leverage a separate healthy clinical cohort with 100 de-identified portal venous phase abdominal CT scans and 20 of the volumetric scans are the testing scans in the MICCAI 2015 Multi-Atlas Abdomen Labeling (BTCV) challenge. All volumetric scans are labeled with 13 multiple organs including: 1) spleen, 2) right kidney, 3) left kidney, 4) gall bladder, 5) esophagus, 6) liver, 7) stomach, 8) aorta, 9) inferior vena cava (IVC), 10) portal splenic vein (PSV), 11) pancreas, 12) right adrenal gland (RAD), 13) left adrenal gland (LAD).

High-Resolution Single Subject Atlas Template: Inspired by (132), we choose the atlas template under several conditions: 1) high-resolution in both in-plane and through-plane, 2) distinctive contrast appearance in pancreas organ morphology with clear boundary, and 3) in healthy condition. The atlas template is provided by Human Biomolecular Atlas Program (HuBMAP) with a high resolution of $0.8 \times 0.8 \times 0.8$ and the corresponding dimension of $512 \times 512 \times 434$. The atlas template is annotated with 13 organs by performing statistical label fusion with the pseudo segmentation from all registered subjects.

12.4.2 Implementation Details

For the BPR model, we pretrain a U-Net like architecture network with a total of 230,625 2D slices from a large population samples of 1030 whole body CT scans (collected from public domain) (198), while the multi-organ labeled portal venous phase CT are leveraged as external validation only. The pretrained U-Net

model is end-to-end optimized with Adam optimizer with a learning rate of 0.0001 and batch size of 4.

After we preprocess all imaging samples with BPR model, we further evaluate our proposed atlas template in multiple perspectives. We first investigate the effectiveness of current state-of-the-art registration tools for abdominal imaging across all contrast phases in both qualitative and quantitative perspective. We performed extensive analysis with tools such as ANTS (232; 5), NIFTYREG (232; 155), and DEEDS (82; 83; 84) as metric-based registration with multi-organ labeled portal venous CT cohort. Before the registration, all preprocessed scans are upsampled to the same resolution with the atlas target and perform the 2-stage hierarchical registration in high-resolution setting. We further perform multiple ablation studies to investigate the effectiveness of the BPR preprocessing to enhnace the successful rate of abdominal organ registration. Moreover, we perform ablation study to search the optimal range of BPR score to define the best FOV correlation between all moving images and the atlas target.

12.4.3 Evaluation Metrics

We adapt two commonly used metrics to evaluate the similarity between the inverse transferred labels from atlas space to moving subject space and the corresponding moving ground truth label: 1) Dice similarity score, and 2) Hausdorff distance (HD). The definition of Dice score is to compute the overlapping ratio between the predicted pixel/voxel-wise label and the ground truth label. The Dice score is defined as follows:

$$Dice(P,G) = \frac{2|P \cap G}{|P| + |G|} \tag{12.2}$$

where P denotes as the label prediction and G is the corresponding ground truth label, while $\|$ denotes as L1 normalization. For computing the Hausdorff distance, we extract the 3-dimensional coordinates of each vertice from the surface rendering of both the predicted label and the ground truth. We compute the Hausdorff distance with the vertices as follows:

$$HD(v_p, v_g) = sup\, inf\, Dist(v_p, v_g) \tag{12.3}$$

where v_p and v_g denotes as the vertice coordinates of the label prediction and ground-truth label respectively. Sup and inf refer to the upper and lower bound of the distance function $Dist$ value.

12.5 Results

12.5.1 Evaluation with Clinical Research Cohort and Multi-Organ Labeled Cohort

After cropping the abdominal regions from the raw data using body part regression, we resampled the cropped volumes and registered them to the reference volume using DEEDs. We then qualitatively assessed the

Table 12.1: Inverse label transfer performance of different registration methods on 13 organs average for Clinical Research Multi-phase CT Cohort (*: p<0.0001, Wilcoxon signed-rank test, A: affine registration, D: deformable registration)

	Non-Contrast		Arterial		Portal Venous		Delayed	
Methods	Dice ↑	HD (mm) ↓	Dice ↑	HD (mm) ↓	Dice ↑	HD (mm) ↓	Dice ↑	HD (mm) ↓
ANTS (A)	0.256±0.115	47.9±22.1	0.242±0.103	39.9±16.8	0.242±0.082	42.2±18.0	0.249±0.078	42.5±20.1
NIFTYREG (A)	0.232±0.141	52.8±22.8	0.200±0.131	47.8±17.3	0.201±0.117	46.4±19.0	0.211±0.103	47.5±15.9
DEEDS (A)	0.192±0.103	48.7±17.8	0.144±0.112	46.4±12.6	0.148±0.111	45.3±14.5	0.150±0.094	49.0±16.5
ANTS (A+D)	0.288±0.121	47.2±22.4	0.302±0.114	38.8±17.1	0.308±0.097	40.1±18.4	0.318±0.065	41.3±20.1
NIFTYREG (A+D)	0.314±0.154	51.7±22.2	0.334±0.180	45.7±16.3	0.350±0.176	45.6±19.1	0.378±0.150	46.8±17.8
DEEDS (A+D)	**0.497±0.076***	**39.2±21.9***	**0.505±0.075***	**32.5±14.7***	**0.494±0.077***	**33.3±17.5***	**0.497±0.086***	**37.1±20.7***

Table 12.2: Inverse label transfer performance of different registration methods on 13 organs for multi-organ labeled portal venous phase CT Cohort (*: p<0.0001, Wilcoxon signed-rank test, A: affine registration, D: deformable registration)

Methods	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	PSV	Panc.	RAG	LAG	Avg
ANTS (A+D)	0.536±0.175	0.498±0.226	0.484±0.206	0.091±0.139	0.183±0.167	0.662±0.158	0.256±0.138	0.431±0.201	0.323±0.143	0.121±0.110	0.204±0.132	0.092±0.110	0.095±0.111	0.306±0.246
NIFTYREG (A+D)	0.544±0.241	0.611±0.290	0.556±0.287	0.113±0.206	0.270±0.201	0.538±0.287	0.175±0.198	0.615±0.264	0.379±0.196	0.111±0.143	0.184±0.170	0.222±0.182	0.162±0.151	0.344±0.293
DEEDS (A+D)	**0.697±0.178***	**0.756±0.207***	**0.729±0.224***	**0.211±0.283***	**0.491±0.166***	**0.746±0.180***	**0.315±0.208***	**0.756±0.200.116***	**0.596±0.173***	**0.329±00.165***	**0.398±0.168***	**0.274±0.169***	**0.253±0.155***	**0.504±0.280***

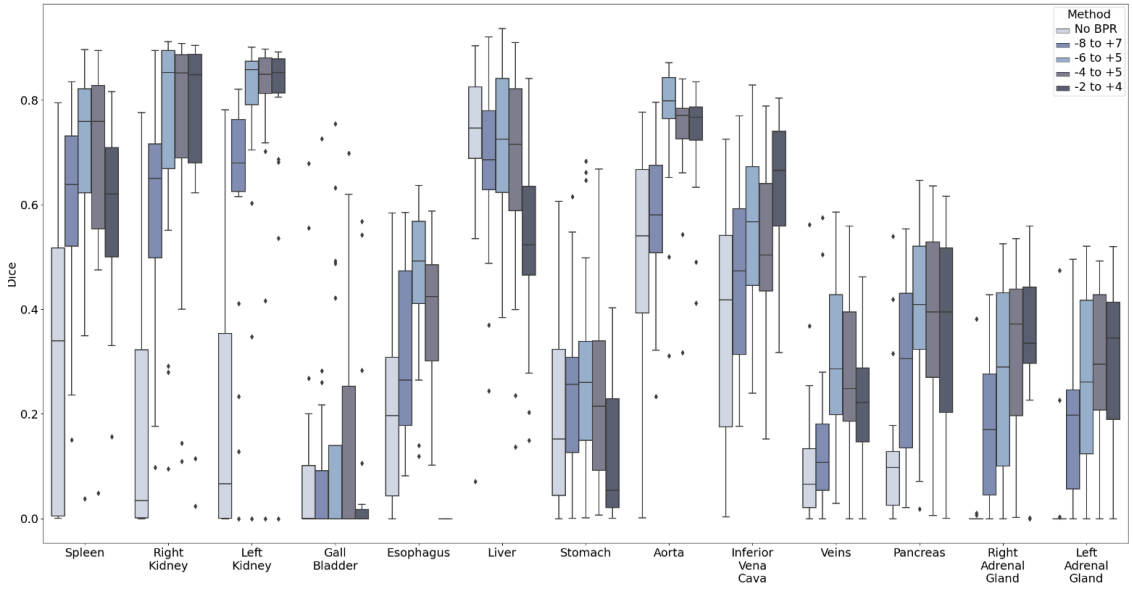


Figure 12.5: We perform ablation study of evaluating inverse label transfer performance with different cropping value range for body part regression. We found that the optimal range for body part regression is between -6 and +5.

registered subjects and removed those with registration failure. The average atlas for the four phases was subsequently obtained from the registered volumes. Fig. 5 presents the tri-planar view of the average atlas, in which the anatomical features of the pancreas can be clearly seen in all directions. Furthermore, the shape of surrounding organs, such as the spleen, is well-preserved, providing a better anatomical context for studying the pancreas. The variance of average templates is shown in Fig. 6. We calculate the log-scale variance between the average template and the registered subjects used to create the atlas, normalizing them to a range

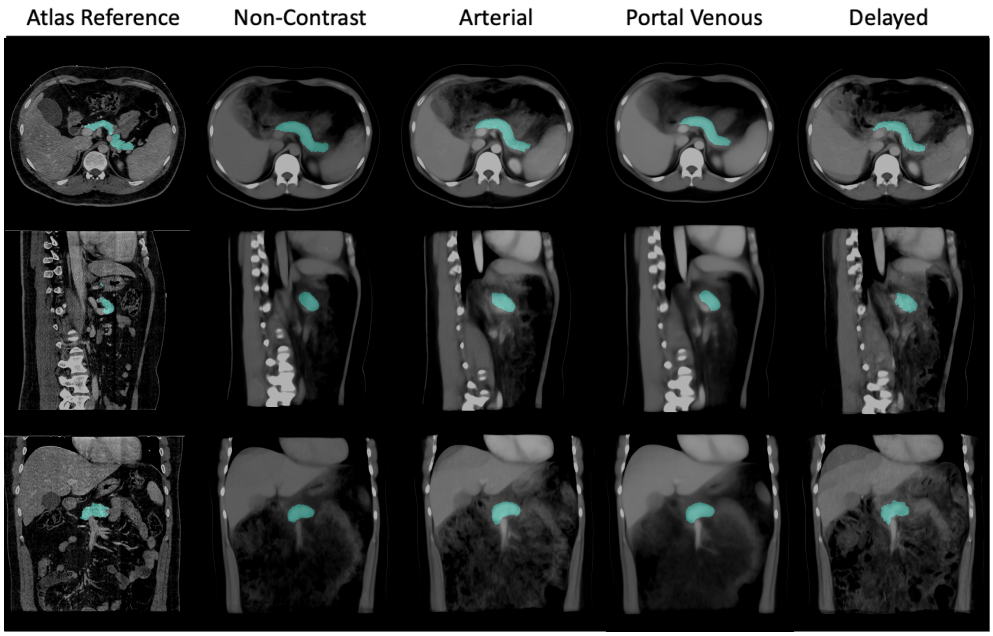


Figure 12.6: We investigate the registration stability across all contrast phase with average mapping. We observe that the morphology of pancreas across all phases are well preserved with clear boundaries.

of 0 to 1. Among all registered regions, the pancreas exhibits relatively low variance for all phases compared to the spleen and spine. The non-contrast and portal venous phases have lower variance, approaching 0.

