MACHINE LEARNING-BASED TECHNIQUES FOR MEDICAL IMAGE REGISTRATION AND SEGMENTATION AND A TECHNIQUE FOR PATIENT-CUSTOMIZED PLACEMENT OF COCHLEAR IMPLANT ELECTRODE ARRAYS

By

Jianing Wang

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

Electrical Engineering

August 31, 2021

Nashville, Tennessee

Approved:

Benoit M. Dawant, Ph.D.
Ipek Oguz, Ph.D.
Jack H. Noble, Ph.D.
Robert F. Labadie, M.D., Ph.D.
Yuankai Huo, Ph.D.

This work is dedicated to my husband, who has been a constant source of encouragement during the challenges of graduate school and life. I am truly thankful for having you in my life. This work is also dedicated to my loving parents, who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve.

ACKNOWLEDGMENTS

First, from the bottom of my heart, I would like to say a big thank you to my advisor Dr. Benoit M. Dawant. I am extremely grateful that you took me on as a student and continued to have faith in me over the years. Your support, guidance, and overall insights in the field of medical image processing have made my Ph.D. journey a valuable experience. You have taught me more than I could ever give you credit for here. You have shown me, by your example, what an extraordinary scientist and person should be. I am thankful for the wonderful experiences you arranged for me and for providing me the opportunities to grow professionally. It is my great honor to be your student.

I owe my deepest gratitude to Dr. Jack H. Noble, who led me to the research area of cochlear implants. I still remember you introducing the projects about cochlear implants to me on my first day in the lab. You guided me through my early years, my first research paper would not have been possible without your help.

I would like to express my appreciation to other members of my dissertation committee: Dr. Ipek Oguz, Dr. Robert F. Labadie, and Dr. Yuankai Huo. I appreciate your support during my dissertation research and the general advice you provided about career development.

I would like to thank Dr. Eric R. Tkaczyk for providing me the opportunities to start my research in deep-learning-based image processing methods. Our paper about skin lesion segmentation is my first published work using deep neural networks. Your passion and dedication to research have greatly inspired me to work hard.

I would also like to extend my sincere thanks to my former supervisor Dr. Yulei Jiang from the University of Chicago. I am grateful to you for giving me, a bold young woman who just graduated from college at that time, the opportunity to work in your team. You provided me a great starting point for a career in scientific research. Thank you for your patience, encouragement, and kind support. I would not be where I am today without you.

My past and present colleagues at Vanderbilt University, thank you so much for all of your support throughout this process. Rui Li, William J. Rodriguez, and Michelle Bukowski, you are always willing to help when I am having problems. Dr. Yuan Liu, Dr. Yiyuan Zhao, Dr. Dongqing Zhang, Dr. Ahmet Cakir, Dr. Xiaochen Yang, Dr. Srijata Chakravorti, Dr. Xiaoqi Liu, Dr. Andrew J. McNeil, Dr. Inga Saknīte, Yubo Fan, Tahsin Reasat, Ziteng Liu, Can Cui, Han Liu, Dingjie Su, Fuyao Chen, Francois-Xavier Carton, and Yujie Chi: Thank you so much for your friendship and help, working with you is a very happy and unforgettable experience to me.

Special thanks to financial support from NIH grants R01-NS095291, R01DC008408, R01DC014037, and R01DC014462. Special thanks also to the Advanced Computing Center for Research and Education (ACCRE) of Vanderbilt University (Nashville, TN) for providing computing resources. I am grateful for the support of the NVIDIA Corporation for providing me with a GPU used to support my research work.

Finally, I wish to thank my parents and my husband. Thank you for your love, support, and encouragement, without you I would never have enjoyed so many opportunities.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

Abbreviation	Description
A1	A Cochlear (Sydney, Australia) Contour Advance array
A2	An Advanced Bionics (Valencia, CA, USA) Mid-Scala array
ABA	Adaptive Bases Algorithm
ACCRE	Advanced Computing Center for Research and Education
ART	Automatic Registration Tools
AutoConfig	An automatic method to generate the configuration of the electrode array
BendE	Bending Energy
cCT	Conventional Computed Tomography
cGANs	Conditional Generative Adversarial Networks
CNC	Consonant-Nucleus-Consonant
CSF	CerebroSpinal Fluid
CT	Computed Tomography
DD	Diffeomorphic Demons
DDF	Dense Deformation Field
DSC	Dice Similarity Coefficients
DVF	Distance-Vs.-Frequency
F3D	Fast Free Form Deformation
FRE	Fiducial Registration Error
GANs	Generative Adversarial Nets
GM	Gray Matter
IB ears	Ears of the cochlear implant recipients who have been implanted bilaterally
IGCIP	Image-Guided Cochlear Implant Programming
IU ears	Ears of the cochlear implant recipients who have been implanted unilaterally
lCT	Low-dose Computed Tomography
MAX	MAXimum
MEA	MEAn
MED	MEDian
MIN	MINimum
MR images	Magnetic Resonance images

MSPDice	Multiscale Soft Probabilistic Dice
MSSIM	Mean Structure SIMilarity
NCC	Normalized Cross-Correlation
P2PE	Point-to-Point Error
Post-CT	Post-implantation Computed Tomography
Pre-CT	Pre-implantation Computed Tomography
RANSAC	RANdom SAmple Consensus
RF	Random Forest
SegPost-BL	A method for segmenting intracochlear anatomy in post-implantation CT images of cochlear implants recipients who have been implanted bilaterally
SegPost-UL	A method for segmenting intracochlear anatomy in post-implantation CT images of cochlear implants recipients who have been implanted unilaterally
SegPre-ASM	A method for segmenting intracochlear anatomy in pre-implantation CT images of cochlear implants recipients
SSIM	Structural SIMilarity
STD	STandard Deviation
SyN	Symmetric Normalization
TPS	Thin Plate Spline
WL1	Weighted L1
WM	White Matter

CHAPTER I

Introduction

1.1 Preregistration Initialization of Non-Rigid Registration Algorithms

Image registration is the process of transforming different sets of images into one coordinate system. Medical image registration aims to find a spatial transformation that best aligns the anatomical structures in two or more images, and it is at the core of many applications. For example, deep-brain stimulation surgery involves placing electrodes within specific deep-brain target nuclei, that is, an atlas-based prediction of the optimal placement point can be made by registering a patient image to an atlas image and then by projecting the optimal placement point from the atlas to the patient using the transformation from the registration algorithm (D’Haese et al., 2005). Image registration methodology can be roughly categorized into affine and deformable. Affine transformation allows translation, rotation, stretching, and shearing or skewing, it moves objects around in space and preserves collinearity (i.e., all points lying on a line initially still lie on a line after transformation) and ratios of distances (e.g., the midpoint of a line segment remains the midpoint after transformation). Deformable transformation allows a non-uniform mapping between images, the term deformable is often used to denote the fact that the observed signals are associated through a non-linear dense transformation, or a spatially varying deformation model (Sotiras et al., 2013). Deformable registration is widely used for the accurate registration of objects with high local variations. Most commonly used deformable registration methods are dependent on a good preregistration initialization (Maintz & Viergever, 1998; Han et al., 2015), i.e., the two images have to be accurately aligned before applying these methods. The initialization can be performed by manually aligning the images, by using intensity-based affine registration algorithms, or by localizing homologous landmarks and calculating a point-based transformation between the images. When landmarks are used, the selection of these landmarks is important. The process of defining a landmark depends on the profound geometrical understanding of the targeted anatomy as well as the modality-specific appearance. While the manual selection of landmarks is possible for small landmark sets, it becomes impractical for larger sets that are required to, for instance, register non-rigidly 3D images. Manual selection of landmarks can also be prone to intra- and inter-observer variations.

These facts led us to explore ways to automatically find a set of robust landmarks to initialize deformable transformations. In this dissertation, we will introduce a machine learning-based method to automatically find a set of robust landmarks to initialize deformable transformations. Specifically, we focus on the T1-weighted MR images of the head, which are commonly used in image-guided deep-brain stimulation (Kahn et al., 2012; D'Haese et al., 2012).

1.2 Cochlear Implants and Image-guided Cochlear Implant Programming

Image registration techniques can also be used for medical image segmentation. For example, atlas-based segmentation methods use image registration to warp the segmentation masks from the labeled atlas to the images to be segmented. Methods for image registration and segmentation have been successfully developed over the past decades, however, the accuracy of these methods can be prone to extreme abnormalities or strong artifacts in the images. Deep learning has achieved tremendous success over the last few years in the field of medical image analysis. Inspired by these, we explore using deep learning techniques to register and segment medical images with defective areas. Specifically, we focus on the segmentation of the CT images of the cochlear implant recipients, which plays a key role in Image-Guided Cochlear Implant Programming (IGCIP) (Noble et al., 2013).

The cochlea is a spiral-shaped cavity that is part of the inner ear. In natural hearing, the cochlea transforms mechanical sound waves into electrical stimulation in the auditory nerves, thus creating a sense of hearing. Figure 1.1 (J. Wang et al., 2019) is an illustration of the intracochlear anatomy. The cochlea contains two main cavities: the scala tympani and the scala vestibuli. The modiolus is a porous bone around which the cochlea is wrapped that hosts the auditory nerves.

Cochlear implants are surgically implanted neural prosthetic devices that are used to treat severe-to-profound hearing loss (National Institute on Deafness and Other Communication Disorders, 2016). Cochlear implants bypass the normal hearing process by replacing it with direct stimulation of neural pathways using an implanted electrode array. Cochlear implants are programmed postoperatively by audiologists for each patient based on their auditory perceptions in response to electrical stimuli. A program or "map", is created for the cochlear implant recipient by setting several parameters to ensure that the electrical pattern generated by the device in response to sound yields optimal speech intelligibility (Vaerenberg et al., 2014). For maximum benefit, recipients of cochlear implants are seen for regularly scheduled intervals to reprogram their cochlear implant throughout their lifetime (Mertes & Chinnici, 2006).

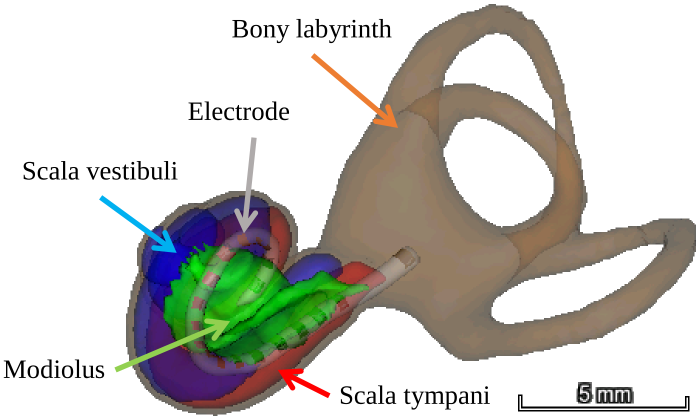


Figure 1.1: An illustration of intracochlear anatomical structures and cochlear implant electrodes.

The spiral ganglion is a group of neuron cell bodies in the modiolus, which are tonotopically ordered by decreasing characteristic frequency from 20000 Hz near the cochlear base to 20 Hz near the cochlear apex. A spiral ganglion cell is stimulated if the incoming sound contains the frequency associated with it (Greenwood, 1990). This stimulation generates hearing impulses that are sent to the brain to induce a sense of hearing. The number of electrodes on the implant electrode array ranges from 12 to 22, depending on the manufacturer, and each electrode activates the spiral ganglion cells that resonate with the sound in a certain band of frequencies (Figure 1.2) (Noble et al., 2014).

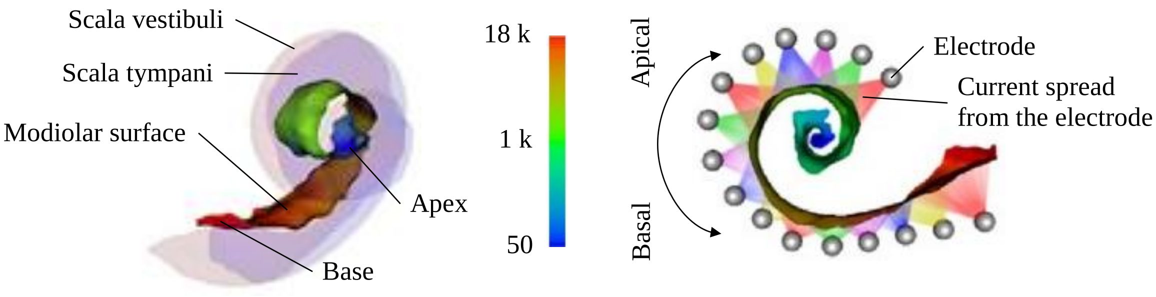


Figure 1.2: An illustration of the modiolar surface and electrodes of the cochlear implant (Noble et al., 2014).

Recent studies have shown those hearing outcomes with cochlear implants are correlated with the spatial relationship between the electrode array and the intracochlear anatomy (Aschendorff et al., 2005; Finley et al., 2008; Holden et al., 2013; Nordfalk et al., 2014; Wanna et al., 2014, 2015). Electrode interaction, which happens when neighboring electrodes are stimulating the same neural area, is an effect that negatively impacts hearing outcomes. Once detected, it can be reduced by deactivating a subset of electrodes. Our IGCIP system has been developed for this purpose. The IGCIP system consists of two stages: (1) before surgery, a Pre-implantation CT (Pre-CT) image of the patient is obtained. The intracochlear anatomy in the Pre-CT image is segmented by using an active shape model-based method (Noble et al., 2011), which is referred to as SegPre-ASM in this dissertation. (2) After surgery, a Post-implantation CT (Post-CT) image is obtained and the electrodes are localized in the Post-CT image (Zhao et al., 2014, 2017, 2018, 2019; Noble & Dawant, 2015). By doing a rigid registration between the Pre-CT image and the Post-CT image, the intracochlear anatomy in the Post-CT image can be segmented by projecting the intracochlear anatomy from the Pre-CT image to the Post-CT image. This permits the determination of the position of the implanted electrodes relative to the intracochlear anatomy. In turn, this assists in selecting the electrode configuration. We have shown that IGCIP can significantly improve hearing outcomes in both adults and children (Noble et al., 2014, 2016).

Our IGCIP system can also generate the configuration of the electrode array of the cochlear implant automatically by using the knowledge of the relative positions of each electrode to the spiral ganglion and their neighboring electrodes. Given the segmentation of the intracochlear anatomy and the localization results of the electrodes, the final output of our IGCIP techniques is the configuration of the electrode array, which is a set of “on” and “off” states of the electrodes. Our automatic method (Zhao et al., 2016), which is referred to as AutoConfig in this dissertation, for generating the configurations of the electrode arrays selects the state of each electrode by optimizing a cost function that is based on the electrode Distance-Vs.-Frequency (DVF) curves. A DVF curve is a 2D plot that captures the patient-specific spatial relationship between the electrodes and the spiral ganglion cells (Noble et al., 2013). Figure 1.3 is an illustration of the DVF curves for a 12-electrode array. Each DVF curve corresponds to one electrode. For each point on a DVF curve, the horizontal value represents the positions along the length of the modiolus in terms of the characteristic frequencies of the spiral ganglion cells, and the vertical value indicates the distance from the electrode to the modiolar surface. The blue DVF curves denote the corresponding electrodes are turned on, and the red ones denote the electrodes are turned off.

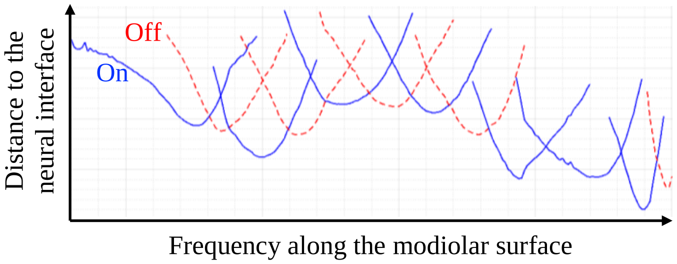


Figure 1.3: An illustration of the DVF curves and the configuration for a 12-electrode array.

1.3 Segmentation of the Intracochlear Anatomy in Post-Implantation CT Images

However, because segmenting the intracochlear anatomy in the Post-CT images is challenging due to the strong artifacts produced by the metallic implant electrodes, our tools for assisting cochlear implant activation and programming do not extend to recipients for whom a Pre-CT image is unavailable, which is the case for long-term recipients who were not scanned before surgery, for recipients for whom images cannot be retrieved, or for recipients implanted at institutions that use pre-operative MR images instead of CT images. To overcome this issue, Reda et al. (Reda, McRackan, et al., 2014; Reda, Noble, et al., 2014) have proposed two methods to segment the intracochlear anatomy in Post-CT images. The first method, which we refer to as SegPost-UL, was developed for segmenting intracochlear anatomy in Post-CT images of recipients who have been implanted unilaterally (Reda, McRackan, et al., 2014). SegPost-UL relies on the intra-subject symmetry in cochlear anatomy across the ears. It first segments the intracochlear anatomy of the contralateral normal ear and then maps the segmented structures to the implanted ear. The second method, which we refer to as SegPost-BL, was developed for segmenting intracochlear anatomy structures in Post-CT images of recipients who have been implanted bilaterally (Reda, Noble, et al., 2014). SegPost-BL first identifies the labyrinth in the Post-CT image by mapping a labyrinth surface that is selected from a library of labyrinth surfaces and then uses the localized labyrinth in the image as a landmark to segment the scala tympani, scala vestibuli, and modiolus with a standard shape model-based segmentation method. But, while using these methods it was observed that they could at times lead to results that lacked accuracy. Thus, it is necessary to improve the accuracy of the segmentation of the intracochlear anatomy when only Post-CT images are available. We will introduce three deep learning-based methods to segment the intracochlear anatomy in post-implantation CT images in this dissertation.

1.4 Intracochlear Placement of the Electrode Arrays

Cochlear implants have produced remarkable results for the majority of recipients with average postoperative word and sentence recognition approximating 60% and 70% correct, respectively, for unilaterally implanted recipients and 70% and 80% correct for bilateral recipients (Litovsky et al., 2006; Buss et al., 2008; Gifford et al., 2008, 2014; Dorman et al., 2011). However, patient outcomes remain highly variable, and there is a substantial fraction of individuals who experience poor speech recognition outcomes (Holden et al., 2013). Depending on the resting state shape of the array, current commercially available electrode arrays can be broadly divided into straight and pre-curved arrays. Straight arrays position the electrodes along the outer "lateral" wall of the cochlea, whereas pre-curved arrays are advanced into the cochlea off of a straightening stylet, or out of a straightening sheath, and coil to attempt to match the shape of the inner "modiolar" wall of the cochlea. Currently, both straight and pre-curved arrays are inserted using "soft" surgical techniques in which the array is threaded at a tangential angle into the scala tympani through either the existing round window membrane or a separate cochleostomy site while attempting to inflict as little trauma as humanly possible on the soft tissue contained within the cochlea (Roland & Wright, 2006). A system of markers on the array lead proximal to the electrodes is used to visually indicate when the generically recommended overall insertion depth of the electrode array has been realized once the markers reach the cochlear entry site (Banalagay et al., 2020). This is a one-size-fits-all approach as the patient-specific cochlear size and shape are not considered. It may have merit if the variations in anatomy are clinically insignificant but it is well-known that considerable variation in cochlear anatomy exists across individuals (Pelosi et al., 2013). As mentioned earlier, the intracochlear placement of the electrodes can affect the outcomes of cochlear implants. However, as the array inside the scala tympani is invisible to the surgeon, the intracochlear placement of the electrodes is generally unknown during and after the surgery and thus the intracochlear positioning of the electrodes has not been evaluated on a large scale yet.