To enhance the identification and visualization of the pancreas region in the atlases, we also acquired segmentation labels by employing a label integration technique that relies on majority voting. As depicted in Figure 5, the pancreas' position and shape in the multi-phase atlases closely resemble the target reference, indicating that the anatomical and contextual information of the pancreas has been effectively transferred into the target space. Even the areas of greatest variation, such as the head and tail of the pancreas, are distinctly visible in the fused labels. Minor differences can be observed between the segmentation labels on atlases across the four phases, suggesting that the anatomical features and contrast properties have been maintained in the multi-phase atlases.

The fairness of the obtained atlas using the DEEDS registration method for different phases is also evaluated with inverse transform. Table 1 displays the performance of various registration tools, including ANTS, NIFTYREG, and DEEDS, to create the average atlas for the pancreas and other abdominal organs in terms of Dice score and symmetric Hausdorff distance. The performance of affine registration alone is significantly lower than that of the two-stage registration. With the two-stage registration, DEEDS achieves the highest Dice score and the lowest Hausdorff distance for the pancreas region. Another 100 subjects in the portal venous phase with 13 multi-organ ground truth segmentation labels are used to evaluate the performance of

our framework, as shown in Table 2. To ensure the alignment of all organs in the atlas, we calculated the Dice score and Hausdorff distance on 100 subjects for the pancreas and the other 12 organs without qualitative assessment. DEEDS achieves the highest performance with a mean Dice score of 0.504 for all 13 organs.

12.5.2 Ablation Study

In our ablation study, we further investigate the influence of the BPR on the performance of DEEDs registration with different range of cropping values. We evaluate five distinct cropping range to establish a similar field of interest between the moving subjects and the reference subject. The raw data without BPR offers a comprehensive view of the pelvis, abdomen, and chest regions. The cropping value of -8 to 7 encompasses parts of the pelvic area and the majority of the chest area. The cropping value of -6 to 5 excludes the pelvic region and some of the chest region. The cropping value of -4 to 5 retains solely the abdominal region, while the cropping value of -2 to 4 effectively preserves the pancreas and kidney regions but does not entirely maintain the spleen region. For the experiments without BPR, -8 to +7, and -6 to +5, we cropped only the moving subjects. For the experiments of -4 to +5 and -2 to +4, we cropped both the reference volume and the moving volume.

Figure 4 displays the Dice scores for 13 abdominal organs on 20 registered subjects in the portal venous phase from the BTCV public dataset. The use of body part regression significantly improves the registration performance for all cropping values. The cropping value of -6 to +5 results in the highest average Dice score on the 13 organs and the highest value on the pancreas. Consequently, we employ the -6 to +5 value to crop multi-phase subjects and compute the average atlas from them. Additionally, the value of -6 to +5 maintains a complete view of all abdominal regions, facilitating a more comprehensive analysis of abdominal organs within an anatomical context.

12.6 Discussion

Population-based tissue maps play a crucial role in studying human organs by generalizing the variability between individuals. In abdominal scans, the heterogeneity of abdominal and retroperitoneal scans presents challenges in constructing an atlas capable of visualizing anatomical features and spatial relationships between organs. This study aims to propose a framework for constructing an abdominal atlas optimized for the pancreas region in multi-contrast CT.

Given the varying anatomical features across different phases of multi-contrast CT, it is essential to generalize the characteristics for each phase rather than using a single atlas template. The non-contrast phase involves acquiring images before injecting the contrast agent and serves as the baseline reference. For the pancreas, the arterial and portal venous phases are frequently used for detecting and characterizing lesions

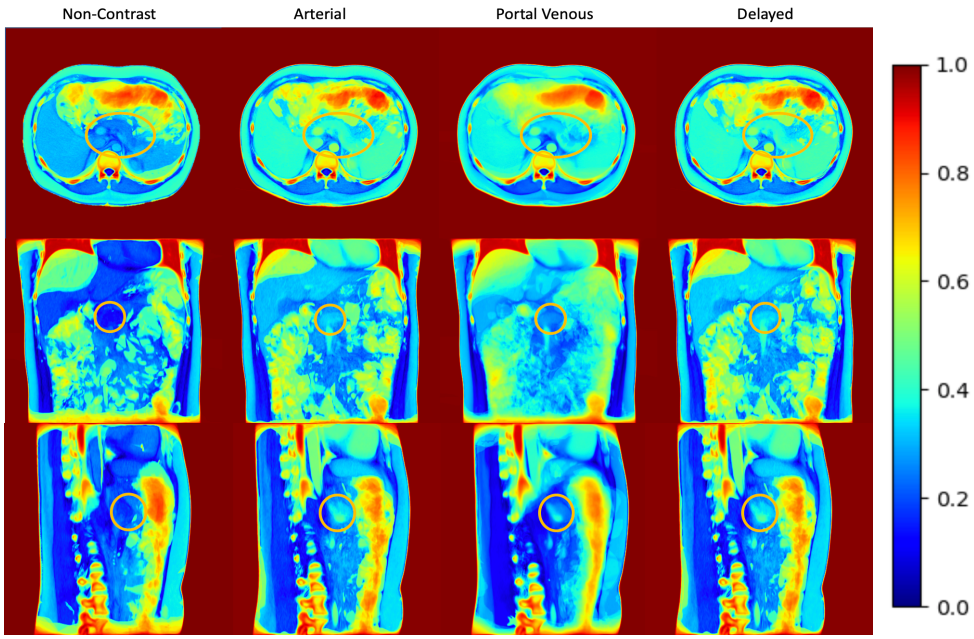


Figure 12.7: We further evaluate the intensity variance across the registration outputs with the average template in each contrast phase. The variance mapping in both non-contrast phase and portal venous phase demonstrates the pancreas context transferal with stability and the variance value near the pancreas region is 0, while the range of variance value is higher in both arterial and delayed phases.

when blood vessels and abdominal organs have been further enhanced by the contrast agent.

To transfer the anatomical characteristics of each phase, we register individual subjects in different phases to the high-resolution atlas template. The robustness of registration tools is vital for the quality of the average atlas. A two-stage registration is used, where the affine registration provides prior information for the deformable registration in the second stage. As shown in Table 2, DEEDs registration achieves an average Dice score of 0.504 on 13 organs, demonstrating accurate transfer of anatomical information for all abdominal organs. ANTs is the least robust, potentially due to its surface-based registration nature. When using ANTs for label registrations, its performance improves, but 3D intensity images are still problematic due to partial matching, especially in aligning the boundaries of abdominal organs. NIFTYREG slightly outperforms ANTs, employing a block-matching approach for non-linear registration, which provides more accurate results in the presence of large deformations but requires longer computation time. DEEDs, a voxel-based method, surpasses the other two methods in performance, possibly because it relies on discrete optimization, allowing greater control over the displacement space (82). One advantage of DEEDs is its use of dense stochastic sampling approaches, sampling random voxels from non-overlapping cubes, and then calculating the displacement on cubes. It ensures accurate registration of small anatomical features undergoing large motion. Additionally, discrete optimization reduces computational complexity, making it more efficient than

continuous optimization.

Nonetheless, registration accuracy is still not ideal and could influence the quality of the average atlas. Challenges for these registration tools include 1) the dissimilar field of view between moving subjects and the target template, and 2) the variation of secondary structures (e.g., muscles, bones) across the population. Many registration failures occur due to mismatched fields of view between subjects, necessitating better preprocessing steps. In the ablation study, we demonstrated the effectiveness of body part regression in matching the field of view. Secondary structures could also distract from the registration of target organs. The variation of these parts could cause undesirable deformation, especially when they occupy a large space in abdominal scans(232). Similarly, small or medium-sized organs could be affected by surrounding organs. For instance, pancreas deformation might be negatively affected by other surrounding target organs with larger sizes, such as the spleen.

In addition to traditional optimization-based registration, learning-based registration methods have gained popularity due to their fast speed. However, concerns remain regarding the use of learning-based methods for constructing average maps in our study. Optimization-based models generally exhibit better expressive power because model parameters are optimized for a specific pair of images, enabling sharper deformation to preserve the details of anatomical features of abdominal organs. On the other hand, learning-based methods optimize the model for the entire dataset, which tends to produce over-smoothed transformations, potentially losing individual characteristics during registration and making them less suitable for high-resolution images (164). Furthermore, learning-based methods typically exhibit less generalizability and are more specific to the dataset.

In addition to constructing average atlases, fused labels on the atlas provide clearer anatomical characteristics and contextualization of the pancreas. As demonstrated in Figure 6, the shape and positions of average segmentation labels closely resemble those in the atlas template. However, limitations still exist. Since the segmentation represents the average of subjects, it is not specific to individual subjects; thus, detailed features and irregular boundary may not be accurately represented on the fused labels, providing only an estimated shape of the pancreas. Another concern regarding segmentation labels on atlases is the accuracy of the segmentation performance of the model. We employed UNesT, which was trained on the BTCV dataset in the portal venous phase. However, due to minor differences in anatomical features between subjects in each phase, segmentation in other phases may not be as accurate as in the portal venous phase.

12.7 Conclusion

This study introduces a high-resolution pancreas atlas framework to generalize the healthy biomarker across population with multi-contrast abdominal CT. By utilizing body part regression to match the field of view

between the target template and moving subjects, and employing the DEEDs registration method to transfer subjects into the target space, our atlas effectively captures population-wide features of the pancreas organ and contextualizes the anatomical characteristics of the pancreas within the entire abdominal scans. Future work involving the use of pancreas atlases could further explore areas such as enhancing the segmentation accuracy of the pancreas and improving the localization of the pancreas in the context of pathological changes.

CHAPTER 13

Unsupervised Registration Refinement for Generating Unbiased Eye Atlas

13.1 Overview

[1]With the confounding effects of demographics across large-scale imaging surveys, substantial variation is demonstrated with the volumetric structure of orbit and eye anthropometry. Such variability increases the level of difficulty to localize the anatomical features of the eye organs for populational analysis. In order to adapt the variability of eye organs with stable registration transfer, we propose an unbiased eye atlas template following with a hierarchical coarse-to-fine approach to provide generalized eye organ context across populations. Furthermore, we retrieved volumetric scans from 1842 healthy patients for generating an eye atlas template with minimal biases. Briefly, we select 20 subject scans and use an iterative approach to generate an initial unbiased template. We then perform metric-based registration to the remaining samples with the unbiased template and generate coarse registered outputs. The coarse registered outputs are further leveraged to train a deep probabilistic network, which aims to refine the organ deformation in unsupervised setting. Computed tomography (CT) scans of 100 de-identified subjects are used to generate and evaluate the unbiased atlas template with the hierarchical pipeline. The refined registration shows the stable transfer of the eye organs, which well- localized in the high-resolution ($0.5mm^3$) atlas space and demonstrated a significant improvement of 2.37% Dice for inverse label transfer performance. The subject-wise qualitative representations with surface rendering successfully demonstrates the transfer details of the organ context and showed the applicability of generalizing the morphological variation across patients.

13.2 Introduction

Significant effort has been invested by the Human BioMolecular Atlas Program (HuBMAP) to relate the molecular findings in organ anatomy across scales (from cellular to organ system level). Previous efforts have been focused mapping the organization and molecular profile in cellular level (183), while several studies have been focused on generating an initial imaging template to adapt contextual information in organ scale. Medical imaging such as computed tomography (CT) and magnetic resonance imaging (MRI) provided the imaging platform to visualize organ anatomy in systematic level. Contrast enhancement is leveraged to emphasize the structural and anatomical context between neighboring organs with the injection of contrast agent and guide to extract information from regions of interest (ROIs). However, the morphology of organs

[1]Published at: Lee, Ho Hin, et al., "Unsupervised Registration Refinement for Generating Unbiased Eye Atlas.", Medical Imaging 2023: Image Processing. Vol. 12464. SPIE, 2023. (127)

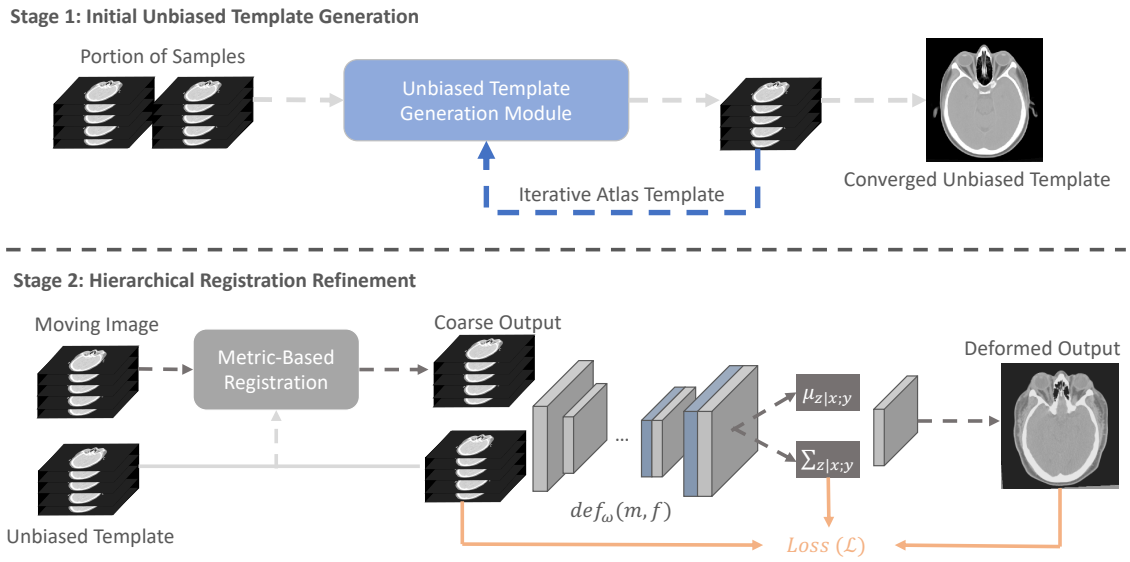


Figure 13.1: The complete pipeline for generating unbiased eye atlas template can be divided into two stages: 1) initial unbiased template generation and 2) hierarchical registration refinement. We leverage the small portion of samples to generate an unbiased template with iterative registration. The average template generated in each iteration is used to be the fixed template for next registration iteration. After the registration is converged , we further use the generated template to perform metric-based registration for the remaining samples. As the coarse output is limited adapt the significant variability of organ structure, a probabilistic refinement network is used to generate large deformation and further aligned the anatomical details across eye organs and head skull.

varies significantly across different demographics, especially in eye organs. The orbital shape and the length of optic nerve varies with age and sex (224). In order to adapt and generalize the population profiles of eye organs, it is important to contextualize the variable anatomy of organs in well-defined reference templates (atlas) to act as a common framework for mapping correspondence across patients (132; 126). From technical perspective, current studies have shown the opportunity to adapt the anatomical variability to one single atlas template with image registrations. However, the anatomical transferred context is biased to the chosen template subject and the registration algorithm is limited to provide large deformation field to localize organ anatomy (156). Therefore, we aim to adapt the conventional information of eye organs with minimal biases and adapt the contextual variability across large scale of patients with robust registration algorithms following the work of (11; 42; 41).

Previous works have been developed an atlas platform with neuroimaging (64; 165). Aging brain atlases are built to visualize the variability of brain organs in both adults and infants population (100; 175; 113). Apart from looking into the anatomical characteristics, single atlas reference is chosen to perform segmentation with unsupervised settings (216; 135). Also, multiple atlas references are randomly picked and perform

registration between the subject moving scans to the multiple atlases' platform (84). Segmentation predictions are computed with joint label fusion using the guidance of multiple registered outputs. Apart from the applicational usage of atlas template, organ-specific atlas template has been proposed to adapt the multi-contrast characteristics and the significant shape variability across large scale of patients. Contrast-specific substructure template has also proposed to look into the fine-grained anatomical details of kidney substructure and stably generalize the variation of small substructures in single template. Furthermore, unbiased template brain atlases have been proposed using brain MRI to reveal the anatomical characteristics of populations across different perspectives such as ages and disease conditions. However, none of works have been proposed in generating unbiased atlas frameworks for the eye organs. Apart from the limited work in generating organ-specific unbiased atlas, challenges are also raised for the robustness of registration pipelines. Significant efforts have been demonstrated in image registration with metric-based and deep learning based approaches (5; 11). VoxelMorph have been proposed as the current network basis to perform deformable registration in unsupervised setting (11). However, the input for the network is needed to be affine aligned and such basis structure is limited to provide diffeomorphism for inverse transformation (42). It is challenging to well align the correspondence between the atlas template and the moving image (41). Therefore, a robust pipeline is demanded to generate an eye-specific template with minimal biases.

In this study, we propose a hierarchical registration approach to construct a unbiased atlas template in unsupervised setting and aim to increase the generalizability of adapting eye organ context across populations. With a total of 100 contrast- enhanced head CT scans without known for ocular diseases, we initial performed iterative registration to generate an unbiased average mapping with small portion of subject scans (20 scans). We further performed a hierarchical registration pipeline to leverage the coarse output from metric-based registration for further registration refinement with deep learning network. Furthermore, we introduce the probabilistic estimation to model the diffeomorphism and compute accurate transformation to enhance the stability of the registration across patients. The generated atlas target and moving subject scans are downsampled to input for the deep registration pipeline. The predicted deformation field is upsampled back to the atlas resolution and perform inverse transformation with atlas label for label evaluation as quantitative measures. Qualitative visualization is further demonstrated the convergence of unbiased template and the proposed registration performance in image level.

13.3 Methods

13.3.1 Initial Unbiased Template Generation

To adapt the variability of eye organs across patients, image registrations are performed to match the anatomical context to single spatial-defined template using different registration tools such as ANTs (5) and NiftyReg

(155). However, the contextual information of each organs from the registered output is then biased and have similar anatomical structure to the single template. Here, we introduce an unbiased template generation module to generate an initial coarse atlas with small portion of data samples and minimize the registration bias instead of using a single fixed template. Specifically, we first input 20 samples and directly generate an average map to coarsely align the morphological structure of head. Hierarchical metric-based registration (following with rigid, affine and deformable registrations) are performed with ANTs and compute an average template with all registered outputs. We use the computed average template in each epoch as the next fixed template and iteratively perform the same hierarchical procedures to the average template until the registration loss across all samples is converged. We hypothesis that such template has the minimal biases and is beneficial to further adapt the population characteristics of eye organs.

13.3.2 Hierarchical Registration Refinement

After generating the initial template with minimal bias, we perform registration to the remaining samples and aim to generalize the anatomical characteristics of eye organs across populations. However, with the visualization of checkerboard overlay in Figure 13.3, we found that the metric-based registration is limited to perform significant deformation to align the head boundaries and the orbital anatomy. Motivated by VoxelMorph and probabilistic network , we introduce a deep probabilistic model $def_w(a,f)$ that leverages convolutional neural network (CNN) with diffeomorphic integration and spatial transform layers. Such registration model is trained in unsupervised setting and demonstrates better ability to generate significant deformation towards specific organ anatomies. We define a and m as the 3D volumetric atlas image and moving images, and z as the latent representation that parametrizes the transformation function $\psi_z : \mathbf{R^3} \rightarrow \mathbf{R^3}$. In our registration scenario, the deformation field is defined as the following differential equation (ODE):

$$\frac{\partial \psi^t}{\partial t} = v(\psi^t), \tag{13.1}$$

where $\psi^0 = Id$ is the identity transformation and t corresponds to time. We compute the stationary velocity field v over $t = [0,1]$ and output the deformation field that are differentiable and invertible (). To model the diffeomorphism with CNN, we model the prior probability of z as following:

$$p(z) = M(z; 0, \sum_z), \tag{13.2}$$

where $M(\cdot; \mu, \sum)$ is the multivariate normal distribution that models with mean μ and covariance $\sum$. In our registration scenario, the representations z is defined as the stationary velocity field that correlates the

diffeomorphism through equation 1. With the basis of above modeling, we can then convert equation 2 to estimate the posterior registration probability with the fixed template and moving image as following:

$$p(z|a;m) = M(a;m \cdot \psi_z), \tag{13.3}$$

With Equation 3, we can approximately predict the diffeomorphic registration field ψ_z to warp moving image m to the atlas template a via MAP estimation. Furthermore, to evaluate $p(z|a;m)$, a variational approach is used and introduce an approximate estimation of the posterior probability $q_\psi(z|a;m)$ parameterized by ψ. We aim to minimize the difference of the predicted posterior probability $p(z|a;m)$ and $q_\psi(z|a;m)$ through KL divergence for unsupervised training as following:

$$\begin{aligned}
\mathscr{L} &= \min_{\phi} KL\left[q_\phi(z|a;m) \| p(z|a;m)\right] \\
&= \min_{\phi} E_q\left[log\ q_\phi(z|a;m) - log\ p(z|a;m)\right] \\
&= \min_{\phi} E_q\left[log\ q_\phi(z|a;m) - log\ p(z,a;m)\right] + log\ p(a;m) \\
&= \min_{\phi} KL\left[q_\phi(z|a;m) \| p(z)\right] - E_q\left[log\ p(z|a;m)\right]
\end{aligned} \tag{13.4}$$

13.4 Data and Experiments

13.4.1 Data and Parameters

To evaluate the unbiased atlas template and our proposed hierarchical registration pipeline, head CT volumetric scans from 1842 patients were retrieved in de-identified form from ImageVU with the approval of the Institutional Review Board. 100 subjects were retrospectively selected to generate and evaluate the atlas template. All selected subject scans consist of 4 organs ground-truth label, which corresponds to optic nerve, rectus muscle, globe and orbital fat. To generate the unbiased atlas, we selected 20 subject scans and resampled to isotropic resolution ($0.8mm \times 0.8mm \times 0.8mm$) for iterative registration with a dimension of $512 \times 512 \times 224$. The criteria of choosing subjects for unbiased template generation are based on the morphological characteristics of eye organs and the high-resolution characteristics of the volumetric image. For the probabilistic registration, both the atlas template and moving subject scans are downsampled to an isotropic voxel resolution of ($1.0mm \times 1.0mm \times 1.0mm$) with a dimension of $256 \times 256 \times 224$. The batch size is 1 and the learning rate is 0.0001 for the network initialization. The predicted deformation field are upsampled back to the original atlas resolution and compute inverse transform for label evaluation.