Our recently developed automated image analysis techniques (Noble et al., 2011; Zhao et al., 2014; Noble & Dawant, 2015) permit using post-implantation CT images to detect the intracochlear position of electrodes relative to intracochlear anatomy. These make it feasible, for the first time, to evaluate electrode positioning on a large scale. In this dissertation, we will use these techniques to retrospectively evaluate the position of pre-curved electrode arrays on a large cohort. We will also explore a technique to improve the placement of the electrode arrays by preoperatively planning.

1.5 Goals and Contributions of this Dissertation

The goals of this dissertation are to (a) develop a method to automatically find a set of robust landmarks in T1-weighted MR images of the head to initialize non-rigid transformations; (b) develop methods to segment the intracochlear anatomy in Post-CT images accurately, and (c) discover the rate at which perimodiolar placement is successfully achieved and to evaluate a new technique we propose to preoperatively plan patient-customized electrode array insertion depths to improve perimodiolar placement at the time of surgery.

The specific contributions of this dissertation are summarized below:

Chapter II presents a learning-based method to automatically find a set of robust landmarks in T1-weighted MR images of the head to initialize non-rigid transformations. Our method involves two steps. First, landmarks that can be reliably localized in the images are identified using a random forest-based method (Breiman, 2001). The subset of landmarks that leads to good initialization transformations, which are computed with a thin plate spline-based method (Rohr et al., 2001), are then selected using a random sample consensus algorithm (Fischler & Bolles, 1981). To show the value of our registration initialization technique, we compare the final registration results obtained with 5 well-established deformable registration algorithms when either an affine transformation or the proposed method is used for preregistration initialization. We show that higher registration accuracy is achieved in the latter case for all 5 registration algorithms. The technique we propose is generic and could be used to initialize non-rigid registration algorithms for other applications.

Chapter III presents a two-step method to segment the intracochlear anatomy in the Post-CT images: (1) generate a synthetic Pre-CT image from a Post-CT image with conditional generative adversarial networks (Goodfellow et al., 2020; Mirza & Osindero, 2014) trained for this purpose and (2) apply SegPre-ASM to the synthetic image. We have shown that this method substantially and significantly improves the results obtained with methods designed to operate directly on Post-CT images, and this method is the current state of the art for segmenting the intracochlear anatomy in Post-CT images. This method has been integrated into an interactive software package that has been deployed to the clinic and is in routine use at Vanderbilt University Medical Center.

Chapter IV evaluates the effect of the method presented in **Chapter III** on the final output of our IGCIP system by assessing the configurations of the electrode arrays obtained with the synthetic Pre-CT images. Results show that around 85% of our configurations generated using the synthetic images can be used for the programming of the cochlear implants, and these configurations are likely to lead to hearing outcomes that are comparable to those achieved using the best possible configurations. Our method provides a good solution for patients for whom the Pre-CT images are unavailable.

Chapter V presents a method, which is an extension of the method proposed in **Chapter III**, to segment the intracochlear anatomy in the Post-CT images. The method uses a multi-task network, which consists of a shared feature extractor, an image synthesis branch, and an image segmentation branch, to remove metal artifacts and generate segmentation masks of the intracochlear anatomy in the Post-CT images simultaneously. Compared to the two-step method proposed in **Chapter III**, this method provides an approach to segment the Post-CT images in one step.

The outputs of SegPre-ASM, SegPost-UL, and SegPost-BL are surface meshes of the scala tympani, scala vestibuli, and modiolus that have a predefined number of vertices. Importantly, each vertex corresponds to a specific anatomical location on the surface of the structures and the meshes are encoded with the information needed for the programming of the implant. However, such point-to-point correspondence is not preserved by the method that we have developed in **Chapter V**, because the method generates pixel-wise segmentation masks of the intracochlear anatomical structures. Our group has developed methods that are based on the marching cubes algorithm (D. Zhang et al., 2019; Fan et al., 2020) which can generate surface meshes from the masks, however, they could introduce additional errors. **Chapter VI** presents an atlas-based method to segment the intracochlear anatomy in the Post-CT images. This method preserves the point-to-point correspondence between the meshes in the atlas and the segmented volumes. First, we use a neural network to generate a dense transformation field to register an atlas image, in which the intracochlear anatomy has been segmented, to the Post-CT image, and then the segmentation of the intracochlear anatomy in the Post-CT image can be obtained by warping the segmentation of the intracochlear anatomy in the atlas image to the Post-CT image. We show that this method produces results that are comparable to the current state of the art. The method also produces results in a fraction of the time needed by the current state of the art, which is of importance for clinical deployment and end-user acceptance.

Chapter VII retrospectively evaluates the position of pre-curved electrode arrays on a large cohort. Further, we propose an image-guided approach to select a patient-customized electrode insertion depth based upon the final resting shape of the electrode arrays and the patient's anatomy as assessed on his or her Pre-CT images. Results show that pre-curved electrode arrays tend to be over inserted leading to displacement away from the modiolus, and better perimodiolar positioning can be achieved using patient-customized insertion depths.

Chapter VIII provides the summary of the work and discusses possible future work.

CHAPTER II

Automatic Selection of Landmarks in T1-weighted Head MR Images with Regression Forests for Image Registration Initialization

Jianing Wang, Yuan Liu, Jack H. Noble, and Benoit M. Dawant

Department of Electrical and Computer Engineering

Vanderbilt University

Nashville, TN, 37232, USA

This chapter is adapted from “Automatic selection of landmarks in T1-weighted head MRI with regression forests for image registration initialization” published in Journal of Medical Imaging, 2017, 4(4): 044005.

The published paper can be found at https://doi.org/10.1117/1.JMI.4.4.044005

It has been reproduced with the permission of the publisher.

Abstract

Medical image registration establishes a correspondence between images of biological structures, and it is at the core of many applications. Commonly used deformable image registration methods are dependent on a good preregistration initialization. In this chapter, we develop a learning-based method to automatically find a set of robust landmarks in 3D MR images of the head. These landmarks are then used to compute a thin plate spline-based initialization transformation. The process involves two steps: (1) identifying a set of landmarks that can be reliably localized in the images and (2) selecting among them the subset that leads to a good initial transformation. To validate our method, we use it to initialize 5 well-established deformable registration algorithms that are subsequently used to register an atlas image to MR images of the head. We compare our proposed initialization method to a standard approach that involves estimating an affine transformation with an intensity-based approach. Results show that for all 5 registration algorithms the final registration results are statistically better when they are initialized with the method we propose than when the standard approach is used. The technique we propose is generic and could be used to initialize non-rigid registration algorithms for other applications.

2.1 Introduction

Medical image registration establishes a correspondence between images of biological structures and it is at the core of many applications. Most deformable registration methods that are commonly used are dependent on a good preregistration initialization (Maintz & Viergever, 1998; Han et al., 2015). The initialization can be performed by manually aligning the images, by localizing homologous landmarks and calculating a point-based transformation between the images, or with intensity-based affine registration techniques.

When landmarks are used, the selection of these landmarks is important. Good landmarks should cover the entire biological structure and should be easy to localize unequivocally, i.e., they should have distinct and salient features. While the manual selection of landmarks is possible for small landmark sets, it becomes impractical for larger sets that are required to, for instance, register non-rigidly 3D image volumes. In this chapter, we propose a learning-based method to automatically find a set of robust landmarks in 3D MR images of the head to initialize non-rigid transformations. Our methods involve two steps. First, landmarks that can be reliably localized in the images are identified using a method based on Random Forest (RF) (Breiman, 2001). The subset of landmarks that leads to good initialization transformations, which are computed with a method based on Thin Plate Spline (TPS) (Rohr et al., 2001), are then selected using a RANdom SAmple Consensus (RANSAC) algorithm (Fischler & Bolles, 1981).

To show the value of our registration initialization technique, we compare the final registration results obtained with 5 well-established deformable registration algorithms, i.e., (1) Adaptive Bases Algorithm (ABA) (Rohde et al., 2003), (2) Automatic Registration Tools (ART) (Ardekani et al., 2005), (3) Diffeomorphic Demons (DD) (Vercauteren et al., 2009), (4) Fast Free Form Deformation (F3D) (Modat et al., 2010), and (5) Symmetric Normalization (SyN) (Avants et al., 2008), when either an affine transformation or the proposed method is used for preregistration initialization. We show that a higher registration accuracy is achieved in the latter case.

2.2 Methods

2.2.1 Overview

Our dataset contains images of 201 individuals from the data repository we have created over a decade for deep brain stimulation surgeries (D'Haese et al., 2012). All these images are T1-weighted MR images with approximately 256×256×170 1 mm^3 isotropic voxels. All these images have been acquired clinically with the subjects in roughly the same position but without special care being taken to position them. The images are randomly partitioned into a first training dataset of 100 images that is used to train RF models that are used to localize a set of landmarks, a second training dataset of 80 images that is used to identify among the set of landmarks which ones are the most robust, and a testing dataset of 20 images that is used to validate our technique and one atlas image.

Our technique includes 4 training steps: (1) the generation of the candidate landmark set using the atlas image, (2) the creation of a series of RF models that are each trained to localize one landmark, (3) the localization of the candidate landmarks in the second training dataset, and (4) the selection of the most reliable landmarks using a RANSAC algorithm and the second training set. In the testing phase, the most reliable landmarks are localized in unknown images and they are used to compute a smoothing TPS transformation that registers the atlas image to each of the testing images. Further refinement of the initial registration is performed with the 5 deformable registration algorithms. Differences between these algorithms are briefly summarized in Table 2.1 (Liu et al., 2014). Each registration algorithm requires values for a set of parameters and we use the method presented by Liu (Liu et al., 2014) to select them.

Algorithm	Deformation Model	Similarity Measure	Regularization
ABA	Radial basis functions	NMI of the whole brain	Transformation symmetry Jacobian threshold
ART	Homeomorphism	NCC of the whole brain	Gaussian smoothing
DD	Diffeomorphic optical flow	SSD of the whole brain	Gaussian smoothing
F3D	Cubic B-splines	NMI of the whole brain	Bending energy
SyN	Symmetric diffeomorphism	CC of the whole brain	Gaussian smoothing Transformation symmetry

Table 2.1: Comparison of the 5 deformable registration algorithms. (NMI – Normalized Mutual Information, SSD – the Sum of Square Differences, CC – Cross-Correlation, NCC – Normalized Cross-Correlation).

2.2.2 The Generation of the Candidate Landmark Set

The brain in the atlas image is extracted with the FSL Brain Extraction Tool (Smith, 2002). 5000 candidate landmarks are randomly placed inside the brain region (Figure 2.1). To find the position of the candidate landmarks on the 180 training images, the atlas image is first registered to each of the training images using a sequence of intensity-based rigid and non-rigid registration steps (Rohde et al., 2003). The accuracy of the registration is visually assessed and each of the landmarks is projected from the atlas image to each of the training images. These projected landmarks are considered to be the ground truth position of the landmarks in the training images, and we denote the true position of landmark L_i in image I_j as $T_{i,j}$.

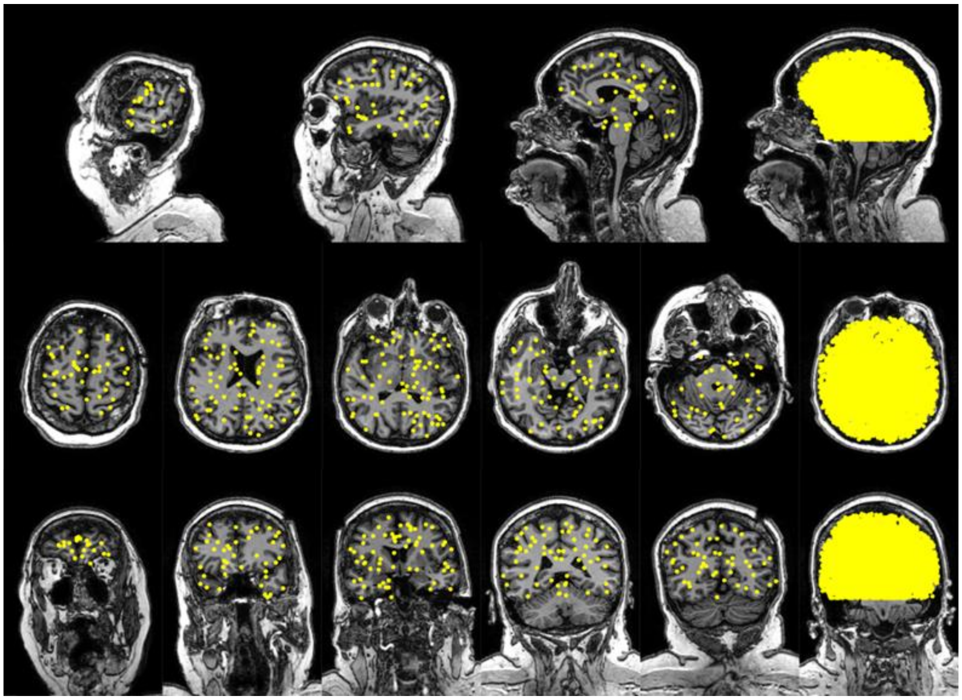


Figure 2.1: Candidate landmarks shown on selected slices of the atlas image in the sagittal (row 1), axial (row 2), and coronal (row 3) directions. The last figure of each row is created by projecting all landmarks on the same slice to show the region of the head covered by the landmarks.

2.2.3 The Creation of the RF Models

To reduce computation cost, we downsample the training images in the first training dataset by a factor of 4. An RF model is trained to localize each landmark in the downsampled images using the methods presented by Glocker (Glocker et al., 2012). Briefly, given a point in a training image, a set of multi-scale long-range textural features are extracted and associated with a probability that this point is at the position of the landmark (Pauly et al., 2011). The RF model is trained to learn the relationship between the features and the probability value. Because one RF model is trained for each landmark, 5000 RF models are trained.

2.2.4 The Localization of the Candidate Landmarks

The images in the second training dataset are first downsampled by a factor of 4. The RF models trained in the previous step are then applied to localize the candidate landmarks in these downsampled images. Given an unknown image and one RF model that is trained to localize a specific landmark L_i, the output of the RF model is a 3D probability map of the same size as the input image (Figure 2.2). In this map, a high value indicates a high probability that the point is the landmark of interest. The local maxima of the 3D probability map are thus the potential positions of the landmark in the image. Because the landmark should be located inside the brain region, the local maxima that are close to the border of the 3D probability map are likely to be false positives. Such local maxima are discarded, and if no local maximum is left, then we consider that the landmark is not localizable in the input image.

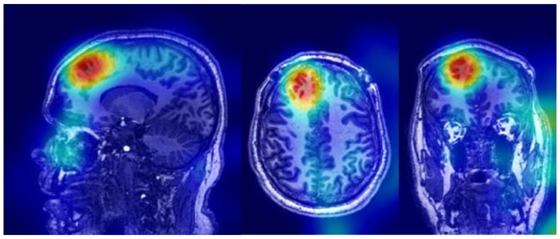


Figure 2.2: Probability map shown on top of a training image in the sagittal (left), axial (middle), and coronal (right) directions. The reddish color indicates higher probability values and the blueish color indicates lower probability values.

We trim the candidate landmark set by removing landmarks that are not localizable in all of the 80 images in the second training dataset. Thus, each of the remaining candidate landmarks has at least one and possibly several corresponding potential landmark points in each training image. Possible reasons for multiple localization include the lack of salient features around a point or anatomic variations. In this study, we select only one point from the multiple potential landmark points. To do so, we define n as the number of images among the 80 images in the second training set in which a landmark has only one corresponding landmark. If $n < 30$ we consider this landmark to be hard to localize unequivocally and we remove it from the candidate landmark set. If $n \geq 30$, then the mean of the landmarks in these n images is calculated, and for each of the 80 –n images, the point that is closest to the mean is selected as the landmark in this image. Because the landmark is localized in the downsampled image, we calculate a coarse landmark position in the full resolution image by upsampling the coordinates of the landmark, and we denote the coarse position of landmark L_i in the full resolution image I_j as $R_{i,j}$. If the distance between $R_{i,j}$ and the true landmark $T_{i,j}$ is greater than 16mm in the x-, y-, or z-direction, we consider that the RF model cannot easily localize L_i, and L_i is discarded to further trim the candidate landmark set. The remaining 1802 landmarks are kept as the candidate landmark set.

The mean of the coarse positions of landmark L_i is calculated as:

$$\bar{R}_i = \frac{1}{80}\sum_{j=1}^{80} R_{i,j}, \tag{2.1}$$

The maximum Euclidean distance between a landmark in the atlas image and the corresponding landmark that is localized by the RF model in an image in the second training dataset is calculated as:

$$MaxDist = \max_i\left(\max_j\left(\left\|L_i - R_{i,j}\right\|\right)\right), \tag{2.2}$$

where i = {1, 2, …, 1802} and j = {1, 2, …, 80}. In our experiments, the value of $MaxDist$ is 70mm, and we use this value to find possible RF model localization errors in the testing phase.

2.2.5 Selection of the Most Reliable Landmarks

We use a RANSAC algorithm to select the most reliable landmarks (denoted as *robust-set*) from the 1802 candidate landmarks (denoted as *whole-set*). The *robust-set* is empty at the beginning. Our algorithm works as follows (Figure 2.3):

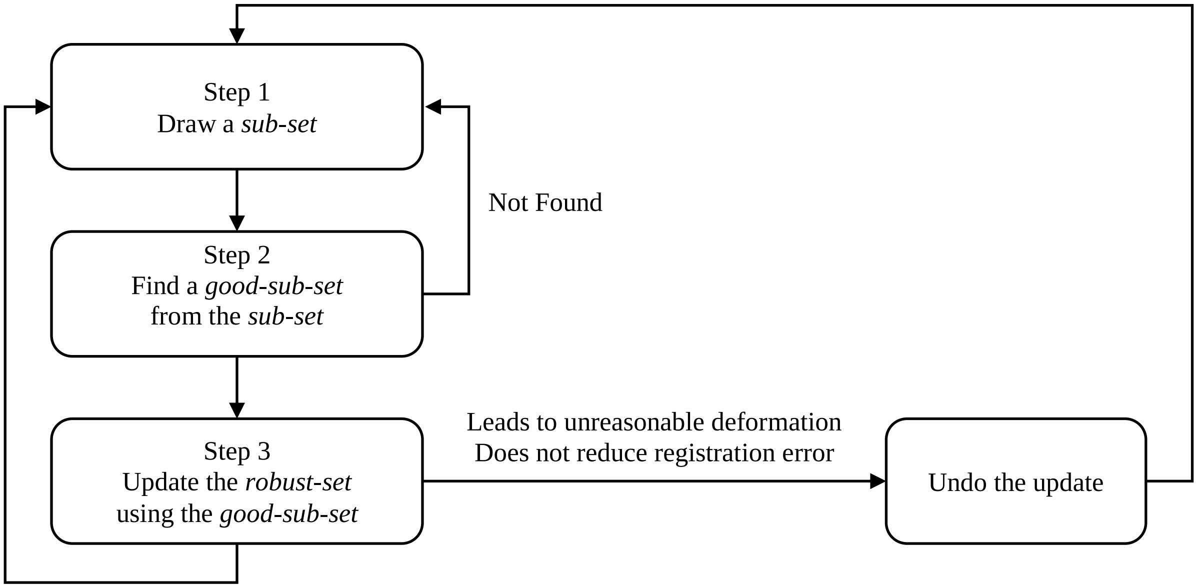


Figure 2.3: The flowchart of our RANSAC algorithm.