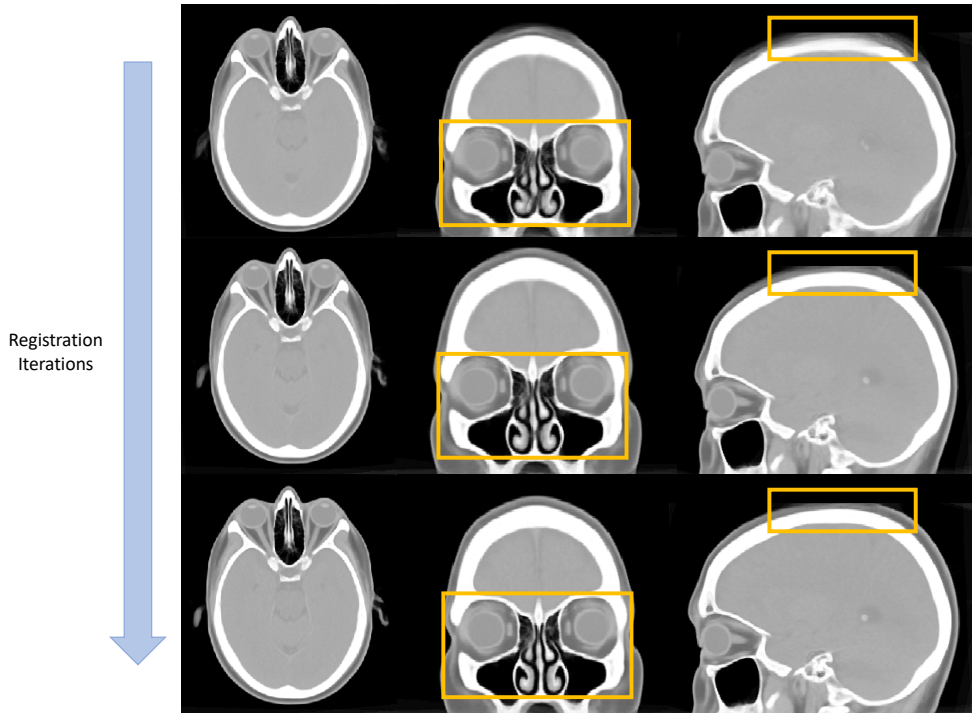


Figure 13.2: The qualitative representations to visualize the convergence of anatomical details across eye organs and the boundaries of head skull in the generation process of unbiased template.

13.4.2 Experiments

13.4.2.1 Iterative Template Comparison

We performed a conventional registration algorithm ANTs as our registration baseline and leverage small portion of subject scans to use the template generation tool (antsMultivariateTempalteConstruction) for coarse atlas construction. We initially applied both rigid and affine registration to align the anatomical location of the head skull and eye organs. SyN registration are then performed for deformable registration with similarity metric of cross-correlation (CC). In total, there are 4 resolution levels for registration and the number of iterations is defined as $100 \times 100 \times 70 \times 20$. The registration loss become converged when the templates updates to 6 iterations. Figure 13.2 further demonstrates the qualitative visualization of the unbiased template in tri-planar view.

13.4.2.2 Hierarchical Registration Refinement

After we generate the unbiased template, we further leverage the unbiased template to register the remaining samples for adapting the eye organ context and evaluate the registration pipeline. We first perform ANTs registration with the remaining 80 subjects following with the same hierarchical manner in atlas generation

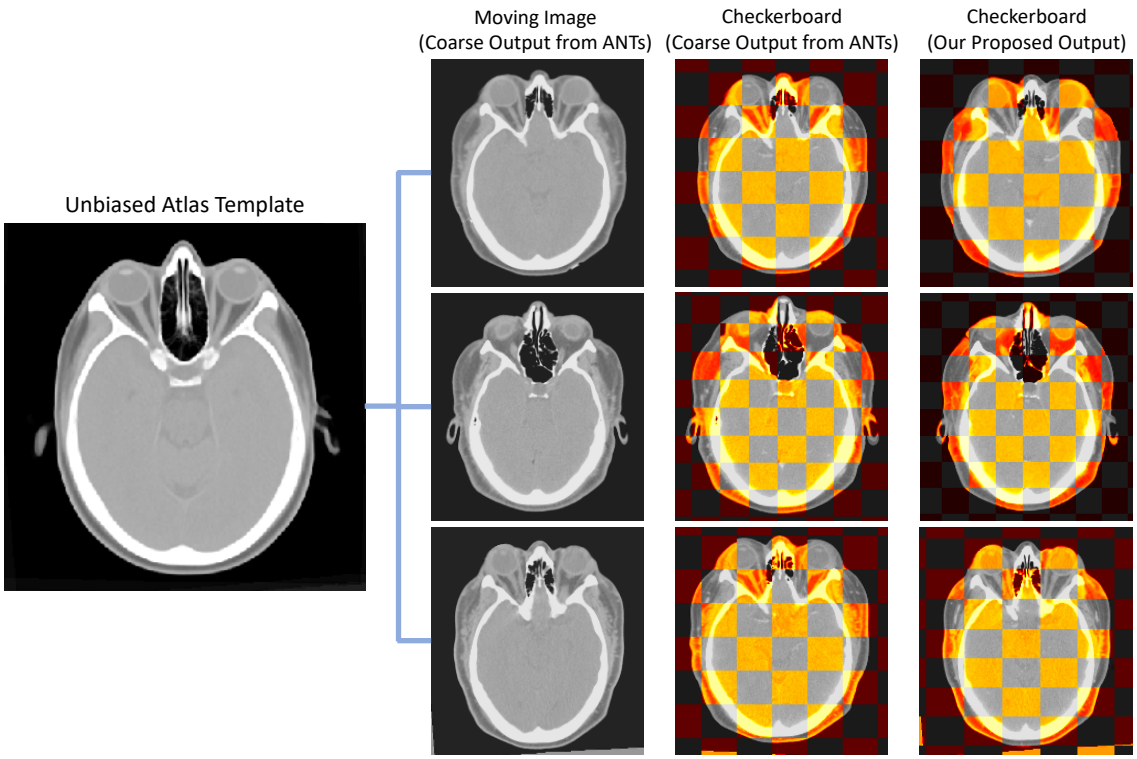


Figure 13.3: The qualitative representation that demonstrates the registration performance comparing with metric-based method as coarse output and the final output with our proposed hierarchical registration pipeline. The checkerboard overlay shows that the second stage refinement allows to further deform significantly and adapt the variability in boundaries and anatomical structure of organs.

Table 13.1: Quantitative evaluation of inverse transferred label for eye organs across all patients (*: p ¡ 0.001)

Methods	Optic Nerve	Rectus muscle	Globe	Orbital Fat	Mean Organ
ANTS ()	0.586±0.0532	0.634±0.0339	0.904±0.0253	0.782±0.0362	0.727±0.132
VoxelMorph ()	0.635±0.0420	0.687±0.0420	0.920±0.0274	0.790±0.0343	0.758±0.140
Ours	**0.651±0.0358**	**0.712±0.0301**	**0.931±0.0227**	**0.810±0.0319**	**0.776±0.115***

(rigid, affine and deformable). We then perform deep learning registration algorithm to enhance the contextual matching in eye organs and the skull boundaries. We perform VoxelMorph as our baseline network to concentrate on enhancing the registration performance. We further integrate the probabilistic ideas with the basis of VoxelMorph to smooth the deformable field and prevent the over-deformation across the eye organs. We compare the registration performance by computing the inverse transform of atlas label with the predicted deformation field and transform back to the moving image space. Dice Coefficient is used to measure the overlapping regions between the transferred label and ground truth labels.

13.5 Results

To minimize the biases of using single atlas template for registration, we want to first investigate the quality and following with the effectiveness of the unbiased atlas template. We evaluate the quality of the unbiased template with qualitative visualization. From Figure 13.3, with the increase of registration iterations, significant changes can be seen in particular anatomies within the orange bounding boxes. The skull boundaries become clearer and the contrast of boundaries is enhanced to provide better guidance to localize the head skull after several iterations. Furthermore, some of the eye organs, especially the retus muscle and the globe, is demonstrated with enhancing contrast level in both the boundaries and the morphology of the organ structure. Regions near the eye organs also demonstrate to be stably localized with clear structural characteristics comparing with the initial iterations. After generating the minimal biased atlas template from small portion of data, we want to further enhance the generalizability of the atlas template and adapt the anatomical context with remaining samples using different registration approaches. We first compare the registration performance qualitatively and manually visualize the quality of the registered eye organs. As shown in Figure 13.3, the registered outputs from ANTs demonstrates a fair registration quality that localize the boundaries of head skull and can align some of the morphological characteristics in eye organs. However, when we further perform the checkerboard overlay to the atlas template, the boundaries of eye organs are not well aligned, especially for the globe and the orbital muscle nearby. The deformable registration of ANTs is limited to generate significant deformation to match the atlas context when significant difference in eye organs morphology or skull shape. By using the hierarchical registration refinement, the anatomical context of both the head skull and eye organs is comparable to the atlas-defined space without over-deformation. The checkerboard with our proposed registration method demonstrates a more distinctive appearance and alignment to the eye organs and match the boundaries well. The contrastive and morphological characteristics can be well transferred with stability across all populations. Apart from the qualitative measures, we further inversely applied the predicted deformation field to the atlas label and transformed back to the original moving image space for label similarity evaluation in Table 13.1. By using ANTs registration only, the transferred label across all organs demonstrates with a mean Dice of 0.727 comparing to the atlas label. After we add the second stage process for registration refinement, the label transferred performance significantly increase from 0.727 to 0.758 across all organs with the basis of VoxelMorph network. Such significant increases demonstrate the possibility of refining coarse output from metric-based registration and further performed deformation to adapt the variability of organs across population. Furthermore, by introducing the probabilistic estimation into registration network, the label transfer performance outperforms VoxelMorph and demonstrates another significant increase of Dice 0.758 to 0.776.

13.6 Discussion and Conclusion

With the qualitative and quantitative representation above, the contrastive and morphological context of the eye organs are stably localized with minimal bias using the hierarchical registration pipeline. The label transfer performance in all sub-organ regions of the eye demonstrates significant improvement with decrease of variance in trend. The deep learning registration network in the second is trained in unsupervised setting and leverage the estimation of probability distribution to enhance the accuracy of the predicted deformation field. With the opportunity of integrating metric-based registration and deep learning registration as hierarchical pipeline, we leverage the characteristics of deep learning network characteristics to generate large deformation and solve the limitation of metric-based registration to adapt morphological variability of both eye organs and the head skull. Furthermore, the probabilistic estimation for registration demonstrates the diffeomorphism in deep learning networks and enhance the stability by comparing the probabilistic representation in the latent space. In this paper, we constructed an unbiased anatomical reference to localize the context of eye organs and further propose a hierarchical registration pipeline integrating both metric-based and deep learning registration. The unbiased average mapping demonstrated the contrastive characteristics of both eye organs and well localized with skull morphology across patients. The atlas target stably adapted the eye organs information with the illustration of checkerboard overlay and demonstrate significant improvement in label transfer performance with our proposed registration pipeline. We aim to create a minimal bias average template that adapt multi-modality imaging for eye organs as our future long- term goal and analyze variability across demographics.

CHAPTER 14

Conclusion & Future Works

14.1 Explore Deep Learning Optimization in Medical Image Segmentation

Medical image segmentation is a important perspective for clinicians to access contextual information of each organ ROIs with pixel/voxel-wise guidance. To enhance the efficiency for further investigation, substantial efforts have integrated deep learning into downstream segmentation task. In this work, we first propose a coarse-to-fine network to adapt deep learning strategy and predict refine segmentation with limited well-annotated samples. Such hierarchical 2-stage network adapts multi-scale knowledge (e.g., anatomical prior as additional channel input) in anatomy and innovate an effective generic backbone for hierarchical segmentation (Chapter 2). However, we observe that the model generalizability is limited across imaging collected from different domains (e.g., contrast-enhanced CT vs non-contrast CT). We further leverage our proposed hierarchical network as the initial backbone and adapt unseen domain samples in unsupervised setting. We employ teacher-student network architecture to enhance the prediction certainty in target domain with unsupervised learning (Chapter 3). In addition to the challenges in domain shift, we tackle the limitation of large population in unlabeled samples. We propose a semi-supervised learning strategy that quantifies the pseudo segmentation quality from unlabeled data as a regression score. We backpropagate it as an unsupervised loss function for training (Chapter 4) to adapt large population of unlabeled samples from different domains.

Instead of only focusing on the data and training strategies, a big step is pushed forward to adapt the concepts of transformer from NLP domain into computer vision. Vision transformers have been introduced to enhance the robustness in different medical tasks. However, we observe that the key contributions of vision transformer can be achieved with CNN modules. Here, we provide a new exploratory direction to investigate the effectiveness of large kernel convolution for medical image segmentation, which simulate the capabilities of vision transformer at the same time (Chapter 5). We further investigate the effectiveness of scaling up the kernel size for convolution. We found that the segmentation performance become saturated or even degraded if the kernel size is too large (e.g., $21 \times 21 \times 21$). We proposed an explainable optimization strategy that models the behavior of spatial frequency in our human visual systems to weight the learning convergence of each elements. The learning convergence of the large kernel convolution are then rescaled from local to global, which further enhances the segmentation performance, instead of degradation.

14.1.1 Technical Innovation

- We innovated a coarse-to-fine network as the genric backbone to adapt multi-scale prior knowledge and refine the intermediate segmentation prediction with supervised learning.
- We leveraged our proposed coarse-to-fine network to generalize unseen imaging domains with significant shift in contrast.
- In addition to addressing the domain shift challenge in data perspective, we adapted large population of unlabeled imaging by quantifying the pseudo segmentation quality as an innovative semi-supervised learning technique.
- For the generic backbone design in deep neural networks, we provided a new exploratory direction to revisit the concepts of convolution and simulated the hierarchical transformer behaviors with large kernel sizes.
- With the observation of how the deep neural network learns, we modeled the receptive field of each neuron as the spatial frequency in our human visual behavior and provided feasibility to adapt significantly large kernels (e.g., $21 \times 21 \times 21$) for volumetric segmentation with robust performance.

14.1.2 Clinical Impact

- We adapted deep learning strategy to enhance the convenience for clinical teams to obtain segmentation across large population samples by leveraging limited samples with good quality annotations.
- We generalized deep neural network models to unseen domain samples (from contrast-enhanced domain to non-contrast domain) without using additional labels for training and demonstrate robust performance.
- We adapted large population of unlabeled samples (in real clinical scenario) to enhance the model generalizability across imaging from different domains.
- We introduced a simple CNNs design and revisited the concepts of large kernel convolution as the generic backbone for medical image segmentation to generate robust segmentation with less parameters.
- We introduced an explainable optimization by modeling the spatial frequency in the human visual behavior to enhance the learning convergence of large kernel convolution and demonstrated the interpretabillity of "how the model learns" with the receptive field.

14.1.3 Future Directions

With the trends from CNNs to vision transformers, we already simulate the scaling behavior and the large receptive field characteristics with our proposed network. However, with the intrinsic ability of limited inductive bias in vision transformers, self-supervised learning become a "must" step to perform and pretrain the model as prior knowledge to enhance both the stability and robustness of the model. Currently, significant efforts have been put to adapt transformer-based self-supervised learning. However, limited studies have been proposed to evaluate the ability of self-supervised learning in large kernel convolution. Meanwhile, it is challenging to directly apply the transformer-based self-supervised strategies into large kernel convolution. Potential direction of innovating self-supervised strategies with large kernel convolution can be one of our future work.

Another challenges in large kernel convolution is the computation efficiency. We observe that the computation time significantly increases if we continue to scale up the kernel size, thus reducing the efficiency for both training and inference stage. To tackle such limitations, sparse neural network can be another potential direction to adapt sparsity in generic network design or in training strategies. With fewer model parameters in both training and inference, we hypothesize that the computation efficiency can be signficantly improved.

14.2 Enhance Feature Interpretability for Medical Segmentation Network

14.2.1 Summary

While we have put significant efforts to enhance the robustness of deep neural network models for medical image segmentation, another important aspect is to know "why the model works". While we can quantify the model performance with different evaluation metric, it is challenging to interpret the model capability in the feature level, especially for multi-organ segmentation. One of the self-supervised learning strategy called contrastive learning demonstrates the feasibility of defining the extracted features into corresponding clusters. In this work, we first adapt the concepts of contrastive learning and introduce multiple semantic meanings as image-level labels to classify the features into corresponding clusters, which hypothesize to be beneficial for downstream segmentation task (Chapter 7). However, when we further investigate the fine-grain details from the imaging samples, a wide range of contrast level difference is demonstrated across organ ROIs. The fixed image-level labels are limited to provide dynamic correlation to classify the features in the latent space. Therefore, instead of leveraging image-level labels, we propose an adaptive contrastive learning strategy that leverage the contrast correlation between organs to evaluate the cosine distance between pairwise features. Even thought the pairwise features are in different modalities, the cosine distance are re-weighted with the contrast difference between organ regions (Chapter 8). Both approaches demonstrate the feasibility of generating meaningful latent space to provide better initialization for downstream segmentation task.

14.2.2 Technical Innovation

- We introduce multiple semantic meanings into the learned features with contrastive learning to enhance the feature interpretability. Meanwhile, such defined latent space is beneficial to enhance model generalizability for downstream segmentation.

- With the introduction of multi-contrast phases CT, we extend the concepts of contrastive learning with adaptive data-driven knowledge. The learned latent space is defined based on the contrast difference between organ ROIs, which adaptively controls the cosine distances between organ-specific features with respect to the contrast correlation.

14.2.3 Clinical Impact

- We enhanced the feature interpretability of the model by introducing semantic meanings to classify features into separable clusters. Performance can be evaluated and explained by interpreting the quality of the defined latent space.

- We further adapted the multi-contrast characteristics in CT by dynamically defining the learn feature with respect to the contrast correlation. Instead of integrating semantic meanings, we weighted the cosine distance between pairwise features to generate a meaningful latent space for generalizing multi-contrast imaging.

14.2.4 Future Directions

We already observed that defining the learned representations into clusters with multiple semantic meanings can benefit both the interpretability and generalizability of segmentation models. With the inspiration of current multi-modal approaches, such as CLIP, we can further extend to investigate the effectiveness of language-wise semantic meanings into segmentation network as one of our future direction in model explainabilty. Furthermore, our proposed contrastive learning strategy are limited to leverage the anatomical prior knowledge as our additional input channel for training. With the inspiration of segmenting anything model, a potential thought of leveraging pseudo anatomical context as the prompt prior knowledge can be another alternative future direction to enhance the interpretability of segmentation model.

14.3 Generalize Population-wise Biomarkers with Organ-Specific Atlas

14.3.1 Summary

After we enhance both the model interpretability and its robustness for organ segmentation towards medical imaging in different domains, we can then perform organ segmentation across large population of unlabeled

samples and observe the population characteristics of specific organs. Meanwhile, with the quantitative measures from the predicted segmentation of certain organs (e.g., kidney), substantial variation is demonstrated due to different demographics across population. Such variation limits the feasibility to standardize the biomarkers of the organs in specific condition (e.g., healthy, metastasis). Previous efforts only demonstrates to standardize the biomarkers in cellular perspective, while limited efforts have been put to adapt multi-scale knowledge with volumetric imaging. Here, we propose a multi-contrast atlas template to generalize the both the contrast and the morphological charactersitics of kidney organs for population analysis. We further observe that there are significant morphological difference in both left and right kidneys across sex (Chapter 10 & 11). We leverage deep learning based algorithm to transfer the anatomical context of kidney substructure in arterial phase CT and generate a kidney substructure atlas to further investigate the anatomical context of each substructure in healthy condition (Chapter 12). In addition to kidney organs, we observe that the eye organs also demonstrate substantial variation in each sub-organ region such as the optic nerve and the orbital. However, previous generated atlas are completely biased to one fixed template that defined with our chosen criteria. We introduce a combined registration framework with metric-based and deep learning based registration to generate a minimal biased eye atlas template and generalize the population feature of eye organs in healthy condition (Chapter 13).

14.3.2 Technical Innovation

- We constructed multi-contrast phase atlas template to generalize the population biomarkers of kidney organs in healthy condition.
- We extended to specialize arterial phase CT to generalize the population biomarkers of kidney substructures in healthy condition.
- Apart from kidney organs, we constructed unbiased atlas template to generalize the population features of eye organs with both metric-based and deep learning based registration.

14.3.3 Clinical Impact

- We identified and distinguished the population biomarkers of kidney organs in healthy conditions across demographics (e.g., sex).
- We further investigated the population characteristics of kidney substructures in healthy condition (e.g., renal cortex, medulla, pelvicalyceal system).
- We constructed an unbiased template to generalize the population characteristics of eye organs and quantify the volumetric biomarkers of eye organs (e.g., orbital fat, optic nerve, globe, retus muscle) in

healthy condition.

14.3.4 Future Directions

With the previous efforts of generating atlas template, one main limitation is to only demonstrate the atlas applicability in computed tomography. The first future direction is to adapt and generalize multi-domain (e.g., MRI in different protocols) anatomical knowledge on single atlas template. However, the resolution of through-plane axis significantly varies, which limits to provide sufficient context for visualizing the distinctive appearance of specific organ ROIs. Integrating robust super-resolution algorithms may further provide additional opportunities to extend and adapt large population of low-resolution imaging.

In addition to the data perspective, having a robust image registration algorithm to transfer anatomical context is also vitally important. Currently, from the qualitative result from our experiments, we leverage metric-based registration (e.g., DEEDS) for abdominal organ registration, rather than using deep learning based algorithms due to over-deformation of neighboring anatomies and the instability performance with domain shift. Instead of adapting metric-based registration by registering each subject scan one-by-one, a potential future opportunity is to adapt deep learning strategy to generate an atlas template, which directly to match the a large population of subject scans in unsupervised setting, thus to facilitate the computation efficiency and enhance the effectiveness of transferring anatomical context.

References

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021.

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[3] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[4] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021.

[5] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[7] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[8] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 541–549. Springer, 2019.

[9] Wenjia Bai, Hideaki Suzuki, Jian Huang, Catherine Francis, Shuo Wang, Giacomo Tarroni, Florian Guitton, Nay Aung, Kenneth Fung, Steffen E Petersen, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nature medicine*, 26(10):1654–1662, 2020.

[10] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

[11] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[12] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[13] Shunxing Bao, Can Cui, Jia Li, Yucheng Tang, Ho Hin Lee, Ruining Deng, Lucas W Remedios, Xin Yu, Qi Yang, Sophie Chiron, et al. Topological-preserving membrane skeleton segmentation in multiplex immunofluorescence imaging. In *Medical Imaging 2023: Digital and Computational Pathology*, volume 12471, pages 62–71. SPIE, 2023.

[14] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[15] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

[16] F Edward Boas, Dominik Fleischmann, et al. Ct artifacts: causes and reduction techniques. *Imaging Med*, 4(2):229–240, 2012.

[17] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. *arXiv preprint arXiv:2204.09817*, 2022.

[18] David J Brenner and Eric J Hall. Computed tomography—an increasing source of radiation exposure. *New England journal of medicine*, 357(22):2277–2284, 2007.

[19] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer methods and programs in biomedicine*, 104(3):e158–e177, 2011.

[20] Leon Y Cai, Ho Hin Lee, Nancy R Newlin, Cailey I Kerley, Praitayini Kanakaraj, Qi Yang, Graham W Johnson, Daniel Moyer, Kurt G Schilling, Francois Rheault, et al. Convolutional-recurrent neural networks approximate diffusion tractography from t1-weighted mri and associated anatomical context. *bioRxiv*, pages 2023–02, 2023.