Step 1: Randomly draw a *sub-set* of 18 landmarks (1% of the *whole-set*) from the *whole-set*.

Step 2: Register the atlas image to each of the 80 images in the second training dataset by calculating the TPS transformations from the atlas image to the training images with the *sub-set* as control points and project all the landmarks in the *whole-set* from the atlas images to each of the 80 training images with these transformations. Here we have used a fixed smoothing parameter value of 0.5 for calculating the TPS transformations. The mean and standard deviation of the registration error for each landmark in the *whole-set* are calculated as:

$$\bar{\epsilon}_i = \frac{1}{80}\sum_{j=1}^{80} \epsilon_{i,j}, \tag{2.3}$$

$$\sigma_i = \sqrt{\frac{1}{80}\sum_{j=1}^{80}\left(\epsilon_{i,j} - \bar{\epsilon}_i\right)^2}, \tag{2.4}$$

in which the registration error $\epsilon_{i,j}$ of the landmark L_i in image I_j is the Euclidean distance between the true landmark $T_{i,j}$ and the point $P_{i,j}$ obtained by projecting the landmark L_i from the atlas to the image I_i via the TPS transformation. All L_i with $\bar{\epsilon}_i$ < 8mm and σ_i < 4mm are considered to be inliers. If the number of inliers is greater than 1261 (70% of the *whole-set*), it suggests that the transformation computed with the landmarks in the current *sub-set* is reasonable.

Landmarks in the current *sub-set* that are also inliers constitute a *good-sub-set*. If there are less than 1261 inliers, then the current *sub-set* is discarded and we go back to Step1 to draw a new *sub-set*.

Step 3: Update the *robust-set* by adding to it the landmarks in the *good-sub-set* that are not yet in the *robust-set*.

Step 4: Check the quality of the *robust-set*. To do so, we place uniform 3D grids in the atlas image, and the grid coordinates are used as the *check-set* (12168 points in total). To assess whether or not the TPS transformations calculated with the *robust-set* as control points lead to unreasonable deformations, the determinant of the Jacobian matrix of the TPS transformation at each point in the *check-set* is calculated. This value measures the volume change of a voxel after the TPS transformation, e.g., a value of 0.5 indicates that a unit volume contracts down to half of its original volume, and a value of 2 indicates that a unit volume expands to twice its original volume. Taking into account the fact that the heads in our 201 images are of a similar size and position but that large anatomical variation may exist among individuals (especially in the ventricles), we use a very loose decision criterion to determine that a transformation is reasonable, and we consider determinant values between 0.2 and 2.2 to be acceptable. A TPS transformation that leads to valid deformations at more than 99.5% of the *check-set* points is considered to be acceptable. If any TPS transformation that registers the atlas image to one of the training images is invalid, then we undo the update of the *robust-set* and go back to Step 1; otherwise, we check if the TPS transformations reduce the registration error. To do so we calculate the sum of the mean and standard registration error as:

$$s = \bar{\epsilon} + \sigma, \tag{2.5}$$

$$\bar{\epsilon} = \frac{1}{1802\times 80}\sum_{i=1}^{1802}\sum_{j=1}^{80}\epsilon_{i,j}, \tag{2.6}$$

$$\sigma = \sqrt{\frac{1}{1802\times 80}\sum_{i=1}^{1802}\sum_{j=1}^{80}\left(\epsilon_{i,j} - \bar{\epsilon}\right)^2}, \tag{2.7}$$

if the value of s decreases, then we keep the update; otherwise we undo the update and go back to Step 1.

We repeat Step 1 to Step 3 until the value of s converges. In our experiments, which have been repeated with various seed points for the random number generator used to draw the *sub-set*, the algorithm converges with about 500 landmarks in the *robust-set*.

Empirically, we found that too many landmarks can affect the results, i.e., too many degrees of freedom may lead to unrealistic transformations when registering the atlas image to some unknown new images. We address this by reducing the size of the *robust-set*. To do so we subdivide the head into a series of 3D boxes (in the experiment presented here we subdivide an image into 4×5×5 boxes) that cover the whole brain in the atlas image. In each of these boxes, we select the landmark that has the smallest sum of the normalized $\bar{\epsilon}_i$ and the normalized σ_i (Equations 2.3, 2.4) if the box contains a landmark. In the end, the trimmed *robust-set* contains 41 points that provide good coverage of the brain (Figure 2.4).

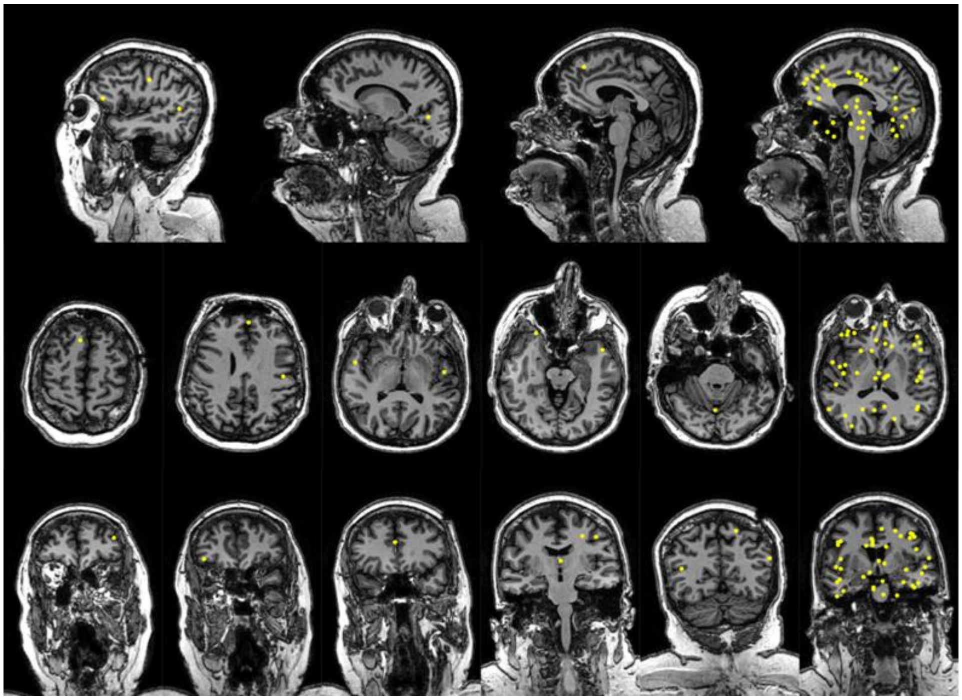


Figure 2.4: The robust-set shown on selected slices of the atlas image in the sagittal (row 1), axial (row 2), and coronal (row 3) directions. The last figure of each row is generated by projecting all landmarks on the same slice to show the region of the head covered by the landmarks.

2.2.6 Method Validation

In the testing phase, first, we use the RF models that are created in Section 2.2.3 to localize the *robust-set* in the 20 testing images. As is done in Section 2.2.4, the images in the testing dataset are downsampled by a factor of 4. Given a downsampled testing image and an RF model that is trained to localize a specific landmark, the output of the RF model is a 3D probability map of the same size as the input image, and the local maxima of the 3D probability map are the potential positions of the landmark in the image. Because these points are localized in the downsampled image, we calculate their coarse positions in the full resolution image by upsampling their coordinates. If, for a landmark L_i, multiple points are localized in a testing image, then the point that is the closest to $\bar{R}_i$ (Equation 2.1) is selected. For a testing image I_j, a small landmark set $\{L_k\}$, in which $\|L_k - R_{k,j}\| > MaxDist$ (Equation 2.2), which may potentially have localization errors are removed from the control points of the TPS registration for this image. For the testing images in our study, the maximum size of $\{L_k\}$ is 4 and the probability of this event is 1%, more frequently the size of $\{L_k\}$ is 2 or 3.

We use the presented method to compute the initialization transformation for the 5 deformable registration algorithms. In our experiments, the atlas image is first registered to each of the 20 testing images with the TPS-based transformations that use the *robust-set* as control points. Next, these transformed atlas images are registered to each of the testing images with each of the 5 deformable registration algorithms for further refinement. These registration methods are referred to as RBS-TPS-ABA, RBS-TPS-ART, RBS-TPS-DD, RBS-TPS-F3D, and RBS-TPS-SyN for simplicity. The only non-deterministic factor in our RBS-TPS- method is the random *sub-set*, which depends on the initial state of the random number generator that performs the random sampling. To test the sensitivity of our method to the initial state of the random number generator, we run our algorithm with 5 different initial states, and we use repeated measures ANOVA (Davis, 2002) to assess whether the performance of the RBS-TPS- approach is consistent across the 5 states.

The validation is performed by comparing our RBS-TPS- approach with four other approaches: (1) applying the deformable registration algorithms without preregistration initialization (referred to as WPI-ABA, WPI-ART, WPI-DD, WPI-F3D, and WPI-SyN), (2) applying the deformable registration algorithms after TPS preregistration initialization using 40 landmarks that are randomly selected from the 1802 candidate landmark set as control points (referred to as RND-TPS-ABA, RND-TPS-ART, RND-TPS-DD, RND-TPS-F3D, and RND-TPS-SyN), (3) applying the deformable registration algorithms after TPS preregistration initialization using the 40 landmarks that have the smallest mean registration error (Equation 2.3) as control points (referred to as MINERR-TPS-ABA, MINERR-TPS-ART, MINERR-TPS-DD, MINERR-TPS-F3D, MINERR-TPS-SyN), and (4) applying the deformable registration algorithms after preregistration with an affine transformation computed with a standard intensity-based registration algorithm that uses mutual information as similarity measure (referred to as AFI-ABA, AFI-ART, AFI-DD, AFI-F3D, and AFI-SyN). The comparison is performed qualitatively and quantitatively. First, we compare the registration results, i.e., the transformed atlas images obtained with the 5 registration approaches to the testing images. The approach that most often results in a better visual correspondence between the transformed atlas images and the testing images is deemed to be superior to the others. Second, we use the Dice Similarity Coefficients (DSC) (Sørensen, 1948) of the Gray Matter (GM), the White Matter (WM), and the CerebroSpinal Fluid (CSF) between the transformed atlas image and the testing image to quantify the similarity of the two images. To calculate the DSC, the brains of the atlas image and the 20 testing images are first segmented into GM, WM, and CSF with the FSL FMRIB's Automated Segmentation Tool (Yongyue Zhang et al., 2001). The segmented atlas is projected onto each of the segmented volumes in the testing set with the transformations computed in each of the 5 aforementioned registration approaches. The DSC of tissue class V is calculated as:

$$DSC_V = \frac{2 \times |V_{trans} \cap V_{test}|}{|V_{trans}| + |V_{test}|}, \tag{2.8}$$

in which V = {GM, WM, CSF}, V_{trans} and V_{test} denote the voxels with tissue label V in the transformed atlas image and the testing image respectively. $|V_{trans}|$ and $|V_{test}|$ are the numbers of voxels in the two groups, and $|V_{trans} \cap V_{test}|$ is the number of overlapping voxels of the two groups. For reference, we also calculate a baseline DSC as:

$$DSC_V = \frac{2 \times |V_{atlas} \cap V_{test}|}{|V_{atlas}| + |V_{test}|}, \tag{2.9}$$

in which V_{atlas} denotes the voxels with tissue label V in the original atlas image. Finally, paired t-tests (Welch, 1947) are performed to assess whether or not the DSC of the RBS-TPS- approach is statistically significantly different from the DSC of each of the other methods.

2.3 Results

Figure 2.5 shows the value of s (Equation 2.5) through iterations of the RANSAC algorithms. Convergence is achieved in about 30 iterations.

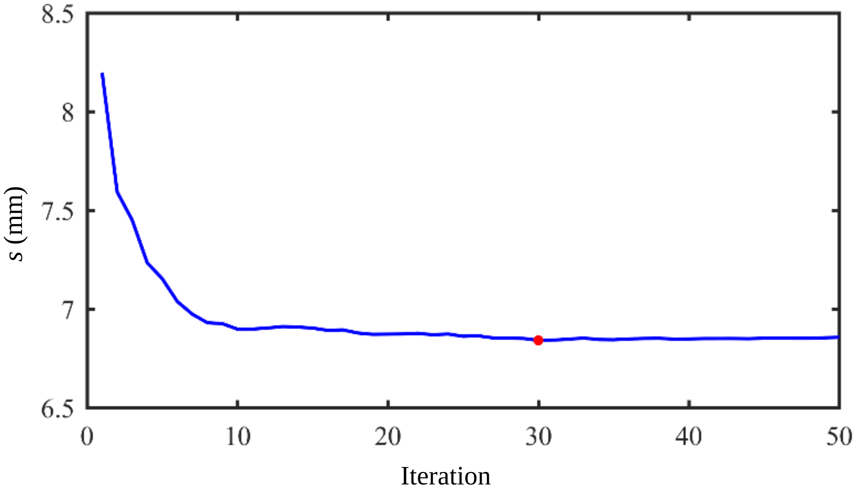


Figure 2.5: Registration error *s* at each iteration.

Results from the various registration approaches on three cases are shown in Figures 2.6−2.8. These image volumes have been selected from the 20 testing images based on their baseline DSC values. Case 1 has the highest baseline DSC among the 20 cases, Case 2 has the baseline DSC value that is closest to the mean baseline DSC value of the 20 cases, and Case 3 has the lowest baseline DSC value among the 20 cases. The accuracy of the registration obtained with WPI-ABA, WPI-ART, WPI-DD, WPI-F3D, and WPI-SyN, i.e., when the deformable registration method is applied without initialization is visually assessed, and we observe a failure rate of at least 20% for each algorithm. This confirms that deformable registration methods require a good preregistration initialization. The failure of the WPI- approaches is apparent in the frontal lobe regions (arrows on Figures 2.7 and 2.8, row 2) of Cases 2 and 3. On the same image volumes, both the AFI- approach and the RBS-TPS- approach lead to good results (Figures 2.7 and 2.8, row 3 and row 4). The ventricular region in the testing image is delineated (Figure 2.6−2.8, row 1, green contours) and the contours are copied to all the other images. This visually shows that the ventricles are accurately registered and that our RBS-TPS- approach produces results that are comparable to the standard AFI- approach for these structures.

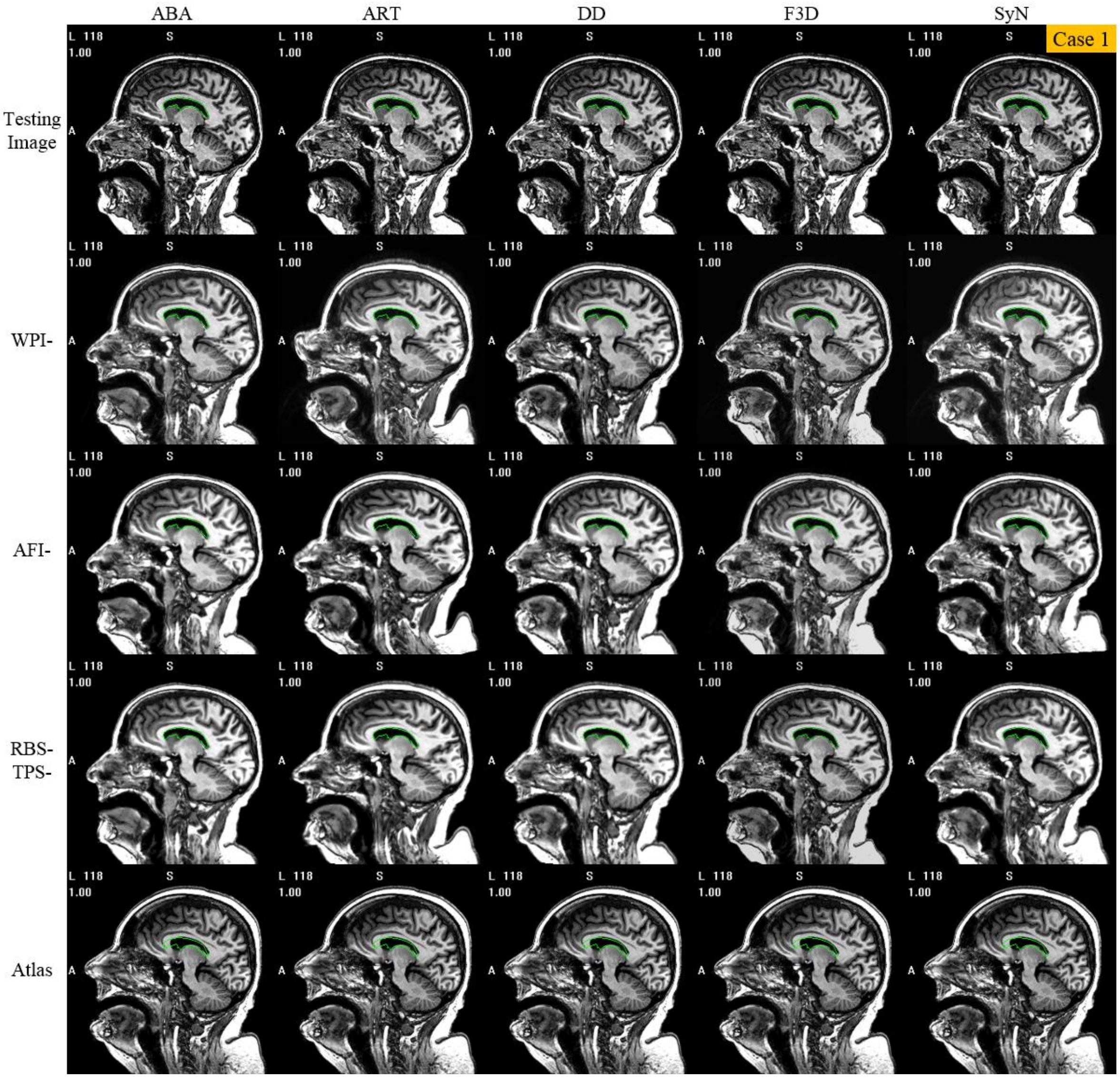


Figure 2.6: Sagittal view of Example Case 1. Shown are the testing image (row 1), transformed atlas images using the WPI-, the AFI-, and the RBS-TPS- approaches (row 2, 3, and 4, respectively); and the original atlas image (row 5). Column 1−5 of row 2−4 show the transformed atlas images when ABA, ART, DD, F3D, and SyN is used as the deformable registration algorithm. The green contours are drawn on the testing image and the contours are copied on the transformed atlas and the original atlas images.