[21] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[23] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

[24] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020.

[25] Anirudh Chandrashekar, Ashok Handa, Natesh Shivakumar, Pierfrancesco Lapolla, Vicente Grau, and Regent Lee. A deep learning approach to generate contrast-enhanced computerised tomography angiography without the use of intravenous contrast agents. *arXiv preprint arXiv:2003.01223*, 2020.

[26] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505, 2020.

[27] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170:446–455, 2018.

[28] Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical image analysis*, 36:135–146, 2017.

[29] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[30] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[31] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[32] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[34] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[35] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[36] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern Recognition*, 113:107826, 2021.

[37] Jooae Choe, Hye Jeon Hwang, Joon Beom Seo, Sang Min Lee, Jihye Yun, Min-Ju Kim, Jewon Jeong, Youngsoo Lee, Kiok Jin, Rohee Park, et al. Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct. *Radiology*, 302(1):187–197, 2022.

[38] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775. Curran

Associates, Inc., 2020.
[39] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
[40] Frederic Commandeur, Markus Goeller, Julian Betancur, Sebastien Cadet, Mhairi Doris, Xi Chen, Daniel S Berman, Piotr J Slomka, Balaji K Tamarappoo, and Damini Dey. Deep learning for quantification of epicardial and thoracic adipose tissue from non-contrast ct. *IEEE transactions on medical imaging*, 37(8):1835–1846, 2018.
[41] Adrian Dalca, Marianne Rakic, John Guttag, and Mert Sabuncu. Learning conditional deformable templates with convolutional networks. *Advances in neural information processing systems*, 32, 2019.
[42] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019.
[43] Adrian V Dalca, Andreea Bobu, Natalia S Rost, and Polina Golland. Patch-based discrete registration of clinical brain images. In *Patch-Based Techniques in Medical Imaging: Second International Workshop, Patch-MI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings 2*, pages 60–67. Springer, 2016.
[44] Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.
[45] Xiang Deng and Guangwei Du. 3d segmentation in the clinic: a grand challenge ii-liver tumor segmentation. In *MICCAI workshop*, 2008.
[46] Laurent Dercle, Jingchen Ma, Chuanmiao Xie, Ai-ping Chen, Deling Wang, Lyndon Luk, Paul Revel-Mouroz, Philippe Otal, Jean-Marie Peron, Hervé Rousseau, et al. Using a single abdominal computed tomography image to differentiate five contrast-enhancement phases: A machine-learning algorithm for radiomics-based precision medicine. *European journal of radiology*, 125:108850, 2020.
[47] Xiaohan Ding, Honghao Chen, Xiangyu Zhang, Kaiqi Huang, Jungong Han, and Guiguang Ding. Re-parameterizing your optimizers rather than architectures. *arXiv preprint arXiv:2205.15242*, 2022.
[48] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022.
[49] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021.
[50] Tilman Donath, Franz Pfeiffer, Oliver Bunk, Christian Grünzweig, Eckhard Hempel, Stefan Popescu, Peter Vock, and Christian David. Toward clinical x-ray phase-contrast ct: demonstration of enhanced soft-tissue contrast in human specimen. *Investigative radiology*, 45(7):445–452, 2010.
[51] Suyu Dong, Gongning Luo, Clara Tam, Wei Wang, Kuanquan Wang, Shaodong Cao, Bo Chen, Henggui Zhang, and Shuo Li. Deep atlas network for efficient 3d left ventricle segmentation on echocardiography. *Medical image analysis*, 61:101638, 2020.
[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
[53] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.
[54] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020.
[55] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7:99065–99076, 2019.
[56] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv*

preprint arXiv:1804.10916, 2018.

[57] Michal Drozdzal, Gabriel Chartrand, Eugene Vorontsov, Mahsa Shakeri, Lisa Di Jorio, An Tang, Adriana Romero, Yoshua Bengio, Chris Pal, and Samuel Kadoury. Learning normalized inputs for iterative estimation in medical image segmentation. *Medical image analysis*, 44:1–13, 2018.

[58] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 58–68. Springer, 2021.

[59] Daniel C Elton, Evrim B Turkbey, Perry J Pickhardt, and Ronald M Summers. A deep learning system for automated kidney stone detection and volumetric segmentation on noncontrast ct scans. *Medical Physics*, 49(4):2545–2554, 2022.

[60] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R Laird, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, 26(8):3508–3526, 2016.

[61] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient belief propagation for early vision. *International journal of computer vision*, 70:41–54, 2006.

[62] Zahra Sedghi Gamechi, Lidia R Bons, Marco Giordano, Daniel Bos, Ricardo PJ Budde, Klaus F Kofoed, Jesper Holst Pedersen, Jolien W Roos-Hesselink, and Marleen de Bruijne. Automated 3d segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced ct. *European radiology*, 29(9):4613–4623, 2019.

[63] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[64] Ali Gholipour, Caitlin K Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, Alireza Akhondi-Asl, Onur Afacan, Cynthia M Ortinau, Sean Clancy, Catherine Limperopoulos, Edward Yang, et al. A normative spatiotemporal mri atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Scientific reports*, 7(1):476, 2017.

[65] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.

[66] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. Niftynet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158:113–122, 2018.

[67] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. Dense image registration through mrfs and efficient linear programming. *Medical image analysis*, 12(6):731–741, 2008.

[68] Ben Glocker, Robert Robinson, Daniel C Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

[69] O Graf, GW Boland, AL Warshaw, C Fernandez-del Castillo, PF Hahn, and PR Mueller. Arterial versus portal venous helical ct for revealing pancreatic adenocarcinoma: conspicuity of tumor and critical vascular anatomy. *AJR. American journal of roentgenology*, 169(1):119–123, 1997.

[70] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[71] Yunhui Guo, Yandong Li, Liqiang Wang, and Tajana Rosing. Depthwise convolution is all you need for learning multiple visual domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8368–8375, 2019.

[72] Zhimeng Han, Muwei Jian, and Gai-Ge Wang. Convunext: An efficient convolution neural network for medical image segmentation. *Knowledge-Based Systems*, page 109512, 2022.

[73] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022.

[74] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman,

Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[75] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[76] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[77] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[78] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. Transformers in medical image analysis: A review. *Intelligent Medicine*, 2022.

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[80] Jay P Heiken, James A Brink, Bruce L McClennan, Stuart S Sagel, Tamara M Crowe, and Mary V Gaines. Dynamic incremental ct: effect of volume and concentration of contrast material and patient weight on hepatic enhancement. *Radiology*, 195(2):353–357, 1995.

[81] Mattias P. Heinrich. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks, 2019.

[82] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE transactions on medical imaging*, 32(7):1239–1248, 2013.

[83] Mattias Paul Heinrich, Mark Jenkinson, Bartlomiej W Papież, Michael Brady, and Julia A Schnabel. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *International conference on medical image computing and computer-assisted intervention*, pages 187–194. Springer, 2013.

[84] Mattias P Heinrich, Oskar Maier, and Heinz Handels. Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. *VISCERAL Challenge@ ISBI*, 1390:27, 2015.

[85] Mattias P Heinrich, Ozan Oktay, and Nassim Bouteldja. Obelisk-net: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions. *Medical image analysis*, 54:1–9, 2019.

[86] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020.

[87] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[88] Ho Hin Lee, Yucheng Tang, Qi Yang, Xin Yu, Shunxing Bao, Leon Y Cai, Lucas W Remedios, Bennett A Landman, and Yuankai Huo. Semantic-aware contrastive learning for multi-object medical image segmentation. *arXiv e-prints*, pages arXiv–2106, 2021.

[89] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 440–450, 2022.

[90] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16291–16301, 2021.

[91] Mu Hu, Junyi Feng, Jiashen Hua, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Online convolutional re-parameterization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 568–577, 2022.

[92] Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–490. Springer, 2021.

[93] Chao Huang, Hu Han, Qingsong Yao, Shankuan Zhu, and S Kevin Zhou. 3d u^2-net: A 3d universal u-net for multi-domain medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–299. Springer, 2019.

[94] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[95] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[96] Yuankai Huo, Zhoubing Xu, Shunxing Bao, Camilo Bermudez, Hyeonsoo Moon, Prasanna Parvathaneni, Tamara K Moyo, Michael R Savona, Albert Assad, Richard G Abramson, et al. Splenomegaly segmentation on multi-modal mri using deep convolutional networks. *IEEE transactions on medical imaging*, 38(5):1185–1196, 2018.

[97] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025, 2018.

[98] Yuankai Huo, Zhoubing Xu, Yunxi Xiong, Katherine Aboud, Prasanna Parvathaneni, Shunxing Bao, Camilo Bermudez, Susan M Resnick, Laurie E Cutting, and Bennett A Landman. 3d whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119, 2019.

[99] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[100] George Andrew James, Onder Hazaroglu, and Keith A Bush. A human brain atlas derived via n-cut parcellation of resting-state and task-based fmri data. *Magnetic resonance imaging*, 34(2):209–218, 2016.

[101] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[102] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022.

[103] Qiran Jia and Hai Shu. Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II*, pages 3–14. Springer, 2022.

[104] Yun Jiang, Yuan Zhang, Xin Lin, Jinkun Dong, Tongtong Cheng, and Jing Liang. Swinbts: a method for 3d multimodal brain tumor segmentation using swin transformer. *Brain Sciences*, 12(6):797, 2022.

[105] Oscar Jimenez-del Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *IEEE transactions on medical imaging*, 35(11):2459–2475, 2016.

[106] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[107] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer, 2017.

[108] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[109] Nahum Kiryati and Yuval Landau. Dataset growth in medical image analysis research. *Journal of Imaging*, 7(8), 2021.

[110] N Kovačević, JT Henderson, E Chan, N Lifshitz, J Bishop, AC Evans, RM Henkelman, and XJ Chen. A

three-dimensional mri atlas of the mouse brain with estimates of the average and variability. *Cerebral cortex*, 15(5):639–645, 2005.

[111] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

[112] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[113] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011.

[114] Janus J Kulikowski, S Marčelja, and Peter O Bishop. Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. *Biological cybernetics*, 43(3):187–198, 1982.

[115] Ravi Kumar and Munish Rattan. Analysis of various quality metrics for medical image processing. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(11):137–144, 2012.

[116] Marc Lalonde, Mario Beaulieu, and Langis Gagnon. Fast and robust optic disc detection using pyramidal decomposition and hausdorff-based template matching. *IEEE transactions on medical imaging*, 20(11):1193–1200, 2001.

[117] B Landman, Z Xu, J Igelsias, M Styner, T Langerak, and A Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[118] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[119] Caltech-UW TMC Cai Long lcai@ caltech. edu 21 b Shendure Jay 9 Trapnell Cole 9 Lin Shin shinlin@ uw. edu 2 e Jackson Dana 9, UCSD TMC Zhang Kun kzhang@ bioeng. ucsd. edu 15 b Sun Xin 15 Jain Sanjay 24 Hagood James 25 Pryhuber Gloria 26 Kharchenko Peter 8, California Institute of Technology TTD Cai Long lcai@ caltech. edu 21 b Yuan Guo-Cheng 35 Zhu Qian 35 Dries Ruben 35, Harvard TTD Yin Peng peng_yin@ hms. harvard. edu 36 37 b Saka Sinem K. 36 37 Kishi Jocelyn Y. 36 37 Wang Yu 36 37 Goldaracena Isabel 36 37, Purdue TTD Laskin Julia jlaskin@ purdue. edu 10 b Ye DongHye 10 38 Burnum-Johnson Kristin E. 39 Piehowski Paul D. 39 Ansong Charles 39 Zhu Ying 39, Stanford TTD Harbury Pehr harbury@ stanford. edu 11 b Desai Tushar 40 Mulye Jay 11 Chou Peter 11 Nagendran Monica 40, et al. The human body at cellular resolution: the nih human biomolecular atlas program. *Nature*, 574(7777):187–192, 2019.