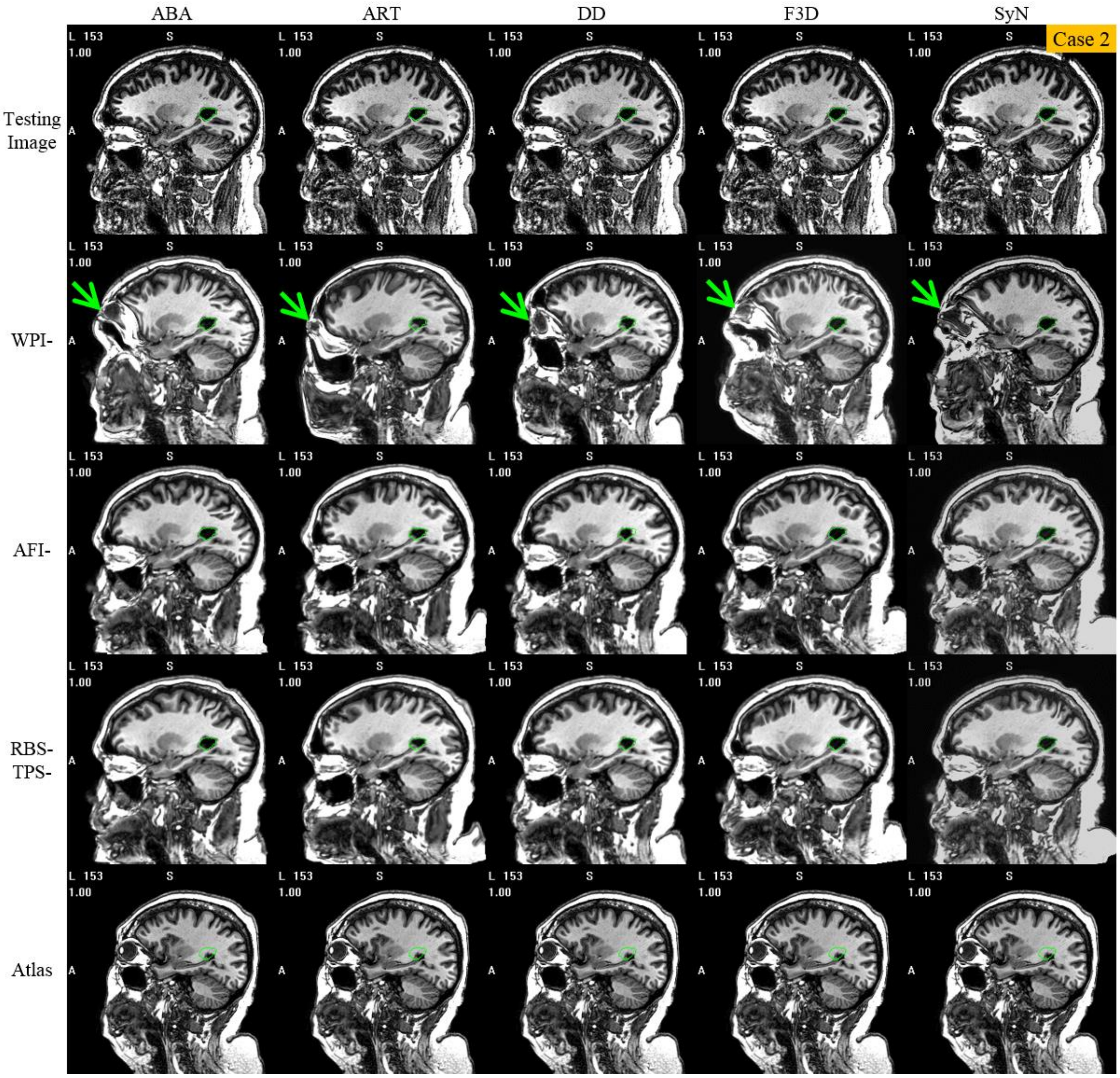


Figure 2.7: Sagittal view of Example Case 2. Shown are the testing image (row 1), transformed atlas images using the WPI-, the AFI-, and the RBS-TPS- approaches (row 2, 3, and 4, respectively); and the original atlas image (row 5). Column 1−5 of row 2−4 show the transformed atlas images when ABA, ART, DD, F3D, and SyN is used as the deformable registration algorithm. The green contours are drawn on the testing image and the contours are copied on the transformed atlas and the original atlas images. The green arrows point to regions where the registration failed.

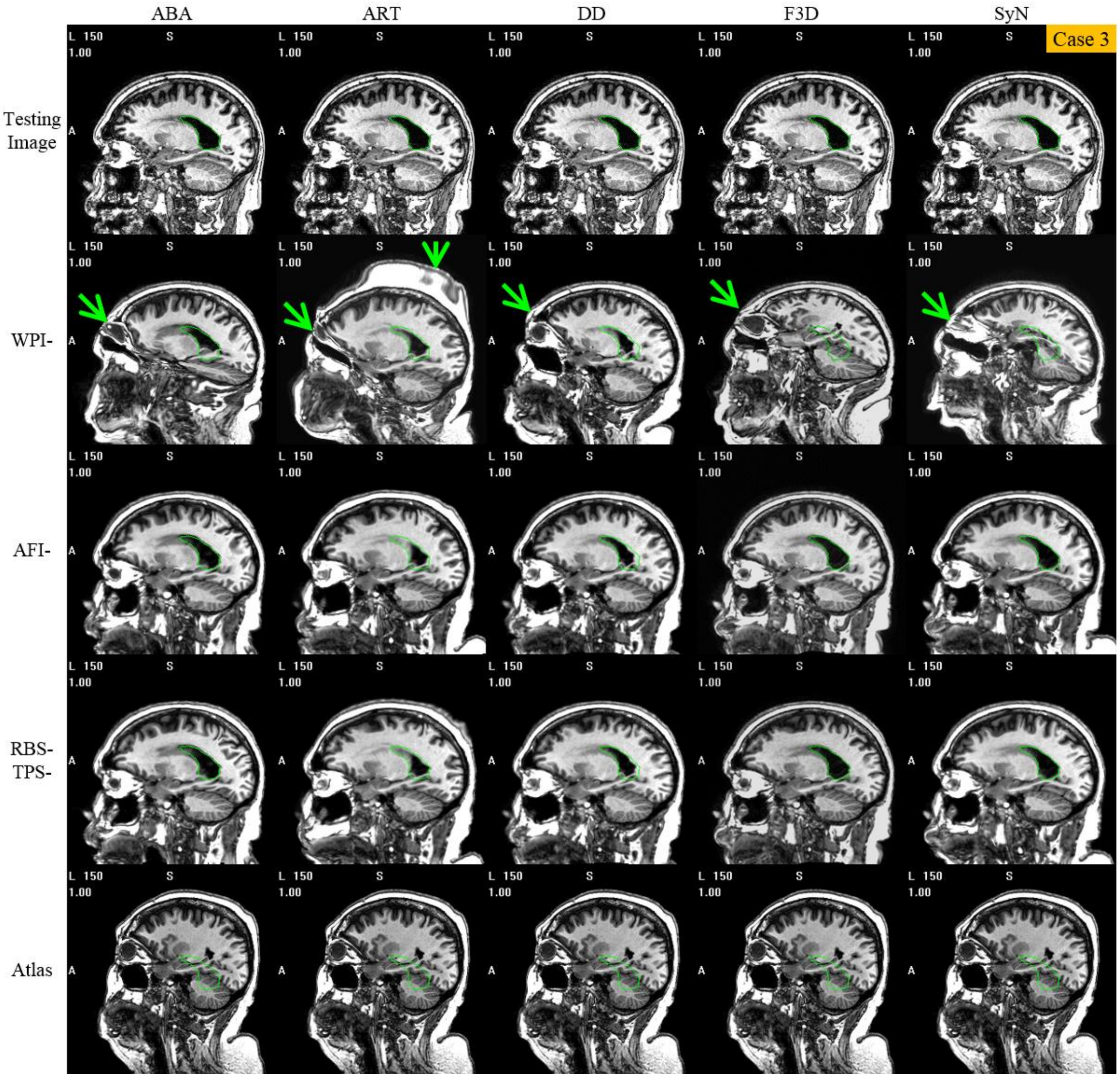


Figure 2.8: Sagittal view of Example Case 3. Shown are the testing image (row 1), transformed atlas images using the WPI-, the AFI-, and the RBS-TPS- approaches (row 2, 3, and 4, respectively); and the original atlas image (row 5). Column 1−5 of row 2−4 show the transformed atlas images when ABA, ART, DD, F3D, and SyN is used as the deformable registration algorithm. The green contours are drawn on the testing image and the contours are copied on the transformed atlas and the original atlas images. The green arrows point to regions where the registration failed.

Figures 2.9−2.13 show box plots of the DSC for CSF (red), GM (green), and WM (blue), for the baseline, the AFI- approach, and the RBS-TPS- approach with 5 different initial states. The same trend is observed for all 5 algorithms, i.e., the DSC values of the RBS-TPS- approaches are higher than those of the AFI- approaches. Paired T-tests (Table 2.2) show that for ABA, F3D, and SyN, the RBS-TPS- approach results in statistically significant higher DSC for WM, GM, and CSF than the AFI- approach ($p < 0.01$ for WM, GM, and CSF); for ART and DD, the RBS-TPS- approach results in statistically significant higher DSC for WM and GM than the AFI- approach ($p < 0.01$ for WM and GM), there is a substantial but not statistically significant difference between the DSC for CSF obtained with the two approaches ($p > 0.01$). Repeated measures ANOVA (Table 2.3) shows that statistically significant differences do not exist in the DSC of the 5 RBS-TPS- trials that are conducted with 5 different initial states, except for ART ($p < 0.01$ for WM). This suggests that our RBS-TPS- approach is robust against the initial state of the random number generator for 4 of the 5 registration methods that we test. For ART despite being statistically significant for the WM, the difference is small.

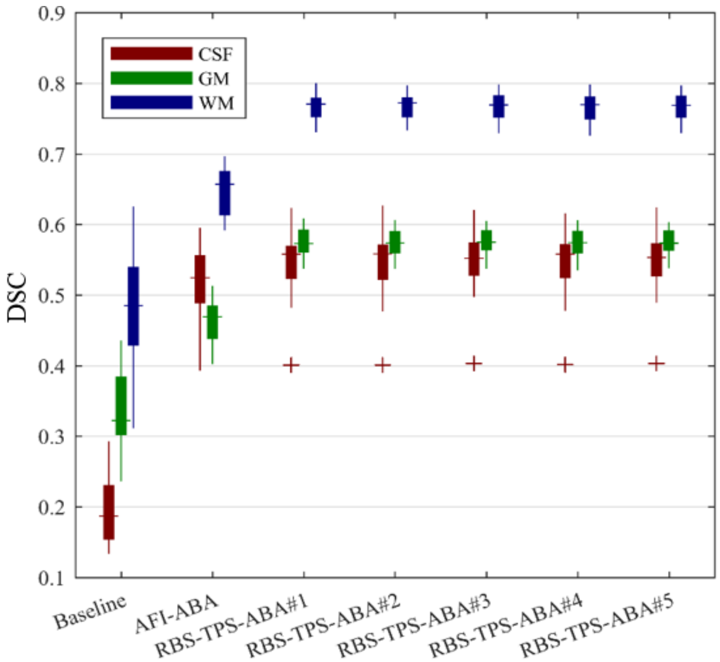


Figure 2.9: Box plots of the DSC of CSF, GM, and WM regions, for the baseline, AFI-ABA, and RBS-TPS-ABA.

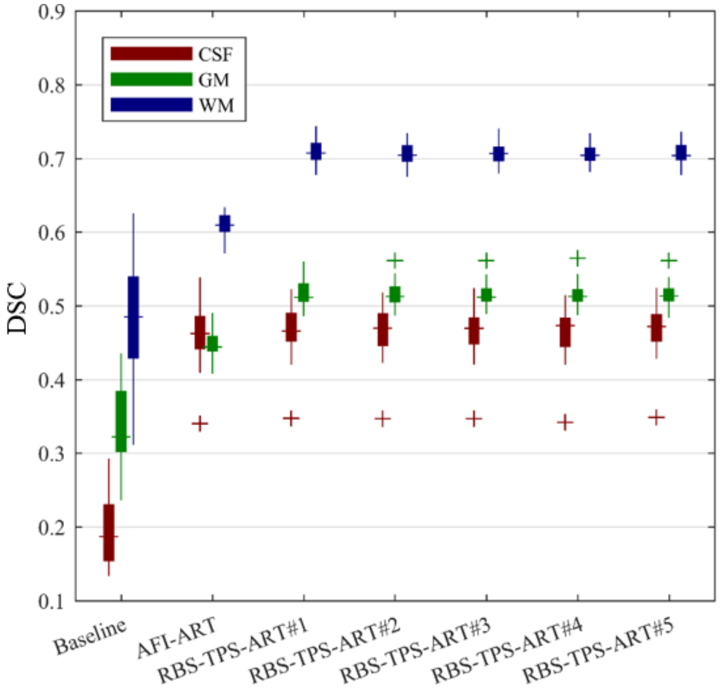


Figure 2.10: Box plots of the DSC of CSF, GM, and WM regions, for the baseline, AFI-ART, and RBS-TPS-ART.

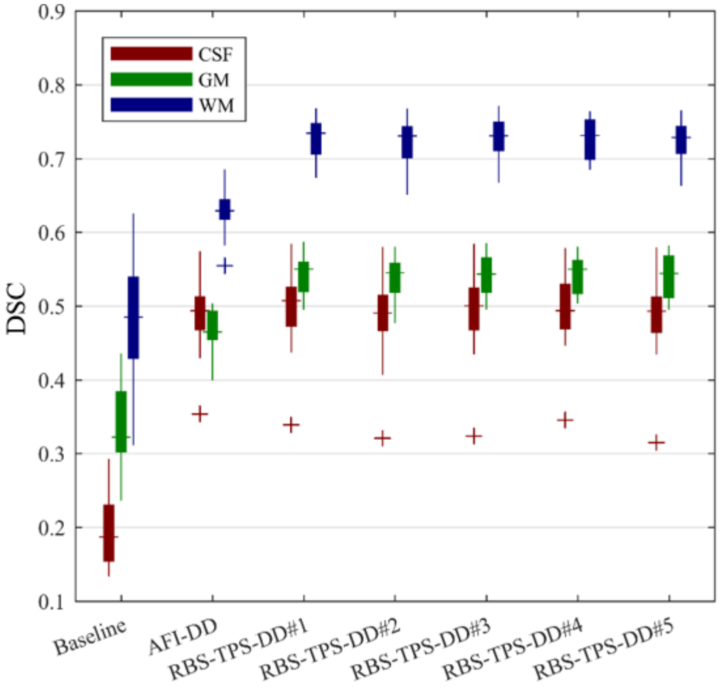


Figure 2.11: Box plots of the DSC of CSF, GM, and WM regions, for the baseline, AFI-DD, and RBS-TPS-DD.

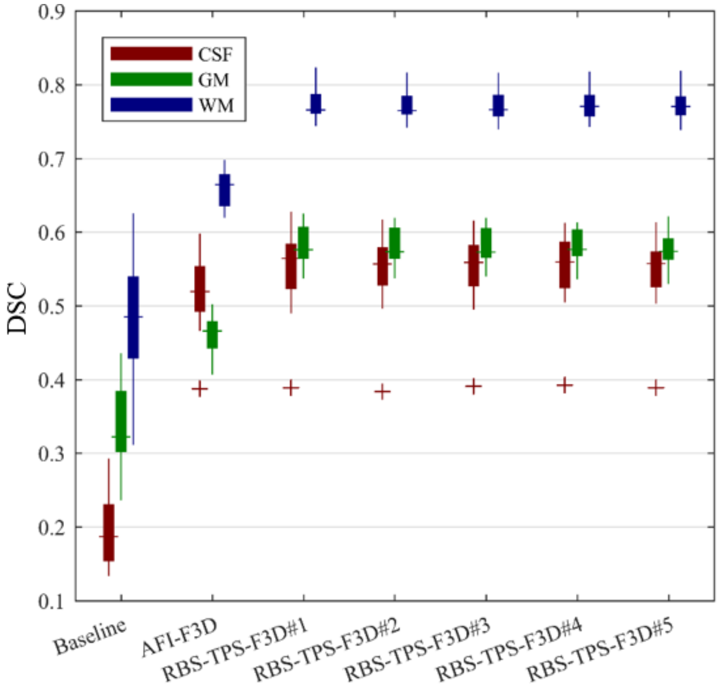


Figure 2.12: Box plots of the DSC of CSF, GM, and WM regions, for the baseline, AFI-F3D, and RBS-TPS-F3D.

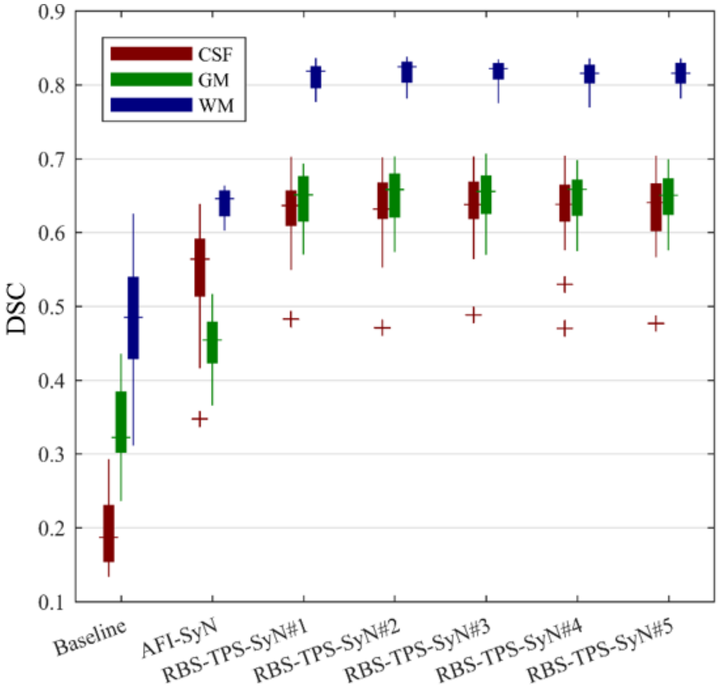


Figure 2.13: Box plots of the DSC of CSF, GM, and WM regions, for the baseline, AFI-SyN, and RBS-TPS-SyN.

RM	RBS-TPS-M#1			RBS-TPS-M#2			RBS-TPS-M#3			RBS-TPS-M#4			RBS-TPS-M#5		
	WM	GM	CSF	WM	GM	CSF	WM	GM	CSF	WM	GM	CSF	WM	GM	CSF
ABA	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
ART	<0.01	<0.01	**0.33**	<0.01	<0.01	**0.62**	<0.01	<0.01	**0.66**	<0.01	<0.01	**0.80**	<0.01	<0.01	**0.38**
DD	<0.01	<0.01	**0.47**	<0.01	<0.01	**0.66**	<0.01	<0.01	**0.86**	<0.01	<0.01	**0.65**	<0.01	<0.01	**0.76**
F3D	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
SyN	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Table 2.2: P-values of the paired t-test between DSC for WM, GM, and CSF for the RBS-TPS-approaches and the AFI- approaches. (RM – Registration Method)

Approach	WM	GM	CSF
RBS-TPS-ABA	0.694	0.725	0.183
RBS-TPS-ART	**<0.01**	0.094	0.274
RBS-TPS-DD	0.127	0.511	0.005
RBS-TPS-F3D	0.055	0.150	0.518
RBS-TPS-SyN	0.095	0.065	0.063

Table 2.3: P-values of the repeated measures ANOVA of the RBS-TPS- approaches with 5 different initial states.