[120] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[121] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076*, 2022.

[122] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman. 3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In *The Eleventh International Conference on Learning Representations*, 2023.

[123] Ho Hin Lee, Quan Liu, Shunxing Bao, Qi Yang, Xin Yu, Leon Y Cai, Thomas Li, Yuankai Huo, Xenofon Koutsoukos, and Bennett A Landman. Scaling up 3d kernels with bayesian frequency reparameterization for medical image segmentation. *arXiv preprint arXiv:2303.05785*, 2023.

[124] Ho Hin Lee, Alberto Santamaria-Pang, Jameson Merkow, Ozan Oktay, Fernando Pérez-García, Javier Alvarez-Valle, and Ivan Tarapov. Region-based contrastive pretraining for medical image retrieval with anatomic query. *arXiv preprint arXiv:2305.05598*, 2023.

[125] Ho Hin Lee, Yucheng Tang, Shunxing Bao, Richard G Abramson, Yuankai Huo, and Bennett A Landman. Rap-net: Coarse-to-fine multi-organ segmentation with single random anatomical prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1491–1494. IEEE, 2021.

[126] Ho Hin Lee, Yucheng Tang, Shunxing Bao, Yan Xu, Qi Yang, Xin Yu, Agnes B Fogo, Raymond Harris, Mark P de Caestecker, Jeffery M Spraggins, et al. Supervised deep generation of high-resolution arterial phase computed tomography kidney substructure atlas. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 736–743. SPIE, 2022.

[127] Ho Hin Lee, Yucheng Tang, Shunxing Bao, Qi Yang, Xin Yu, Kevin L Schey, Jeffery M Spraggins,

Yuankai Huo, and Bennett A Landman. Unsupervised registration refinement for generating unbiased eye atlas. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 470–476. SPIE, 2023.

[128] Ho Hin Lee, Yucheng Tang, Riqiang Gao, Qi Yang, Xin Yu, Shunxing Bao, James G Terry, J Jeffrey Carr, Yuankai Huo, and Bennett A Landman. Pseudo-label guided multi-contrast generalization for non-contrast organ-aware segmentation. *arXiv preprint arXiv:2205.05898*, 2022.

[129] Ho Hin Lee, Yucheng Tang, Han Liu, Yubo Fan, Leon Y Cai, Qi Yang, Xin Yu, Shunxing Bao, Yuankai Huo, and Bennett A Landman. Adaptive contrastive learning with dynamic correlation for multi-phase organ segmentation. *arXiv preprint arXiv:2210.08652*, 2022.

[130] Ho Hin Lee, Yucheng Tang, Olivia Tang, Yuchen Xu, Yunqiang Chen, Dashan Gao, Shizhong Han, Riqiang Gao, Michael R Savona, Richard G Abramson, et al. Semi-supervised multi-organ segmentation through quality assurance supervision. In *Medical Imaging 2020: Image Processing*, volume 11313, page 113131I. International Society for Optics and Photonics, 2020.

[131] Ho Hin Lee, Yucheng Tang, Kaiwen Xu, Shunxing Bao, Agnes B Fogo, Raymond Harris, Mark P de Caestecker, Mattias Heinrich, Jeffery M Spraggins, Yuankai Huo, et al. Construction of a multi-phase contrast computed tomography kidney atlas. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 501–508. SPIE, 2021.

[132] Ho Hin Lee, Yucheng Tang, Kaiwen Xu, Shunxing Bao, Agnes B Fogo, Raymond Harris, Mark P de Caestecker, Mattias Heinrich, Jeffrey M Spraggins, Yuankai Huo, et al. Multi-contrast computed tomography healthy kidney atlas. *Computers in Biology and Medicine*, 146:105555, 2022.

[133] Ho Hin Lee, Yucheng Tang, Qi Yang, Xin Yu, Shunxing Bao, Leon Y Cai, Lucas W Remedios, Bennett A Landman, and Yuankai Huo. Semantic-aware contrastive learning for multi-object medical image segmentation. *arXiv preprint arXiv:2106.01596*, 2021.

[134] Wenhui Lei, Wei Xu, Ran Gu, Hao Fu, Shaoting Zhang, Shichuan Zhang, and Guotai Wang. Contrastive learning of relative position regression for one-shot object localization in 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 155–165. Springer, 2021.

[135] Dengwang Li, Li Liu, Jinhu Chen, Hongsheng Li, Yong Yin, Bulat Ibragimov, and Lei Xing. Augmenting atlas-based liver segmentation for radiotherapy treatment planning by incorporating image features proximal to the atlas contours. *Physics in Medicine & Biology*, 62(1):272, 2016.

[136] Hao Li, Yang Nan, Javier Del Ser, and Guang Yang. Large-kernel attention for 3d medical image segmentation. *arXiv preprint arXiv:2207.11225*, 2022.

[137] Hao Li, Yang Nan, and Guang Yang. Lkau-net: 3d large-kernel attention-based u-net for automatic mri brain tumor segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, pages 313–327. Springer, 2022.

[138] Jun Li, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A Landman, and S Kevin Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, page 102762, 2023.

[139] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[140] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. Recent developments of content-based image retrieval (cbir). *Neurocomputing*, 452:675–689, 2021.

[141] Ailiang Lin, Jiayu Xu, Jinxing Li, and Guangming Lu. Contrans: Improving transformer with convolutional attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 297–307. Springer, 2022.

[142] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 475–485. Springer, 2020.

[143] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022.

[144] Weizhe Liu, David Ferstl, Samuel Schulter, Lukas Zebedin, Pascal Fua, and Christian Leistner. Domain adaptation for semantic segmentation via patch-wise contrastive learning, 2021.

[145] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021.

[146] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[147] Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yang Liu, Jiayu Zhou, and Yao Zhao. Margin preserving self-paced contrastive learning towards domain adaptation for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(2):638–647, 2022.

[148] Peter Lorenzen, Marcel Prastawa, Brad Davis, Guido Gerig, Elizabeth Bullitt, and Sarang Joshi. Multi-modal image set registration and atlas formation. *Medical image analysis*, 10(3):440–451, 2006.

[149] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[150] Xuesong Lu, Qinlan Xie, Yunfei Zha, and Defeng Wang. Fully automatic liver segmentation combining multi-dimensional graph cut with shape information in 3d ct images. *Scientific reports*, 8(1):10700, 2018.

[151] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[152] Judith Mackay, George A Mensah, and Kurt Greenlund. *The atlas of heart disease and stroke*. World Health Organization, 2004.

[153] Jian-Xun Mi, An-Di Li, and Li-Fang Zhou. Review study of interpretation methods for future interpretable machine learning. *IEEE Access*, 8:191969–191985, 2020.

[154] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[155] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[156] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 211–221. Springer, 2020.

[157] Eric D Morris, Ahmed I Ghanem, Ming Dong, Milan V Pantelic, Eleanor M Walker, and Carri K Glide-Hurst. Cardiac substructure segmentation with deep learning for improved cardiac sparing. *Medical physics*, 47(2):576–586, 2020.

[158] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[159] Gerd Muehllehner and Joel S Karp. Positron emission tomography. *Physics in Medicine & Biology*, 51(13):R117, 2006.

[160] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[161] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[162] Yoshiharu Nakayama, Kazuo Awai, Yoshinori Funama, Masahiro Hatemura, Masanori Imuta, Takeshi Nakaura, Da Ryu, Shoji Morishita, Shamima Sultana, Natsuko Sato, et al. Abdominal ct with low tube voltage: preliminary observations about radiation dose, contrast enhancement, image quality, and noise. *Radiology*, 237(3):945–951, 2005.

[163] Akash Nayak, Esha Baidya Kayal, Manish Arya, Jayanth Culli, Sonal Krishan, Sumeet Agarwal, and Amit Mehndiratta. Computer-aided diagnosis of cirrhosis and hepatocellular carcinoma using multi-phase abdomen ct. *International journal of computer assisted radiology and surgery*, 14:1341–1352, 2019.

[164] Abdullah Nazib, Clinton Fookes, and Dimitri Perrin. A comparative analysis of registration tools: Traditional vs deep learning approach on high resolution tissue cleared data. *arXiv preprint arXiv:1810.08315*, 2018.

[165] Kenichi Oishi, Susumu Mori, Pamela K Donohue, Thomas Ernst, Lynn Anderson, Steven Buchthal, Andreia Faria, Hangyi Jiang, Xin Li, Michael I Miller, et al. Multi-contrast human neonatal brain

atlas: application to normal neonate development analysis. *Neuroimage*, 56(1):8–20, 2011.

[166] Toshiyuki Okada, Ryuji Shimada, Masatoshi Hori, Masahiko Nakamoto, Yen-Wei Chen, Hironobu Nakamura, and Yoshinobu Sato. Automated segmentation of the liver from 3d ct images using probabilistic atlas and multilevel statistical shape model. *Academic radiology*, 15(11):1390–1403, 2008.

[167] Djeane Debora Onthoni, Ting-Wen Sheng, Prasan Kumar Sahoo, Li-Jen Wang, and Pushpanjali Gupta. Deep learning assisted localization of polycystic kidney on contrast-enhanced ct images. *Diagnostics*, 10(12):1113, 2020.

[168] Şaban Öztürk. Class-driven content-based medical image retrieval using hash codes of deep features. *Biomedical Signal Processing and Control*, 68:102601, 2021.

[169] Dinesh D Patil and Sonal G Deore. Medical image segmentation: a review. *International Journal of Computer Science and Mobile Computing*, 2(1):22–27, 2013.

[170] Nick Pawlowski, Sofia Ira Ktena, Matthew C. H. Lee, Bernhard Kainz, Daniel Rueckert, Ben Glocker, and Martin Rajchl. Dltk: State of the art reference implementations for deep learning on medical images, 2017.

[171] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, et al. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1):1–14, 2021.

[172] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

[173] Emmanuel Pintelas, Ioannis E Livieris, and Panagiotis Pintelas. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. *Algorithms*, 13(1):17, 2020.

[174] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8–20, 2017.

[175] Deepthi Rajashekar, Matthias Wilms, M Ethan MacDonald, Jan Ehrhardt, Pauline Mouches, Richard Frayne, Michael D Hill, and Nils D Forkert. High-resolution t2-flair and non-contrast ct brain atlas of the elderly. *Scientific Data*, 7(1):56, 2020.

[176] Ashwin Raju, Chi-Tung Cheng, Yunakai Huo, Jinzheng Cai, Junzhou Huang, Jing Xiao, Le Lu, Chien-Huang Liao, and Adam P Harrison. Co-heterogeneous and adaptive segmentation from multi-source and multi-phase ct imaging data: A study on pathological liver and lesion segmentation, 2021.

[177] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[178] Bernard Rosner, Robert J Glynn, and Mei-Ling T Lee. The wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, 62(1):185–192, 2006.

[179] HR Roth, L Lu, A Farag, A Sohn, and RM Summers. Spatial aggregation of holistically-nested networks for automated pancreas segmentation in international conference on medical image computing and computer-assisted intervention, 2016.

[180] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pages 556–564. Springer, 2015.

[181] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018.