Experiments show that the RND-TPS- approach and the MINERR-TPS-approach are not feasible. The TPS transformation fails or results in unreasonable deformation for some testing images when randomly selected landmarks are used as control points. The TPS transformation fails on 100% of our testing images when the landmarks that have the smallest mean registration error are used as control points. This is not unexpected because we have observed that these landmarks are typically located near the midbrain, thus they cannot provide good coverage of the brain (Figure 2.14).

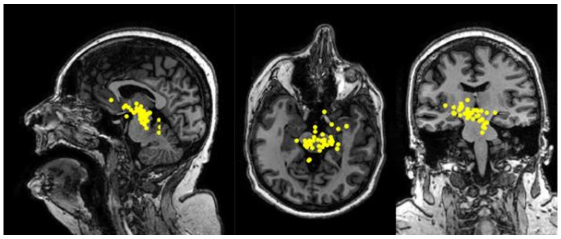


Figure 2.14: Landmarks that have the smallest mean registration error are shown on the same slice of the atlas image in the sagittal (left), axial (middle), and coronal (right) directions. All these landmarks are projected on the same slice to show their coverage range.

2.4 Conclusions

We present a novel approach for the selection of reliable landmarks for deformable registration initialization that uses a random forest approach followed by a random sample consensus step. The method we propose is fully automatic and generic. It could be applied to other registration problems. We have evaluated our approach by using it to initialize five well-established deformable registration algorithms and our results show that the same trend is observed for all 5 algorithms, i.e., the final registration results are statistically better with our approach than with a standard approach that relies on the estimation of an affine transformation computed with an intensity-based approach.

Because this approach operates on principles that are different from most non-rigid registration methods in routine use, it could also be used as an error detection mechanism. In this context, it could be run in parallel with existing processing pipelines and differences observed between methods in either deformation fields or landmark positions could trigger alerts; this will be explored in our future studies.

Acknowledgments

This study was supported by grant NIH R01-NS095291. The work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN. The content is solely the responsibility of the authors and does not necessarily represent the official views of these institutes.

CHAPTER III

Metal Artifact Reduction for the Segmentation of the Intracochlear Anatomy in CT Images of the Ear with 3D Conditional GANs

Jianing Wang, Jack H. Noble, and Benoit M. Dawant

Department of Electrical and Computer Engineering

Vanderbilt University

Nashville, TN, 37232, USA

This chapter is adapted from “Metal artifact reduction for the segmentation of the intra cochlear anatomy in CT images of the ear with 3D-conditional GANs” published in Medical Image Analysis, 2019, 58: 101553.

The published paper can be found at https://doi.org/10.1016/j.media.2019.101553

It has been reproduced with the permission of the publisher.

Abstract

Cochlear implants are surgically implanted neural prosthetic devices that are used to treat severe-to-profound hearing loss. These devices are programmed postoperatively and precise knowledge of the implant position with respect to the intracochlear anatomy can help the programming audiologists. Over the years, we have developed algorithms that permit determining the position of implanted electrodes relative to the intracochlear anatomy using pre- and post-implantation CT image pairs. However, these do not extend to recipients for whom Pre-implantation CT (Pre-CT) images are not available. This is so because post-operative images are affected by strong artifacts introduced by the metallic implant. To overcome this issue, we have proposed two methods to segment the intracochlear anatomy in Post-implantation CT (Post-CT) images, but they lead to segmentation errors that are substantially larger than errors obtained with Pre-CT images. We have proposed an approach that uses 2D-conditional generative adversarial nets to synthesize pre-operative images from post-operative images. This permits to use of segmentation algorithms designed to operate on Pre-CT images even when these are not available. We have shown that it substantially and significantly improves the results obtained with methods designed to operate directly on Post-CT images. In this chapter, we expand on our earlier work by moving from a 2D architecture to a 3D architecture. We perform a large validation and comparative study that shows that the 3D architecture improves significantly the quality of the synthetic images measured by the commonly used mean structural similarity index. We also show that the segmentation results obtained with the 3D architecture are better than those obtained with the 2D architecture although differences have not reached statistical significance.

3.1 Introduction

The cochlea is a component of the inner ear. It is a spiral-shaped cavity located inside the bony labyrinth that contains two main cavities: the scala vestibuli and the scala tympani. The modiolus is a porous bone around which the cochlea is wrapped that hosts the cochlear nerve and the spiral ganglions. These structures, which are the ones this chapter focuses on, are shown in Figure 3.1 and will be referred to together as intracochlear anatomy. Cochlear implants are surgically implanted neural prosthetic devices that are used to treat severe-to-profound hearing loss (National Institute on Deafness and Other Communication Disorders, 2016). Cochlear implants bypass the normal acoustic hearing process by replacing it with direct stimulation of neural pathways using an implanted electrode array. After implantation cochlear implants are programmed by audiologists who adjust a processor’s settings to send the appropriate signals to each of the implant’s electrodes. The efficacy of the programming is sensitive to the spatial relationship between the cochlear implant’s electrodes and the intracochlear anatomy. Providing accurate information about the position of the contacts with respect to these structures can thus help audiologists to fine-tune and customize the programming of the cochlear implants (Noble et al., 2013). To provide this information we have developed several algorithms that permit determining the position of implanted electrodes relative to the intracochlear anatomy using pre- and post-implantation CTs.

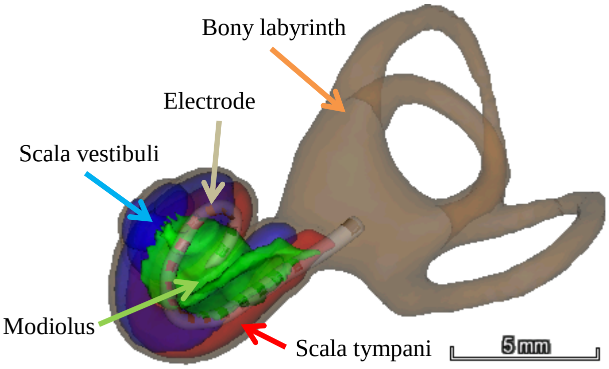


Figure 3.1: An illustration of intracochlear anatomical structures and electrodes of the cochlear implant.

Pre-implantation CT (Pre-CT) images and Post-implantation CT (Post-CT) images of the ear are acquired before and after the surgery, respectively. The cochlear implant’s electrodes are localized in the Post-CT images using the automatic methods proposed by Zhao et al. (Zhao et al., 2018, 2019). It is difficult to directly localize the intracochlear anatomy in the Post-CT images due to the strong artifacts produced by the metallic electrodes. The intracochlear anatomy is thus localized in the Pre-CT images and the position of the electrodes relative to the intracochlear anatomy is obtained by registering the Pre-CT images and the Post-CT images. To localize the intracochlear anatomy in the Pre-CT images, where the intracochlear anatomy is only partially visible, Noble et al. (Noble et al., 2011) have developed a method, which we refer to as SegPre-ASM. SegPre-ASM relies on a weighted active shape model created with high-resolution microCT scans of the cochlea acquired ex-vivo in which the intracochlear anatomy is visible. The model is fitted to the partial information available in the Pre-CT images and used to estimate the position of structures not visible in these images.

This approach does not extend to the recipients for whom a Pre-CT image is unavailable. To overcome this issue, Reda et al. (Reda, McRackan, et al., 2014; Reda, Noble, et al., 2014) have proposed two methods to segment the intracochlear anatomy in Post-CT images. The first method, which we refer to as SegPost-UL for Post-CT segmentation unilateral, was developed for segmenting the intracochlear anatomy in Post-CT images of recipients who have been implanted unilaterally (Reda, McRackan, et al., 2014). SegPost-UL relies on the intra-subject symmetry in cochlear anatomy across ears. It first segments the intracochlear anatomy of the contralateral normal ear and then maps the segmented structures to the implanted ear. The second method, which we refer to as SegPost-BL for Post-CT segmentation bilateral, was developed for segmenting the intracochlear anatomy in Post-CT images of recipients who have been implanted bilaterally (Reda, Noble, et al., 2014). SegPost-BL first identifies the labyrinth in the Post-CT image by mapping a labyrinth surface that is selected from a library of labyrinth surfaces and then uses the localized labyrinth in the image as a landmark to segment the scala tympani, the scala vestibuli, and the modiolus with a standard shape model-based segmentation method.

But, while using these methods it was observed that they could at times lead to results that lacked accuracy compared to other components of the processing pipeline on which we rely to provide programming guidance to the audiologists. For instance, we can localize contacts in electrode arrays with an average accuracy better than 0.15mm (Zhao et al., 2018, 2019) when the segmentation error of SegPost-UL applied to the unilateral cases, which we refer to as IU for implanted unilaterally, included in this study is 0.26mm. The segmentation error of SegPost-BL applied to the bilateral cases, which we refer to as IB for implanted bilaterally, is larger and reaches 0.44mm. These observations led us to explore ways to improve our segmentation accuracy.

Generative Adversarial Nets (GANs) (Goodfellow et al., 2020) have been applied to various computer vision tasks and have produced impressive results. In particular, conditional Generative Adversarial Networks (cGANs) (Mirza & Osindero, 2014) have emerged as a general-purpose solution to image-to-image translation problems. Inspired by these, we proposed an alternative for localizing the intracochlear anatomy in the Post-CT images (J. Wang et al., 2018). First, we train 2D-cGANs introduced by Isola et al. (Isola et al., 2017) to synthesize artifact-free images from the Post-CT images, i.e., we train a network whose input is a 2D slice in a volume in which the artifacts are present and whose output is the corresponding synthetic artifact-free image. Once this is done for all slices, the synthetic 2D images are stacked to each other, and SegPre-ASM is applied to the synthetic volume. Results obtained with SegPre-ASM on the synthesized volumes and the real Pre-CT volumes can then be compared to assess the efficacy of the artifact removal method. In this earlier study performed on 74 ears, we show that this approach produces segmentation errors that are about half the segmentation errors that were obtained with the methods designed to operate on the post-operative images. In this chapter, we increase the size of our testing dataset to 124 ears and we explore ways to improve our method further by (1) using a 3D architecture rather than a 2D architecture and (2) modifying the training objective of the 3D-cGANs, which is a sum of an adversarial loss and a weighted L1 reconstruction loss. The quality of the artifact-corrected images is evaluated quantitatively by computing the surface error between the segmentation of the intracochlear anatomy obtained with SegPre-ASM applied to the real Pre-CT images and to the artifact-corrected CT images. We further validate our method by comparing the results obtained with SegPre-ASM applied to the artifact-corrected CT images and those which are obtained with SegPost-UL and SegPost-BL directly applied to the Post-CT images. Finally, as is commonly done to assess the quality of images, we compare the Mean Structural SIMilarity (MSSIM) index (Z. Wang et al., 2004) between the real images and the synthetic images obtained with the 2D and 3D architectures.

3.2 Material and Methods

3.2.1 Training Objectives

3.2.1.1 Adversarial Loss

Typically, GANs are implemented by a system of a generative network (G) and a discriminative network (D) that are competing with each other. G learns a mapping between a latent space and a particular data distribution of interest, while D discriminates between instances from the true data distribution and candidates produced by G. The training objective of G is to increase the error rate of D, i.e., to fool D by producing synthesized candidates that appear to come from the true data distribution (Goodfellow et al., 2020). cGANs are a special case of GANs in which both G and D are conditioned on additional information that is used to direct the data generation process. This makes cGANs suitable for image-to-image translation task, where G is conditioned on an input image and generates a corresponding output image (Mirza & Osindero, 2014; Isola et al., 2017).

For our purpose, which is to eliminate the artifacts produced by the cochlear implants, we use cGANs that are conditioned on the artifact-affected Post-CT images. G thus produces an artifact-free image $G(post)$ from a Post-CT image $post$, and $G(post)$ should not be distinguishable from the real artifact-free Pre-CT image pre by D, which is trained to do as well as possible to detect G's "fakes". The input of D is the concatenation of $post$ and a real or synthetic Pre-CT image. The output of D can be interpreted as the probability of the input Pre-CT image to be generated by G rather than real. Therefore, the training objective of D is to assign a high value to $G(post)$ and a low value to pre. Conversely, the training objective of G is to fool D to assign a low value to $G(post)$ and a high value to pre. Thus, the adversarial loss of the cGANs can be expressed as:

$$L_{\mathrm{cGAN}}(G, D) = \min_{G} \max_{D} \mathbb{E}_{post,pre}\left[\log\left(D(post, pre)\right)\right] + \mathbb{E}_{post}\left[\log\left(1 - D\left(post, G(post)\right)\right)\right] \tag{3.1}$$

3.2.1.2 Reconstruction Loss

Previous research suggests that it is beneficial to mix the adversarial loss with a more traditional reconstruction loss, such as the L1 distance between $G(post)$ and pre (Isola et al., 2016), which is defined as:

$$L_{\mathrm{L}_1}(G) = \mathbb{E}_{post,pre}[\|pre - G(post)\|_1] \tag{3.2}$$

For our ultimate purpose, which is to localize the intracochlear anatomy in the Post-CT images, we are more concerned about the quality of the image content in the small region that encompasses the cochlea than in the other regions in an artifact-corrected CT image, therefore we assign a higher weight to the voxels inside this region when calculating the L1 loss. To do so, we first create a bounding box that encloses the cochlea. With the number of voxels inside the bounding box equal to N_{in} and the number of voxels outside of the bounding box equal to N_{out}, we assign weights to the voxels inside and outside of the bounding box to $(N_{in} + N_{out})/N_{in}$ and 1, respectively. The Weighted L1 (WL1) loss can then be expressed as shown in Equation 3.3:

$$L_{\mathrm{WL}_1}(G) = \mathbb{E}_{post,pre}\left[\left\|W \circ \left(pre - G(post)\right)\right\|_1\right] \tag{3.3}$$

in which W is the weighting matrix and $\circ$ is the element-wise multiplication operation.

3.2.1.3 Total Loss

The total loss can be expressed as a combination of the adversarial loss and the reconstruction loss:

$$L = \arg\min_{G}\max_{D} L_{\mathrm{cGAN}}(G, D) + \alpha L_{\mathrm{WL}_1}(G) \tag{3.4}$$

wherein α is the weight of the WL1 term.

3.2.2 Architecture of the 3D-cGANs

Figure 3.2 shows the architecture of our 3D-cGANs. The generator is a 3D network that consists of 3 convolutional blocks followed by 6 ResNet blocks (He et al., 2015), and another 3 convolutional blocks (Figure 3.2, the sub-network on the left). As is done in Isola et al. (Isola et al., 2017), dropout is applied to introduce randomness into the training of the generator. The input of the generator is a 1-channel 3D Post-CT image, and the output is a 1-channel 3D synthetic Pre-CT image. The discriminator is a fully convolutional network (Figure 3.2, the sub-network on the right) that maps the input, which is the concatenation of a Post-CT image and the corresponding Pre-CT image (or a Post-CT image and the synthetic Pre-CT image), to a 3D array d, in which each $d_{i,j,k}$ captures whether the (i, j, k)-th 3D patch of the input is real or fake. The ultimate output of the discriminator is a scalar obtained by averaging d.

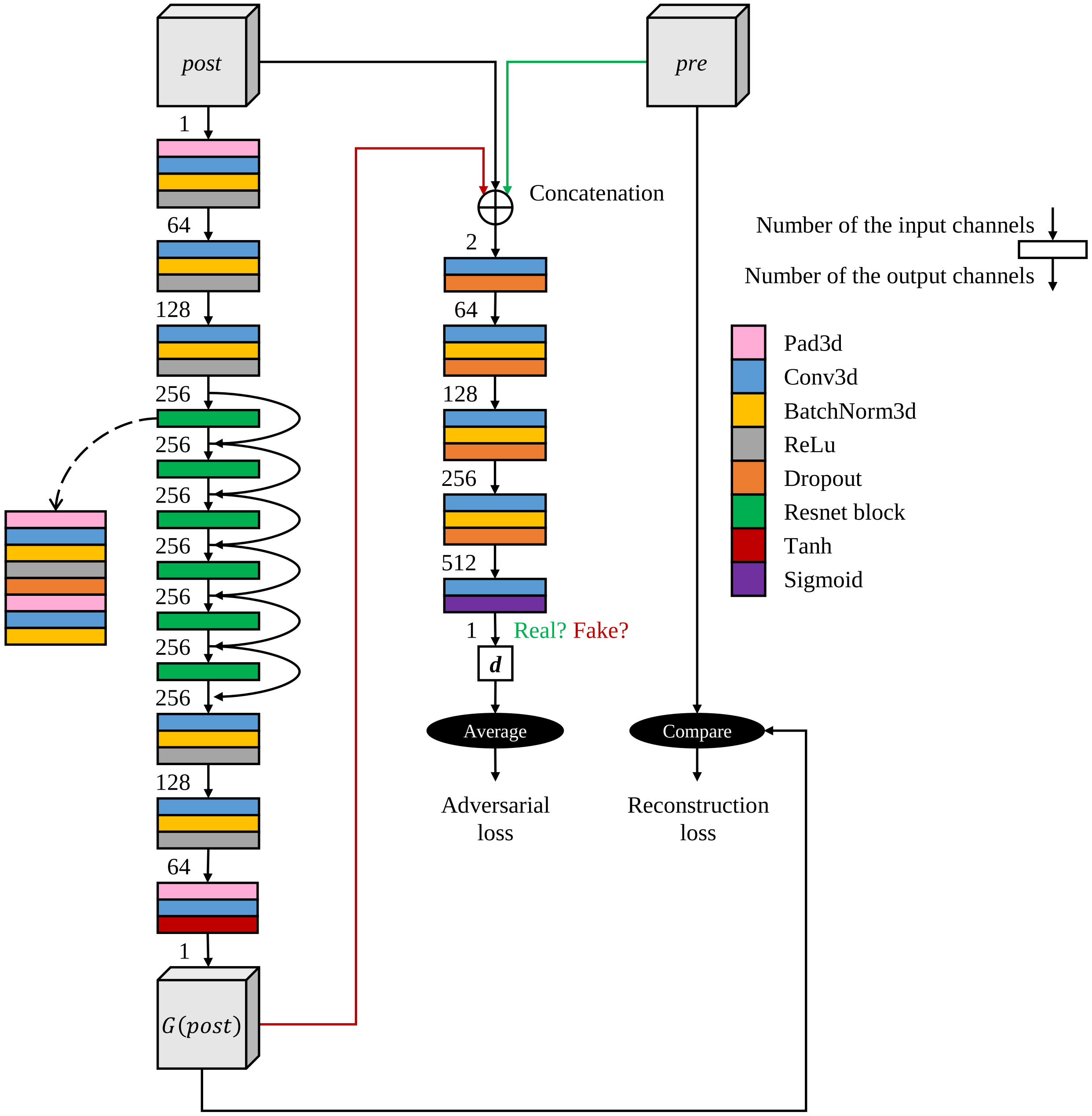


Figure 3.2: An illustration of the architecture of the 3D-cGANs, in which *post* is a Post-CT image, *pre* is a real Pre-CT image and $G(post)$ is a synthetic Pre-CT image.