[182] Holger R Roth, Chen Shen, Hirohisa Oda, Takaaki Sugino, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 417–425. Springer, 2018.

[183] Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.

[184] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[185] László Ruskó, György Bekes, and Márta Fidrich. Automatic segmentation of the liver from multi-and single-phase contrast-enhanced ct images. *Medical Image Analysis*, 13(6):871–882, 2009.

[186] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[187] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.

[188] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 2023.

[189] Elaine H Shen, Caroline C Overly, and Allan R Jones. The allen human brain atlas: comprehensive gene expression mapping of the human brain. *Trends in neurosciences*, 35(12):711–714, 2012.

[190] Feng Shi, Pew-Thian Yap, Guorong Wu, Hongjun Jia, John H Gilmore, Weili Lin, and Dinggang Shen. Infant brain atlases from neonates to 1-and 2-year-olds. *PloS one*, 6(4):e18746, 2011.

[191] Akinobu Shimizu, Rena Ohno, Takaya Ikegami, Hidefumi Kobatake, Shigeru Nawano, and Daniel Smutek. Segmentation of multiple organs in non-contrast 3d abdominal ct images. *International journal of computer assisted radiology and surgery*, 2:135–142, 2007.

[192] Xiu Shu, Yunyun Yang, and Boying Wu. Adaptive segmentation model for liver ct images based on neural network and level set method. *Neurocomputing*, 453:438–452, 2021.

[193] Salim Si-Mohamed, Nicolas Dupuis, Valérie Tatard-Leitman, David Rotzinger, Sara Boccalini, Matthias Dion, Alain Vlassenbroek, Philippe Coulon, Yoad Yagil, Nadav Shapira, et al. Virtual versus true non-contrast dual-energy ct imaging for the diagnosis of aortic intramural hematoma. *European radiology*, 29:6762–6771, 2019.

[194] Tzu-Hsi Song, Victor Sanchez, Hesham EIDaly, and Nasir M Rajpoot. Dual-channel active contour model for megakaryocytic cell segmentation in bone marrow trephine histology images. *IEEE transactions on biomedical engineering*, 64(12):2913–2923, 2017.

[195] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *arXiv preprint arXiv:2006.03829*, 2020.

[196] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[197] Yibo Tang, Yaxiong Chen, and Shengwu Xiong. Deep semantic ranking hashing based on self-attention for medical image retrieval. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4960–4966. IEEE, 2022.

[198] Yucheng Tang, Riqiang Gao, Shizhong Han, Yunqiang Chen, Dashan Gao, Vishwesh Nath, Camilo Bermudez, Michael R Savona, Shunxing Bao, Ilwoo Lyu, et al. Body part regression with self-supervision. *IEEE Transactions on Medical Imaging*, 40(5):1499–1507, 2021.

[199] Yucheng Tang, Riqiang Gao, Ho Hin Lee, Yunqiang Chen, Dashan Gao, Camilo Bermudez, Shunxing Bao, Yuankai Huo, Brent V Savoie, and Bennett A Landman. Phase identification for dynamic ct enhancements with generative adversarial network. *Medical Physics*, 48(3):1276–1285, 2021.

[200] Yucheng Tang, Riqiang Gao, Ho Hin Lee, Shizhong Han, Yunqiang Chen, Dashan Gao, Vishwesh Nath, Camilo Bermudez, Michael R Savona, Richard G Abramson, et al. High-resolution 3d abdominal segmentation with random patch network fusion. *Medical Image Analysis*, 69:101894, 2021.

[201] Yucheng Tang, Riqiang Gao, Ho Hin Lee, Zhoubing Xu, Brent V Savoie, Shunxing Bao, Yuankai Huo, Agnes B Fogo, Raymond Harris, Mark P de Caestecker, et al. Renal cortex, medulla and pelvicaliceal system segmentation on arterial phase ct images with random patch-based networks. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 379–386. SPIE, 2021.

[202] Yucheng Tang, Yuankai Huo, Yunxi Xiong, Hyeonsoo Moon, Albert Assad, Tamara K Moyo, Michael R Savona, Richard Abramson, and Bennett A Landman. Improving splenomegaly segmentation by learning from heterogeneous multi-source labels. In *Medical Imaging 2019: Image Processing*, volume 10949, pages 53–60. SPIE, 2019.

[203] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh

Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[204] Qiaoying Teng, Zhe Liu, Yuqing Song, Kai Han, and Yang Lu. A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, pages 1–21, 2022.

[205] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[206] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[207] Jun-Ichiro Toriwaki, Yasuhito Suenaga, Toshio Negoro, and Teruo Fukumura. Pattern recognition of chest x-ray images. *Computer Graphics and Image Processing*, 2(3-4):252–271, 1973.

[208] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[209] Martin Urschler, Manuel Werlberger, Eva Scheurer, and Horst Bischof. Robust optical flow based deformable registration of thoracic ct images. *Medical Image Analysis for the Clinic: A Grand Challenge*, pages 195–204, 2010.

[210] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 547–556. IEEE, 2018.

[211] Gijs van Tulder and Marleen de Bruijne. Learning cross-modality representations from multi-modal images. *IEEE transactions on medical imaging*, 38(2):638–648, 2018.

[212] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[213] Marinus T Vlaardingerbroek and Jacques A Boer. *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013.

[214] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Machine Learning for Healthcare Conference*, pages 755–769. PMLR, 2021.

[215] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

[216] Jinke Wang, Hongliang Zu, Haoyan Guo, Rongrong Bi, Yuanzhi Cheng, and Shinichi Tamura. Patient-specific probabilistic atlas combining modified distance regularized level set for automatic liver segmentation in ct. *Computer Assisted Surgery*, 24(sup2):20–26, 2019.

[217] Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, et al. The allen mouse brain common coordinate framework: a 3d reference atlas. *Cell*, 181(4):936–953, 2020.

[218] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[219] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021.

[220] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[221] Yan Wang, Jianpeng Zhang, Hengfei Cui, Yanning Zhang, and Yong Xia. View adaptive learning for pancreas segmentation. *Biomedical Signal Processing and Control*, 66:102347, 2021.

[222] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55:88–102, 2019.

[223] Haruo Watanabe, Masayuki Kanematsu, Toshiharu Miyoshi, Satoshi Goshima, Hiroshi Kondo, Noriyuki Moriyama, and Kyongtae T Bae. Improvement of image quality of low radiation dose ab-

dominal ct by increasing contrast enhancement. *American Journal of Roentgenology*, 195(4):986–992, 2010.

[224] Ashley A Weaver, Kathryn L Loftis, Josh C Tan, Stefan M Duma, and Joel D Stitzel. Ct based three-dimensional measurement of orbit and eye anthropometry. *Investigative ophthalmology & visual science*, 51(10):4892–4897, 2010.

[225] Min Wei, Yongjin Zhou, and Mingxi Wan. A fast snake model based on non-linear diffusion for medical image segmentation. *Computerized Medical Imaging and Graphics*, 28(3):109–117, 2004.

[226] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

[227] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[228] Yuxin Wu and Justin Johnson. Rethinking" batch" in batchnorm. *arXiv preprint arXiv:2105.07576*, 2021.

[229] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

[230] Zhaohu Xing, Lequan Yu, Liang Wan, Tong Han, and Lei Zhu. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 140–150. Springer, 2022.

[231] Yan Xu, Yang Li, Mingyuan Liu, Yipei Wang, Maode Lai, and Eric I-Chao Chang. Gland instance segmentation by deep multichannel side supervision. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 496–504. Springer, 2016.

[232] Zhoubing Xu, Christopher P Lee, Mattias P Heinrich, Marc Modat, Daniel Rueckert, Sebastien Ourselin, Richard G Abramson, and Bennett A Landman. Evaluation of six registration methods for the human abdomen on clinically acquired ct. *IEEE Transactions on Biomedical Engineering*, 63(8):1563–1572, 2016.

[233] Yuan Xue, Hui Tang, Zhi Qiao, Guanzhong Gong, Yong Yin, Zhen Qian, Chao Huang, Wei Fan, and Xiaolei Huang. Shape-aware organ segmentation by predicting signed distance maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12565–12572, 2020.

[234] Ke Yan, Le Lu, and Ronald M Summers. Unsupervised body part regression via spatially self-ordering convolutional neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1022–1025. IEEE, 2018.

[235] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018.

[236] Hao-Yu Yang, Junling Yang, Yue Pan, Kunlin Cao, Qi Song, Feng Gao, and Youbing Yin. Learn to be uncertain: Leveraging uncertain labels in chest x-rays with bayesian neural networks. In *CVPR Workshops*, pages 5–8, 2019.

[237] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022.

[238] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5812–5823. Curran Associates, Inc., 2020.

[239] Zhao Yu-Qian, Gui Wei-Hua, Chen Zhen-Cheng, Tang Jing-Tian, and Li Ling-Yun. Medical images edge detection based on mathematical morphology. In *2005 IEEE engineering in medicine and biology 27th annual conference*, pages 6492–6495. IEEE, 2006.

[240] Gabriel Tozatto Zago, Rodrigo Varejao Andreao, Bernadette Dorizzi, and Evandro Ottoni Teatini Salles. Retinal image quality assessment using deep learning. *Computers in biology and medicine*, 103:64–70, 2018.

[241] Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang,

Jingtong Hu, and Yiyu Shi. Positional contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 221–230. Springer, 2021.

[242] Le Zhang, Ali Gooya, Bo Dong, Rui Hua, Steffen E Petersen, Pau Medrano-Gracia, and Alejandro F Frangi. Automated quality assessment of cardiac mr images using convolutional neural networks. In *Simulation and Synthesis in Medical Imaging: First International Workshop, SASHIMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, pages 138–145. Springer, 2016.

[243] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[244] Yuyao Zhang, Feng Shi, Guorong Wu, Li Wang, Pew-Thian Yap, and Dinggang Shen. Consistent spatial-temporal longitudinal atlas construction for developing infant brains. *IEEE transactions on medical imaging*, 35(12):2568–2577, 2016.

[245] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3417–3425, 2022.

[246] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10600–10610, 2019.

[247] Xiangyun Zhao, Raviteja Vemulapalli, Philip Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label-efficient semantic segmentation. *arXiv preprint arXiv:2012.06985*, 2020.

[248] Xiangyun Zhao, Raviteja Vemulapalli, Philip Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label-efficient semantic segmentation, 2021.

[249] Yu-qian Zhao, Zhen Yang, Yan-jin Wang, Fan Zhang, Ling-li Yu, Xiao-bin Wen, et al. Target organ non-rigid registration on abdominal ct images via deep-learning based detection. *Biomedical Signal Processing and Control*, 70:102976, 2021.

[250] Aoxiao Zhong, Xiang Li, Dufan Wu, Hui Ren, Kyungsang Kim, Younggon Kim, Varun Buch, Nir Neumark, Bernardo Bizzo, Won Young Tak, et al. Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in covid-19. *Medical Image Analysis*, 70:101993, 2021.

[251] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[252] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[253] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

[254] Xiangrong Zhou, Takaaki Ito, Ryosuke Takayama, Song Wang, Takeshi Hara, and Hiroshi Fujita. Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting. In *Deep Learning and Data Labeling for Medical Applications*, pages 111–120. Springer, 2016.

[255] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L. Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[256] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K Fishman, and Alan L Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I*, pages 693–701. Springer, 2017.

[257] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.

[258] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

[259] Zhuotun Zhu, Yingda Xia, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In *International conference on medical image computing and computer-assisted intervention*, pages 3–12. Springer, 2019.

[260] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–428. Springer, 2019.

[261] Hasib Zunair and A Ben Hamza. Sharp u-net: depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, 136:104699, 2021.