3.3 Experiments

3.3.1 Dataset

Our dataset consists of Post-CT and Pre-CT image pairs of 252 ears, all these CT volumes have been acquired with the recipients of the cochlear implants in roughly the same position. 24 Post-CT images and all of the 252 Pre-CT images were acquired with several conventional scanners referred to as cCT scanners (GE BrightSpeed, LightSpeed Ultra; Siemens Sensation 16; and Philips Mx8000 IDT, iCT 128, and Brilliance 64). The other 228 Post-CT images were acquired with a low-dose flat-panel volumetric CT scanner referred to as lCT scanner (Xoran Technologies xCAT® ENT). The typical voxel size is and 0.25×0.25×0.3 mm^3 for the cCT images, and 0.4×0.4×0.4 mm^3 for the lCT images. The 252 ears are randomly partitioned into a set of 90 ears for training, a set of 25 ears for validation, and a set of 137 ears for testing. After random assignment, there are 13 bilateral cases for which one ear has been assigned to the training (or validation) set and the other ear has been assigned to the testing set, the 13 ears are removed from the testing set so that no image from the same patient are used for both training and testing. Details about our image set can be found in Table 3.1.

Usage	Total number of the ears		Number of Post- and Pre-CT pairs	
			lCT-cCT	cCT-cCT
Training	90		82	8
Validation	25		21	4
Testing	124	88 IB ears	78	10
		36 IU ears	34	2

Table 3.1: The number of ears and the type of CT scanner used to acquire the images in the training, validation, and testing sets. "lCT-cCT" denotes that the ear has been scanned by the lCT scanner postoperatively and a cCT scanner preoperatively, and "cCT-cCT" denotes that the ear has been scanned by a cCT scanner postoperatively and preoperatively.

The Pre-CT images are registered to the Post-CT images using intensity-based affine registration techniques (Wells et al., 1996; Maes et al., 1997). The registrations have been visually inspected and confirmed to be accurate. We apply image augmentation to the training set by rotating each image by 20 small random angles in the range of -10 and 10 degrees about the x-, y-, and z-axis, such that 60 additional training images are created from each original image. This results in a training set that is expanded to 5490 volumes.

Because in our dataset the typical voxel size of the Post-CT images is 0.4×0.4×0.4mm^3, we first resample the CT images to 0.4×0.4×0.4mm^3, so that all of the images have the same resolution. 3D patch pairs that contain the cochlea are cropped from the Pre-CT and Post-CT images, i.e., paired patches contain the same cochlea; one patch with and the other without the implant (Figure 3.3). The size of the patches is 38.4×38.4×38.4mm^3 (96×96×96 voxels). Each patch is clamped to the range 0.1 to 99.9-th percentiles of its intensity values. Then the patches are rescaled to the -1 to 1 range.

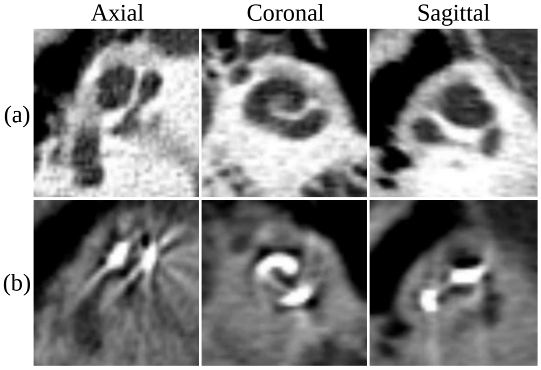


Figure 3.3: Three orthogonal views of (a) the Pre-CT image and (b) the Post-CT image of an example ear.

3.3.2 Optimization and Inference

Our PyTorch implementation of the 3D-cGANs is adapted from the 2D implementation provided by Zhu et al. (Zhu et al., 2017). α introduced in Equation 3.4 is set to its default value 100. In practice, the cochlea is at the center of each 3D patch and we simply use the central 56×56×56 voxels of the 3D patch as the bounding box for calculating the weights. The 3D-cGANs are trained alternatively between one stochastic gradient descent step on the discriminator, then one step on the generator, using a minibatch size of 1 and the Adam solver (Kingma & Ba, 2014) with a momentum of 0.5. The 3D-cGANs are trained for 200 epochs in which a fixed learning rate of 0.0002 is applied in the first 100 epochs and a learning rate that is linearly reduced to zero in the second 100 epochs. At the inference phase, given an unseen Post-CT patch, the generator produces an artifact-corrected image.

The MSSIM, which will be introduced in Section 3.3.3.2, inside the 56×56×56 bounding box of the true Pre-CT images and the artifact-corrected images generated by the cGANs has been used to select the number of training epochs. To do so we run inference on the validation set every 5 epochs, the MSSIM is calculated for each of the ears, and the epoch where it achieves the highest median MSSIM is selected as the optimal epoch.

3.3.3 Evaluation

The proposed method is compared to the published baseline methods SegPost-UL and SegPost-BL as well as to our earlier 2D-cGANs-based method. As we have done in our earlier study (J. Wang et al., 2018), we upsample the voxel size of the CT images to $0.1\times0.1\times0.1\text{mm}^3$ to train and test the 2D-cGANs. This was done to improve slice-to-slice consistency. Due to memory limitations, this is not possible for the 3D-cGANs that are trained with volumes.

To evaluate the quality of the synthetic images independently from the segmentation results we compare the MSSIM between the original Pre-CT images and the images produced with the 2D and 3D architectures. We also compare the performance of the 3D-cGANs trained using the weighted L1 loss and those which are trained using the original L1 loss.

3.3.3.1 Point-to-point Errors

The effect of artifact reduction on segmentation accuracy is evaluated quantitatively by comparing the segmentation of the structures of interest obtained with SegPre-ASM applied to the real Pre-CT images with the results obtained when applying SegPre-ASM to the artifact-corrected CT images. Because SegPre-ASM is based on an active shape model approach, the outputs of SegPre-ASM are surface meshes of the scala tympani, the scala vestibuli, and the modiolus that have a pre-defined number of vertices, and each vertex corresponds to an anatomical location on the surface of the structures. There are 3344, 3132, and 17947 vertices on the scala tympani, scala vestibuli, and modiolus surfaces, respectively, for a total of 24423 vertices. Point-to-Point Errors (P2PEs), computed as the Euclidean distance in millimeter, between the corresponding vertices on the meshes generated from the real Pre-CT images and the meshes generated from artifact-corrected images are calculated to quantify the quality of the artifact-corrected images.

To compare the P2PEs of the 3D-cGANs-based method to results obtained with the published baseline methods, we segment the scala tympani, the scala vestibuli, and the modiolus with SegPost-UL and SegPost-BL in the Post-CT images of the IU ears and the IB ears, respectively. The output of SegPost-UL and SegPost-BL are surface meshes for the scala tympani, the scala vestibuli, and the modiolus that have the same anatomical correspondences as the meshes generated by SegPre-ASM. The P2PEs between the corresponding vertices on the meshes generated with SegPre-ASM in the real Pre-CT images and the meshes generated with SegPost-UL and SegPost-BL in the Post-CT images serve as baselines for comparison.

To compare the 3D-cGANs-based method to our earlier 2D-cGANs-based method, the P2PEs between the corresponding vertices on the meshes generated from the real Pre-CT images with SegPre-ASM and the meshes generated from artifact-corrected images generated by the 2D-cGANs with SegPre-ASM are also calculated.

3.3.3.2 Mean Structural Similarity

To compare the quality of the artifact-corrected images produced by our earlier 2D-cGANs and those which are generated by the 3D-cGANs, we compare the MSSIM inside the 56×56×56 bounding box of the true Pre-CT images and the artifact-corrected images generated by the 2D- and the 3D-cGANs. The MSSIM between the true Pre-CT images and the Post-CT images serves as a baseline for comparison. The MSSIM between the artifact-corrected CT image $G(post)$ and the true Pre-CT image pre can be expressed as:

$$\mathrm{MSSIM}(G(post), pre) = \frac{1}{M}\sum_{j=1}^{M} \mathrm{SSIM}(g_j, pre_j) \tag{3.5}$$

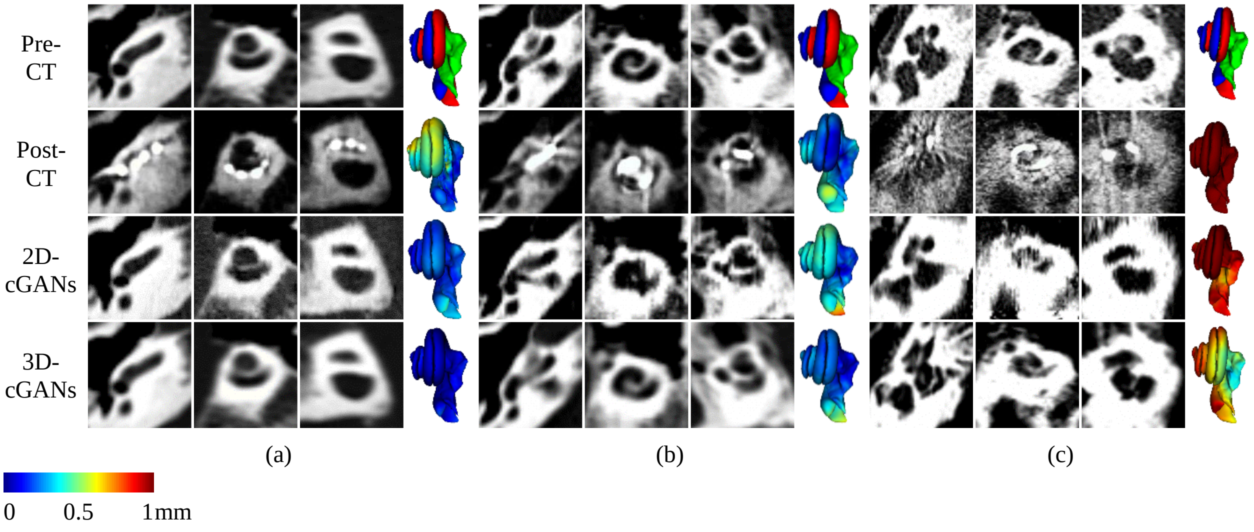
wherein $\mathrm{SSIM}(g_j, pre_j)$ is the local Structural SIMilarity (SSIM) between g_j and pre_j, which are the image contents at the jth local window of $G(post)$ and pre, and M is the number of local windows in the image. The local SSIM can be expressed as:

$$\mathrm{SSIM}(g_j, pre_j) = \frac{\left(2\mu_{g_j}\mu_{pre_j}+C_1\right)\left(2\sigma_{g_j pre_j}+C_2\right)}{\left(\mu_{g_j}^2+\mu_{pre_j}^2+C_1\right)\left(\sigma_{g_j}^2+\sigma_{pre_j}^2+C_2\right)} \tag{3.6}$$

in which μ_{g_j}, μ_{pre_j}, σ_{g_j}, σ_{pre_j}, and $\sigma_{g_j pre_j}$ are the local means, standard deviations, and cross-covariance of g_j and pre_j, and C_1 and C_2 are constants to avoid instability when $\mu_{g_j}^2 + \mu_{pre_j}^2$ or $\sigma_{g_j}^2 + \sigma_{pre_j}^2$ are close to zero (Z. Wang et al., 2004).

3.4 Results

Figure 3.4 shows 3 example cases in which our proposed method leads to (a) good, (b) average, and (c) poor results. For each case, the first row shows three orthogonal views of the Pre-CT image and the meshes generated when applying SegPre-ASM to this image. The scala tympani, the scala vestibuli, and the modiolus surfaces are shown in red, blue, and green, respectively. The second row shows the Post-CT image and the meshes generated when applying SegPost-BL (or SegPost-UL) to this image. The third and the last rows show the outputs of the 2D- and the 3D-cGANs and the meshes generated when applying SegPre-ASM to these images. The meshes from the second to the last rows are color-coded with the P2PE at each vertex on the meshes. Notably, even in the worst case, segmentation errors of the 3D-cGANs are lower than those of the baseline and the 2D-cGANs. Note also the severity of the artifact in this case and the failure of the segmentation method designed to operate on the Post-CT images. The values of the MSSIM and the type of Post- and Pre-CT pairs for these examples are listed in Table 3.2. In all cases, the MSSIM between the original images and the synthetic images produced by the 3D networks is higher than between the original images and the synthetic images produced by the 2D networks. This is consistent with the visual appearance of the synthetic images as can be appreciated by comparing rows 3 and 4 of Figure 3.4.



(a)

(b)

(c)

Figure 3.4: Three example cases in which our proposed method leads to (a) a good, (b) average, and (c) poor results (see text for details).

Image name	Baseline	2D-cGANs	3D-cGANs	Type of the Post- and Pre-CT pairs
Figure 3.4a	0.771	0.891	0.971	lCT-cCT
Figure 3.4b	0.499	0.780	0.931	lCT-cCT
Figure 3.4c	0.348	0.473	0.552	lCT-cCT

Table 3.2: The values of the MSSIM between the true Pre-CT images and the artifact-corrected CT images generated by the 2D- and the 3D-cGANs. “lCT-cCT” denotes that the ear has been scanned by the lCT scanner postoperatively and a cCT scanner preoperatively.

3.4.1 Point-to-point Errors

For each testing ear, we calculate the P2PEs of the 24423 vertices, and we calculate the MAXimum (MAX), MEAn (MEA), MEDian (MED), STandard Deviation (STD), and MINimum (MIN) of the P2PEs.

Figures 3.5a and 3.5b show the boxplots of these statistics for the 88 IB ears and the 36 IU ears. SegPost-BL and SegPost-UL serve as the baseline method for the bilateral and unilateral cases, respectively. Figure 3.5a shows that both the 2D- and the 3D-cGANs-based methods substantially reduce the P2PEs obtained with SegPost-BL in the Post-CT images. The median of P2PEs of the baseline method is 0.439mm, the medians of the P2PEs of the 2D- and the 3D-cGANs-based approaches are 0.233mm and 0.198mm, which are about half of the baseline method. We perform a Wilcoxon signed-rank test (McDonald, 2014) between the MAX, MED, MEA, STD, and MIN values obtained with the baseline method and the cGANs-based methods, and the resulting p-values are corrected using the Holm-Bonferroni method (Holm, 1979). The results show that the cGAN-based methods significantly reduce the P2PEs compared to the baseline method ($p < 0.05$) (Table 3.3. 88 IB ears, row 1 and 2). We also perform a Wilcoxon signed-rank test between the 2D- and the 3D-cGANs-based approaches that show that despite being visible the difference between the results of the 2D- and the 3D-cGANs are not statistically significant ($p > 0.05$) (Table 3.3. 88 IB ears, row 3). Figure 3.5b shows that both the 2D- and the 3D-cGANs-based methods reduce the P2PEs obtained with SegPost-UL in the Post-CT images. The median of the P2PEs of the baseline method is 0.260mm, whereas the medians of the P2PEs of the 2D- and the 3D-cGANs are 0.194mm and 0.188mm, respectively. A Wilcoxon signed-rank test shows that the cGANs-based methods significantly reduce the P2PEs compared to the baseline method for

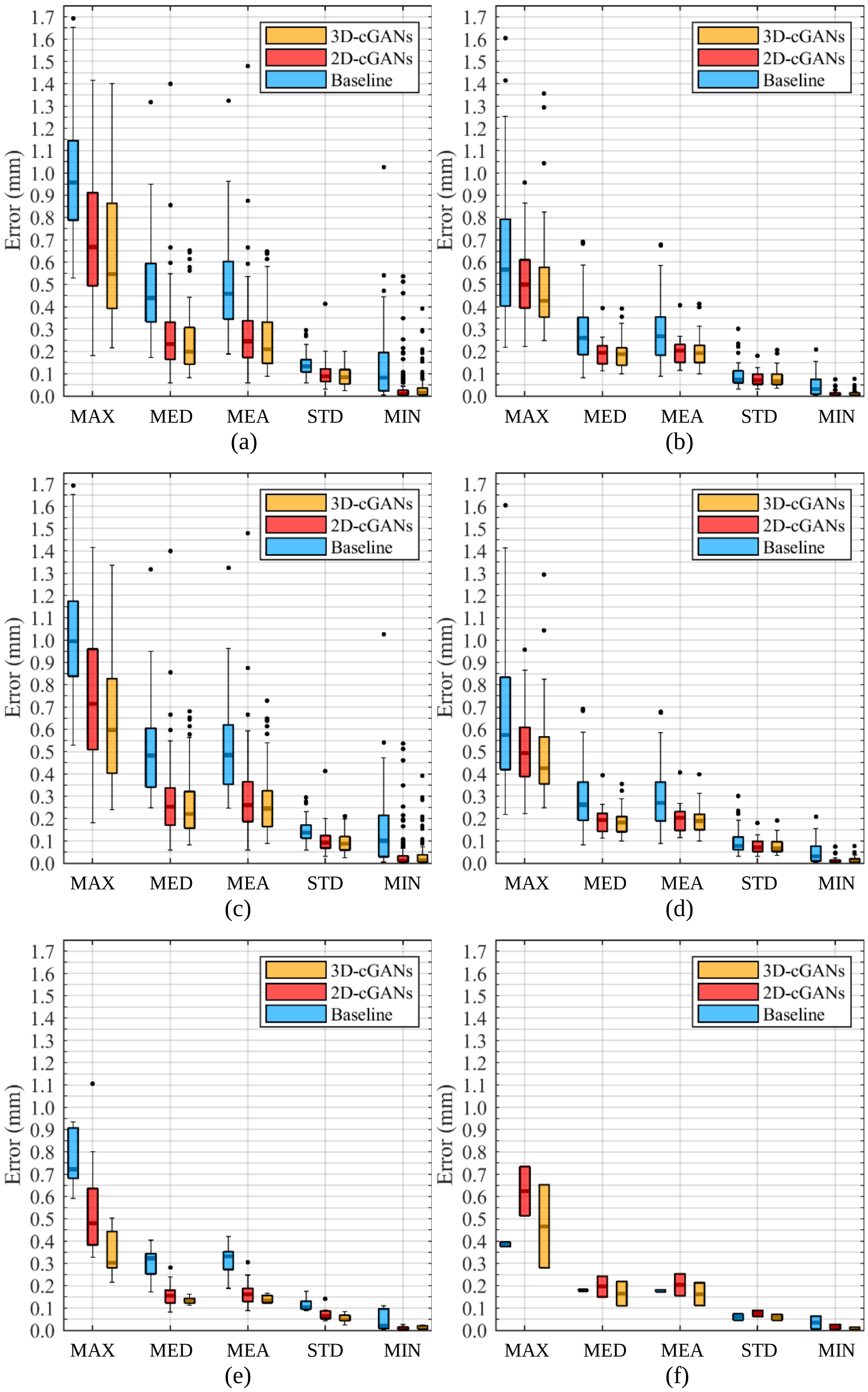
MED and MEA (p < 0.05) (Table 3.3. 36 IU ears, row 1 and 2). There is a visible but not statistically significant difference between the MAX of the cGANs-based method and the baseline (p > 0.05) (Table 3.3. 36 IU ears, row 1 and 2). There is a visible but not statistically significant difference between the results of the 2D- and the 3D-cGANs (p > 0.05) (Table 3.3. 36 IU ears, row 3).

Figures 3.5c and 3.5d show the boxplots of the statistics of the 78 IB ears and the 34 IU ears that have been scanned with the lCT scanner postoperatively and the cCT scanners preoperatively. Table 3.4 shows the results of the Wilcoxon signed-rank tests. These show the same trend as Figures 3.5a and 3.5b and Table 3.3.

Figures 3.5e and 3.5f show the boxplots of the statistics of the 10 IB ears and the 2 IU ears that have been scanned with the cCT scanners postoperatively and preoperatively. At the time of writing, we are not able to draw strong conclusions from these two plots because we only have a very limited number of such images but the trends are similar to those obtained with the other datasets.

Testing ears	Approaches to compare	MAX		MED		MEA		STD		MIN	
		Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
88 IB ears	Post-CT + SegPost-BL 2D-cGANs + SegPre-ASM	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	Post-CT + SegPost-BL 3D-cGANs + SegPre-ASM	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	2D-cGANs + SegPre-ASM 3D-cGANs + SegPre-ASM	0.117	–	0.174	–	0.179	–	0.628	–	1.483	–
36 IU ears	Post-CT + SegPost-UL 2D-cGANs + SegPre-ASM	0.136	**–**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.371	–	**0.002**	**< 0.001**
	Post-CT + SegPost-UL 3D-cGANs + SegPre-ASM	0.115	–	**0.002**	**0.001**	**0.002**	**0.001**	0.221	–	0.005	–
	2D-cGANs + SegPre-ASM 3D-cGANs + SegPre-ASM	0.729	–	0.825	–	0.949	–	0.937	–	1.482	–

Table 3.3: P-values of the two-sided and one-sided Wilcoxon signed-rank tests of the five statistics for the P2PEs of the 88 IB ears and the 36 IU ears. Note: Bold indicates cases that are significantly different (p-value less than 0.05). The p-values have been corrected using the Holm-Bonferroni method.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.5: Boxplot of P2PEs for (a) the 88 IB ears, (b) the 36 IU ears, (c) the 78 IB ears scanned by the lCT scanner postoperatively and the cCT scanners preoperatively, (d) the 34 IU ears scanned by the lCT scanner postoperatively

and the cCT scanners preoperatively, (e) the 10 IB ears scanned by the cCT scanners postoperatively and preoperatively, (f) the 2 IU ears scanned by the cCT scanners postoperatively and preoperatively.

Testing ears	Approaches to compare	MAX		MED		MEA		STD		MIN	
		Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
78 IB ears	Post-CT + SegPost-BL 2D-cGANs + SegPre-ASM	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	Post-CT + SegPost-BL 3D-cGANs + SegPre-ASM	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	2D-cGANs + SegPre-ASM 3D-cGANs + SegPre-ASM	0.136	–	0.163	–	0.172	–	0.770	–	1.110	–
34 IU ears	Post-CT + SegPost-UL 2D-cGANs + SegPre-ASM	0.080	–	**0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**0.285**	–	**0.002**	**< 0.001**
	Post-CT + SegPost-UL 3D-cGANs + SegPre-ASM	0.075	–	**0.002**	**0.001**	**0.002**	**< 0.001**	**0.285**	–	**0.006**	–
	2D-cGANs + SegPre-ASM 3D-cGANs + SegPre-ASM	0.993	–	0.590	–	0.675	–	0.857	–	1.110	–

Table 3.4: P-values of the two-sided and one-sided Wilcoxon signed-rank tests of the five statistics of the P2PEs for the 78 IB ears and the 34 IU ears that have been scanned with the lCT scanner postoperatively and the cCT scanners preoperatively. Note: Bold indicates cases that are significantly different (p-value less than 0.05). The p-values have been corrected using the Holm-Bonferroni method.

Figure 3.6 shows the boxplots of the statistics for P2PEs of the 124 testing ears processed by the 3D-cGANs that are trained using L1 and WL1. Visually, the medians of the MAX, MED, and MEA values obtained with WL1 (yellow bars) are lower than those obtained with L1 (black bars). Wilcoxon signed-rank tests reported in Table 3.5 show that these differences are significant for the MAX, MED, MEA, and STD ($p < 0.05$).

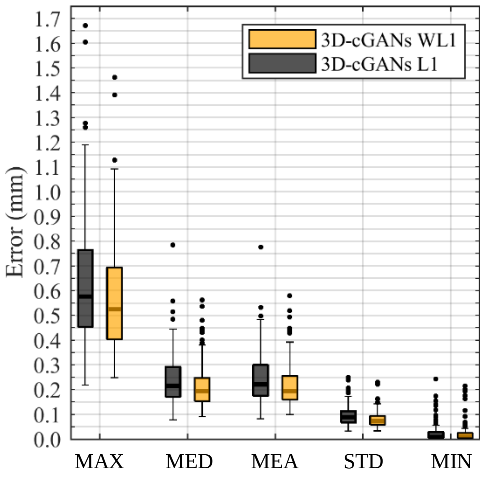


Figure 3.6: Boxplot of P2PEs for the 124 testing ears. “3D-cGANs WL1” and “3D-cGANs L1” denote the results obtained with the 3D-cGANs which are trained using the weighted L1 loss and original L1 loss, respectively.

Reconstruction loss to compare	MAX		MED		MEA		STD		MIN	
	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
WL1 L1	**0.004**	**0.002**	**< 0.001**	**< 0.001**	**0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**0.945**	–

Table 3.5: P-values of the two-sided and one-sided Wilcoxon signed-rank tests of the five statistics for the P2PEs of the 124 testing ears. Note: Bold indicates cases that are significantly different (p-value less than 0.05). The p-values have been corrected using the Holm-Bonferroni method.

3.4.2 Mean Structural Similarity

Figure 3.7 shows boxplots of the MSSIM for the 124 testing ears. Wilcoxon signed-rank tests show that all of the cGANs-based methods achieve statistically significant higher MSSIM compared to the baseline ($p < 0.05$). Table 3.6 shows the p-values of the Wilcoxon signed-rank tests between the results of the 2D-cGANs and the 3D-cGANs that are trained using a different reconstruction loss. The 3D strategies achieve statistically significantly higher MSSIM compared to the 2D approach ($p < 0.05$). The 3D-cGANs trained with the weighted L1 loss produce a significantly higher MSSIM than those trained with the non-weighted L1 loss ($p < 0.05$). We also observe that the 3D-cGANs reach the optimal epoch at the 15-th training epoch when the weighted L1 loss is applied. However, they need 70 training epochs to reach the optimal epoch when the non-weighted L1 loss is applied. This suggests that using weights can accelerate the optimization of the networks.

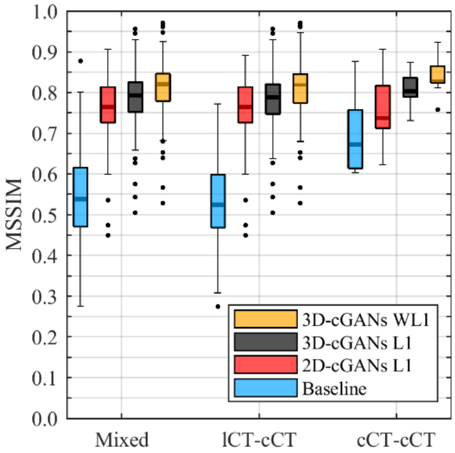


Figure 3.7: Shown on the left, middle, and right are boxplots of the MSSIM for the 124 testing ears (Mixed), the 112 ears scanned by the lCT scanner postoperatively and the cCT scanners preoperatively (lCT-cCT), and the 12 ears scanned by the cCT scanners postoperatively and preoperatively (cCT-cCT). “Baseline” denotes the MSSIM between the Post-CT images and the true Pre-CT images; “2D-cGANs L1” denotes the results produced by our previous 2D-cGANs trained with the pure L1 loss; “3D-cGANs L1” and “3D-cGANs WL1” denote the results produced by the 3D-cGANs which are trained using the pure L1 loss and the weighted L1 loss, respectively.

Approaches to compare	Mixed (124 ears)		lCT-cCT (112 ears)		cCT-cCT (12 ears)	
	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
2D-cGANs L1 3D-cGANs L1	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
2D-cGANs L1 3D-cGANs WL1	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
3D-cGANs L1 3D-cGANs WL1	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**

Table 3.6: The p-values of the two-sided and one-sided Wilcoxon signed-rank tests between the MSSIM of the true Pre-CT images and the synthetic images produced by the 2D-cGANs and the 3D-cGANs trained using different reconstruction losses. “lCT-cCT” denotes that the ear has been scanned by the 1CT scanner postoperatively and a cCT scanner preoperatively, and “cCT-cCT” denotes that the ear has been scanned by a cCT scanner postoperatively and preoperatively. Note: Bold indicates cases that are significantly different (p-value less than 0.05). The p-values have been corrected using the Holm-Bonferroni method.

3.5 Discussion and Conclusions

As discussed in the recent review article by Yi et al. (Yi et al., 2019), GANs have been extensively used to solve medical imaging related problems such as classification, detection, image synthesis, low dose CT denoising, reconstruction, registration, and segmentation. However, at the time of writing and to the best of our knowledge, GANs have not been proposed to eliminate or reduce metallic artifacts in CT images. There is also a large body of work aiming at reducing artifacts in CT images (Gjesteby et al., 2016). But, compared to the current leading methods, which generally necessitate the raw data from CT scanners, our approach is a post-reconstruction processing method for which the raw data is not required. Compared to other published machine learning-based methods proposed for the removal of metallic artifacts that either depend on existing traditional methods or require post-processing of the outputs produced by machine learning models (Gjesteby et al., 2017; Park et al., 2017; Yanbo Zhang & Yu, 2018), ours is unique in being able to synthesize directly an artifact-free image from an image in which artifacts are present. Although we have not investigated it yet, we hypothesize that our method could be applied to other types of images affected by the same type of artifacts if sufficient training images consisting of pairs of images with and without artifacts were available. We also hypothesize that if the problems are similar enough transfer learning could be used to reduce the size of the dataset needed for training.

The results we have generated show that the quality of the images produced by the 3D networks is better than that of the images produced by the 2D networks when the MSSIM is used to compare them. This is confirmed by the visual appearance of the synthetic images produces by these two architectures as shown in Figure 3.4. There is also a small but not statistically significant difference in the segmentation results produced with the images generated with the 3D and the 2D networks; this difference is especially noticeable for the maximum error. The fact that the segmentation results improve only modestly when the quality of the images improves more substantially suggests that the constraints imposed by the active shape model can compensate for imperfections in the synthetic images. Likely, segmentation methods that do not impose strong constraints on the shape of the intracochlear anatomy would be more sensitive to those errors. As discussed earlier, we also note that the technique we have developed to assist audiologists in programing the implant depends on the position of the contacts with respect to the anatomy (Noble et al., 2013). Any improvement in segmentation accuracy, even small, may have a positive impact on the programming recommendations we provide to the audiologists. Finally, the methods we have developed to segment the anatomy, localize the contacts, and provide programming guidance have been integrated into an interactive software package that has been deployed to the clinic and is in routine use at our institution. Without further optimization of our current implementation of the cGANs, the speed of execution for the 3D version is 1.5s when it is 60s for the 2D version, which is important for the integration of our methods into the clinical workflow. Overall, the study we have conducted shows that cGANs are an effective way to eliminate metallic artifacts in CT images and that the 3D version of our proposed method should be preferred over the 2D version.

Acknowledgments

This study has been supported in parts by NIH grants R01DC014037 and R01DC014462 and by the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN. The content is solely the responsibility of the authors and does not necessarily represent the official views of these institutes. We are grateful for the support of the NVIDIA Corporation for providing us with a GPU used to support this work. We thank

the anonymous reviewers of Medical Image Analysis for their constructive comments, their feedback has greatly improved the quality of this chapter.

CHAPTER IV

Validation of a Metal Artifact Reduction Method Based on 3D Conditional GANs for CT Images of the Ear

Jianing Wang[1], Srijata Chakravorti[1], Yiyuan Zhao[2], Jack H. Noble[1], and Benoit M. Dawant[1]

[1]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, 37232, USA

[2]Siemens Medical Solutions USA Inc, Malvern, PA USA 19355

This chapter is adapted from "Validation of a metal artifact reduction method based on 3D conditional GANs for CT images of the ear" published in Proc. SPIE 11315, Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling, 1131507.

The published paper can be found at https://doi.org/10.1117/12.2549398

It has been reproduced with the permission of the publisher.

Abstract

We have proposed an approach that uses conditional Generative Adversarial Nets (cGANs) to synthesize Pre-implantation CT (Pre-CT) images from Post-implantation CT (Post-CT) images of cochlear implant recipients in **Chapter III**. This method permits to use of algorithms designed to segment Pre-CT images even when these are unavailable, thus provides an approach to obtain the segmentation of the intracochlear anatomy of the cochlear implant recipients for whom the Pre-CT images are unavailable. Our group has developed Image-Guided Cochlear Implant Programming (IGCIP) techniques to assist audiologists with the programming of the implanted electrodes. Given the segmentation of the intracochlear anatomy and the location information of the implanted electrodes, our IGCIP techniques can produce the configuration of the cochlear implant, which is a set of activation states of the electrodes. In this chapter, we evaluate the effect of the method proposed in **Chapter III** on the final output of our IGCIP techniques by assessing the configurations of the cochlear implant generated using the synthetic Pre-CT images. We also modify the training loss functions used in **Chapter III** to further improve the performance of our networks. Results show that, for cGANs trained using the loss functions proposed in this chapter, 85.1% of the configurations generated using the synthetic images can be used for the programming of the cochlear implants, and these configurations are likely to lead to hearing outcomes that are comparable to those achieved using the best possible configurations. The rate is 83.9% for the cGANs trained using the loss functions proposed in **Chapter III**. A configuration, for which the states of more than one electrode needed to be adjusted to make it feasible for the programming, is considered to be a poor case. The rate of poor cases is 2.3% for the cGANs trained using the loss functions proposed in this chapter, and the number is 8.0% for the cGANs trained using the loss functions proposed in **Chapter III**. Overall, the success rate we have achieved is good, which confirms that our techniques could greatly benefit patients for whom Pre-CT images are not available. Using the training loss functions proposed in **Chapter III** and this chapter can achieve a similar success rate, however, the new loss functions can lead to a lower rate of poor cases.

4.1 Introduction

Cochlear implants are surgically implanted neural prosthetic devices that are used to treat severe-to-profound hearing loss and are programmed postoperatively by audiologists to optimize outcomes. The efficacy of the programming is sensitive to the spatial relationship between the implanted electrodes and the intracochlear anatomy. Providing accurate information about the position of the electrodes with respect to the intracochlear structures can thus help audiologists to fine-tune and customize the programming. Our group has developed Image-Guided Cochlear Implant Programming (IGCIP) methods that use image analysis techniques to determine the position of the implanted electrodes relative to the intracochlear anatomy to assist audiologists with programming by selecting a subset of active electrodes (Noble et al., 2013). Our IGCIP techniques require the segmentation of the scala tympani, the scala vestibuli, and the modiolus in the CT images of the recipients. Noble et al. (Noble et al., 2011) have developed an active shape model-based method, which we refer to as SegPre-ASM, to segment the intracochlear anatomy in Pre-implantation CT (Pre-CT) images. However, SegPre-ASM cannot be directly applied to Post-implantation CT (Post-CT) images due to the strong artifacts produced by the electrodes. Reda et al. have proposed two methods, which we refer to as SegPost-UL (Reda, McRackan, et al., 2014) and SegPost-BL (Reda, Noble, et al., 2014), to segment the intracochlear anatomy in the Post-CT images of recipients who have been implanted unilaterally and of recipients who have been implanted bilaterally, but these methods could lead to segmentation errors that are substantially larger than errors obtained with Pre-CT images. We have proposed an approach in **Chapter III** of this dissertation, that first, we use conditional Generative Adversarial Nets (cGANs) to generate an artifact-free Pre-CT image from a given Post-CT image, then we apply SegPre-ASM to the resulting synthetic Pre-CT image to get the segmentation of the intracochlear anatomy (J. Wang et al., 2018, 2019). We have shown that this approach substantially and significantly reduces the segmentation errors obtained with SegPost-UL and SegPost-BL.

As introduced in **Chapter I**, our IGCIP system can generate the configuration of the electrodes automatically by using the knowledge of the relative positions of each electrode to the spiral ganglion and their neighboring electrodes. We refer to our current automatic method (Zhao et al., 2016) for generating the configuration of the electrodes as “AutoConfig” for simplicity. AutoConfig selects the state of each electrode by optimizing a cost function that is based on the electrode Distance-Vs.-Frequency (DVF) curves, which are a set of 2D plots that captures the patient-specific spatial relationship between the electrodes and the spiral ganglion nerves (Noble et al., 2013). Figure 4.1 is an example of DVF curves for a 12-electrode array. Each DVF curve corresponds to one electrode and is labeled with its corresponding electrode number. For each point on a DVF curve, the horizontal value represents the positions along the length of the modiolus in terms of the characteristic frequencies of the spiral ganglion cells, and the vertical value indicates the distance from the electrode to the modiolar surface. The solid curves and the dashed curves denote that the corresponding electrodes are selected to be active and inactive by AutoConfig, respectively.

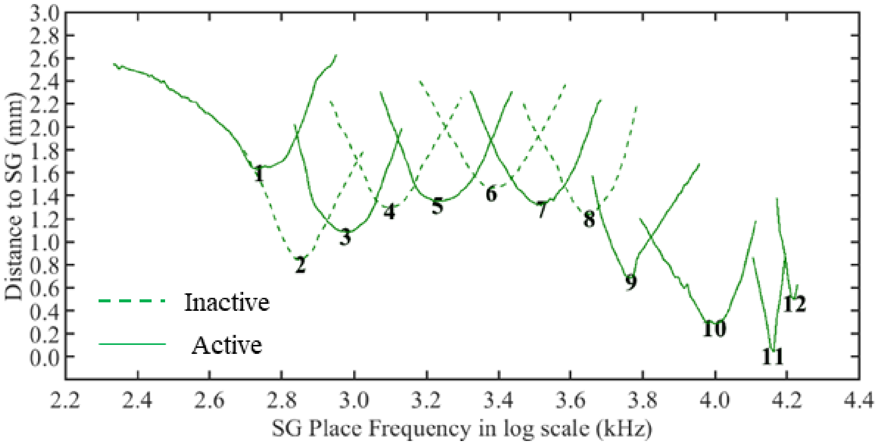


Figure 4.1: An example of DVF curves for a 12-electrode array. Each DVF curve is labeled with its corresponding electrode number. (SG – spiral ganglion)

The construction of the DVF curves depends on the location knowledge of the implanted electrodes that can be obtained by applying an automatic method (Zhao et al., 2019) to the Post-CT images, and the segmentation of the intracochlear anatomy that can be acquired by applying SegPre-ASM to the Pre-CT images when they are available or by applying SegPre-ASM to the synthetic Pre-CT images generated by our cGANs when only the Post-CT images are available. Thus, in this chapter, we evaluate the effect of our image synthesis method proposed in **Chapter III** on the final output of our IGCIP techniques by comparing configurations of the electrodes obtained using the real and the synthetic Pre-CT images. We also explore ways to improve the performance of our networks by modifying the training loss functions used in **Chapter III**.

4.2 Materials and Methods

4.2.1 Dataset

Our dataset consists of Pre-CT and Post-CT image pairs of 208 ears. All of these images have been acquired with the recipients of the cochlear implants in roughly the same position. The 208 ears are randomly partitioned into a set of 116 ears for training, a set of 5 ears for validation, and a set of 87 for testing. No images from the same patient are used for both training and testing. The Pre-CT images are registered to the Post-CT images using intensity-based affine registration techniques. As we have done in **Chapter III**, each image is clamped to the range of 0.1 to 99.9-th percentiles of its intensity values. Then the intensity values of the images are rescaled to the -1 to 1 range. We apply image augmentation to the training and validation set by rotating each image by 20 small random angles in the range of -30 and 30 degrees about the x-, y-, and z-axis, such that 60 additional images are created from each original image. This results in a training set that is expanded to 7076 volumes, and a validation set that is expanded to 310 volumes.

4.2.2 Network Architecture

As described in **Chapter III**, our cGANs (Figure 4.2) are a 3D expansion of the 2D framework proposed by Isola et al. (Isola et al., 2017). The input of the generator G is a Post-CT image $post$ of size 96×96×96 (Figure 4.3a). The output of G is a synthetic Pre-CT image $G(post)$ of the same size. The discriminator D maps the input, which is the concatenation of $post$ and its corresponding real Pre-CT image pre (Figure 4.3b) (or the concatenation of $post$ and $G(post)$), to a scalar that can be interpreted as the probability of the input Pre-CT image to be synthetic rather than real.

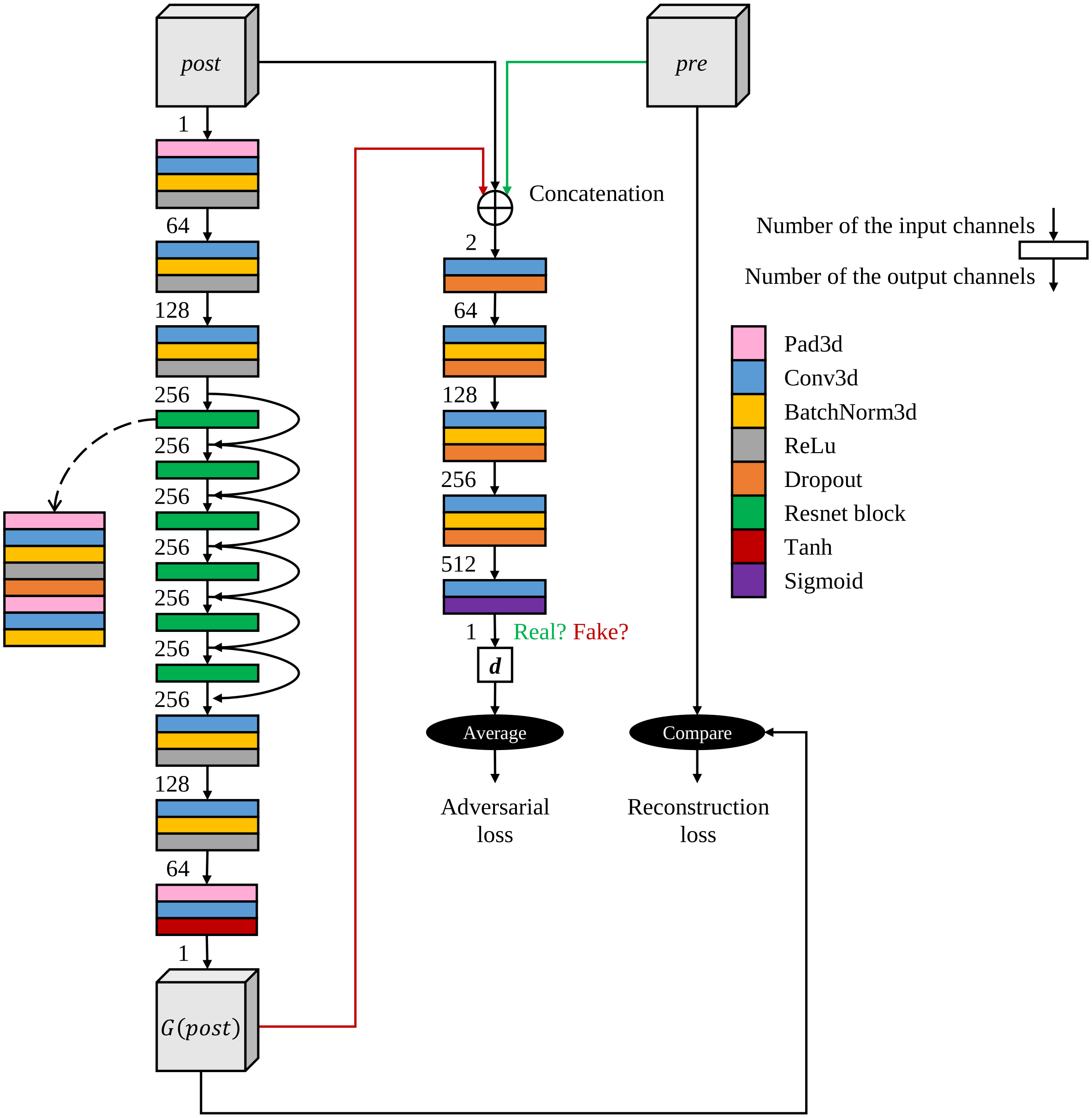


Figure 4.2: An illustration of the architecture of the 3D-cGANs.

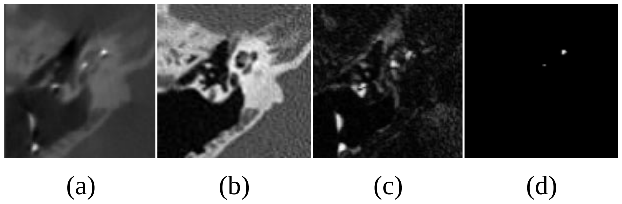


Figure 4.3: An example of (a) Post-CT and (b) Pre-CT patches. (c) The dissimilarity map of (a) and (b). (d) The regions affected by the implanted electrodes in the Post-CT are shown in white.

4.2.3 Training Objectives

We use the combination of an adversarial loss term and a weighted L1 loss term to train the cGANs in **Chapter III**, here we modify the previous training loss by adding a first loss term that emphasizes the difference between the image content of the Pre-CT image and the Post-image, and a second loss term that penalizes specific regions in the synthetic Pre-CT images. Our current training loss is the weighted sum of the 4 terms listed below.

4.2.3.1 Adversarial Loss

As defined in **Chapter III**, the adversarial loss $L_{\mathrm{cGAN}}(G, D)$ can be expressed as:

$$L_{\mathrm{cGAN}}(G, D) = \min_{G} \max_{D} \mathbb{E}_{post,pre}\left[\log\left(D(post, pre)\right)\right] + \mathbb{E}_{post}\left[\log\left(1 - D\left(post, G(post)\right)\right)\right] \quad (4.1)$$

4.2.3.2 L1 Loss Weighted by a Simple Weighting Matrix

As done in **Chapter III**, we use an L1 loss term that is weighted by a simple weighting matrix. The weighted L1 loss $L_{\mathrm{Box}}(G)$ is defined as:

$$L_{\mathrm{Box}}(G) = \mathbb{E}_{post,pre}\left[\left\|W_{\mathrm{Box}} \circ \left(pre - G(post)\right)\right\|_1\right] \quad (4.2)$$

in which W_{box} is the weighting matrix and $\circ$ is the element-wise multiplication operation. To calculate W_{box}, we first create a bounding box of size 56×56×56 that encloses the cochlea. With the number of voxels inside the bounding box equal to N_{in} and the number of voxels outside of the bounding box equal to N_{out}, we assign weights to the voxels inside and outside of the bounding box to $(N_{in} + N_{out})/N_{in}$ and 1, respectively.

4.2.3.3 L1 Loss Weighted by a Dissimilarity Map

We use an L1 loss weighted by a dissimilarity map between the Pre-CT image and the Post-CT image to lay stress on the difference between the image content of the Pre-CT and Post-CT image pair. This weighted L1 loss $L_{\mathrm{Dissim}}(G)$ can be expressed as:

$$L_{\mathrm{Dissim}}(G) = \mathbb{E}_{post,pre}\left[\left\|W_{\mathrm{Dissim}} \circ \left(pre - G(post)\right)\right\|_1\right] \tag{4.3}$$

wherein W_{Dissim} is the dissimilarity map between the Pre-CT image and the Post-CT image (Figure 4.3c). To calculate W_{Dissim}, first, we adjust the intensity values of the Post-CT image by matching its histogram to that of its paired Pre-CT image. W_{Dissim} is then calculated by subtracting the Pre-CT image from the intensity-adjusted Post-CT image and rescaling the resulting image to the 0 to 1 range. Higher values in W_{Dissim} indicates a higher degree of disparity between the Pre-CT image and the Post-CT image, such that more attention should be given to those areas.

4.2.3.4 Region-specific Loss

In our specific case, areas that are affected by the implanted electrodes in the Post-CT image should be more likely to be mapped to the intensity values of the "intracochlear fluid" in the Pre-CT image. Therefore we use a region-specific loss term $L_{\mathrm{Region}}(G)$ to incorporate this prior knowledge into our training objectives. $L_{\mathrm{Region}}(G)$ is calculated as:

$$L_{\mathrm{Region}}(G) = \mathbb{E}_{post,pre}[\|G(post)_{\mathrm{e}} - v_{\mathrm{f}}\|_2] \tag{4.4}$$

wherein $G(post)_{\mathrm{e}}$ denotes voxels in the synthetic Pre-CT image that are mapped from the areas affected by the implanted electrodes in the Post-CT image. These voxels can be roughly identified by thresholding the Post-CT image with the 99.95-th percentile of its intensity values (Figure 4.3d). v_{f} is the intensity value of the intracochlear fluid, which is set to -1 for our experiments.

4.2.3.5 Total Loss

The total training loss used in **Chapter III** can be expressed as:

$$\arg\min_G \max_D L_{\mathrm{cGAN}}(G,D) + \alpha_{\mathrm{III}} L_{\mathrm{Box}}(G) \tag{4.5}$$

and our modified training loss can be expressed as:

$$\arg\min_G \max_D L_{\mathrm{cGAN}}(G,D) + \alpha_{\mathrm{IV}} L_{\mathrm{Box}}(G) + \beta L_{\mathrm{Dissim}}(G) + \gamma L_{\mathrm{Intens}}(G) \tag{4.6}$$

wherein α_{III}, α_{IV}, β, and γ are the weighting factors of each loss term, which are set to 100, 60, 30, and 10 empirically.

4.2.4 Evaluation

We train two models using the training loss which is used in **Chapter III** (Equation 4.5) and our modified loss (Equation 4.6). The training has been performed as done in **Chapter III**. For simplicity, we denote the cGANs trained with the modified loss and the cGANs trained with the previous loss as “cGANs-proposed” and “cGANs-previous”, respectively.

cGANs-proposed and cGANs-previous are evaluated using the 87 testing cases, the quality of the synthetic Pre-CT images is evaluated by (1) comparing the segmentation results of the intracochlear anatomy obtained with SegPre-ASM applied to the synthetic Pre-CT images and those obtained using the real Pre-CT images, and (2) evaluating the configurations of the implanted electrodes obtained with our IGCIP system applied to the synthetic Pre-CT images.

4.2.4.1 Comparison of the Segmentation of the Intracochlear Anatomy Obtained with the Real and the Synthetic Pre-CT Images

As described in **Chapter III**, the output of SegPre-ASM is surface meshes of the intracochlear anatomy that have 24423 pre-defined vertices. Point-to-Point Errors (P2PEs), which are computed as the Euclidean distance in millimeter, between the corresponding vertices on the meshes generated from the real Pre-CT images and the meshes generated from the synthetic Pre-CT images are calculated to quantify the quality of the synthetic images. For each testing ear, we calculate the P2PEs of the 24423 vertices, and we calculate the MAXimum (MAX), MEAn (MEA), MEDian (MED), STandard Deviation (STD), and MINimum (MIN) of the P2PEs.

4.2.4.2 Comparison of the Configuration of Electrodes Obtained with the Real and the Synthetic Pre-CT Images

We compare the configurations of the electrodes obtained using the real Pre-CT images with those obtained with the synthetic Pre-CT images. The configuration of an electrode array can be represented by a state vector of 1s and 0s, in which 1 denotes an electrode is active and 0 denotes an electrode is deactivated. We use the Hamming distance (Hamming, 1950) between two state vectors to capture the disagreement between the two configurations. A Hamming distance of zero indicates that a synthetic Pre-CT image is identical to the real Pre-CT image for our application.

Comparing the case for which the Hamming distance is not zero is more difficult because different configurations may lead to essentially the same neural activation patterns. Therefore we conduct an expert evaluation study to assess these cases visually. As shown in Figure 4.1, the configuration of a cochlear implant is displayed as the “on” and “off” states of the electrodes on the corresponding DVF curves. The DVF curves were generated with the real Pre-CT image of the recipient, whereas the configuration can be generated with the real Pre-CT image, or generated with a synthetic Pre-CT image that is generated from the Post-CT image of the recipient, or a control plan. A control plan of an implanted electrode array is a suboptimal configuration that seems workable but actually cannot be used for the programming of the electrodes, it has been generated by an expert independently by manually changing the states of few electrodes in a configuration that has been generated using the real Pre-CT image. The control plans are added to the study to avoid evaluation bias.

Our expert evaluation study is conducted as follows: Each time a configuration is displayed to an expert (not the one who made the control plans), the expert is asked to rate whether the configuration is good, fair, or poor. A good case is a configuration that can be used for the programming of the cochlear implant, and this configuration is likely to lead to hearing outcomes that are comparable to those achieved using the best possible configuration. A fair case is an inferior configuration that can turn to be “good” by changing the state of only one electrode. A poor case is a configuration that cannot be used for the programming unless the states of more than one electrode are adjusted. The orders of the configurations are randomized during the visual assessment so that the expert who reviews these DVF curves is blind to the origin of the configurations.

4.3 Results

4.3.1 Point-to-point Errors

Figure 4.4 shows the boxplots of the MAX, MEA, MED, STD, and MIN of the 87 testing ears, wherein “Previous” and “Proposed” denote the results of cGANs-previous and cGANs-proposed, respectively. The medians of MAX of cGANs-previous and cGANs-proposed are 0.564mm and 0.556mm. The medians of MED of cGANs-previous and cGANs-proposed are 0.207mm and 0.199mm. The medians of STD of cGANs-previous and cGANs-proposed are

0.080mm and 0.077mm. cGANs-proposed achieves a notable interquartile range of MAX. However, the difference between the results of cGANs-previous and cGANs-proposed are not statistically significant. Overall the two approaches achieve similar performance in terms of P2PEs.

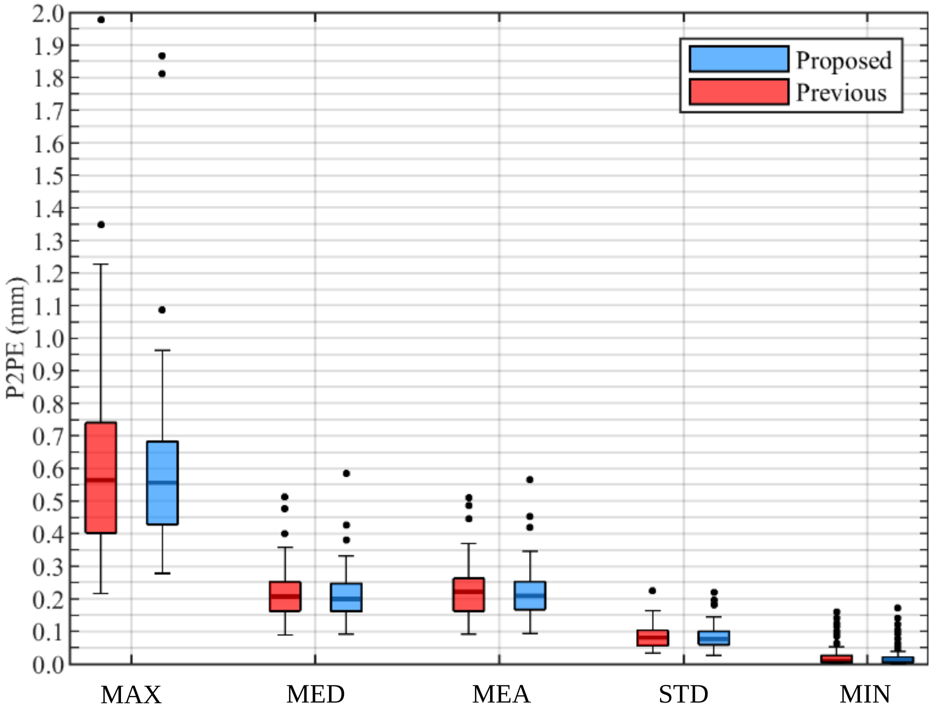


Figure 4.4: Boxplots of the MAX, MEA, MED, STD, and MIN of the 87 testing ears.

4.3.2 Hamming Distances

Figure 4.5 shows the histogram of the Hamming distances between the configurations generated using the synthetic Pre-CT images and those generated using the corresponding real Pre-CT images. The numbers of cases of zero Hamming distance for cGANs-previous and cGANs-proposed are 18 and 22. We also observed that there is no strong linear correlation between the mean P2PE and the Hamming distance (Pearson correlation coefficient ~0.4, p-value < 0.005). For example, Figure 4.6 shows 3 cases with average (Figure 4.6a) or small (Figure 4.6b, Figure 4.6c) mean P2PE values but for which the Hamming distance varies from 5 to 11.

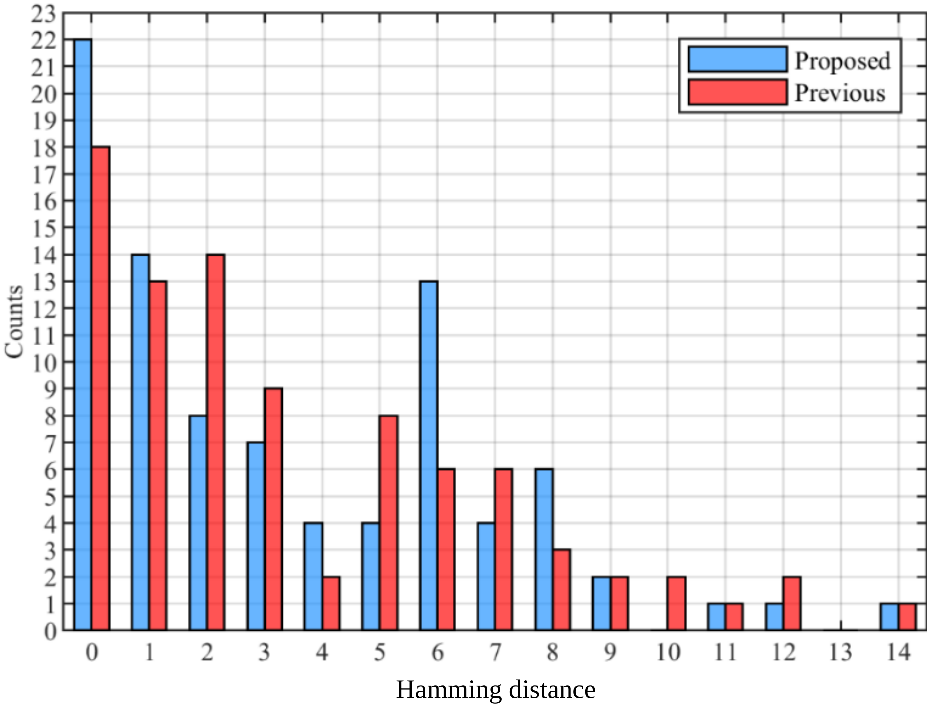


Figure 4.5: Histogram of the Hamming distances between the configurations generated using the synthetic Pre-CT images and those generated with the corresponding real Pre-CT images.

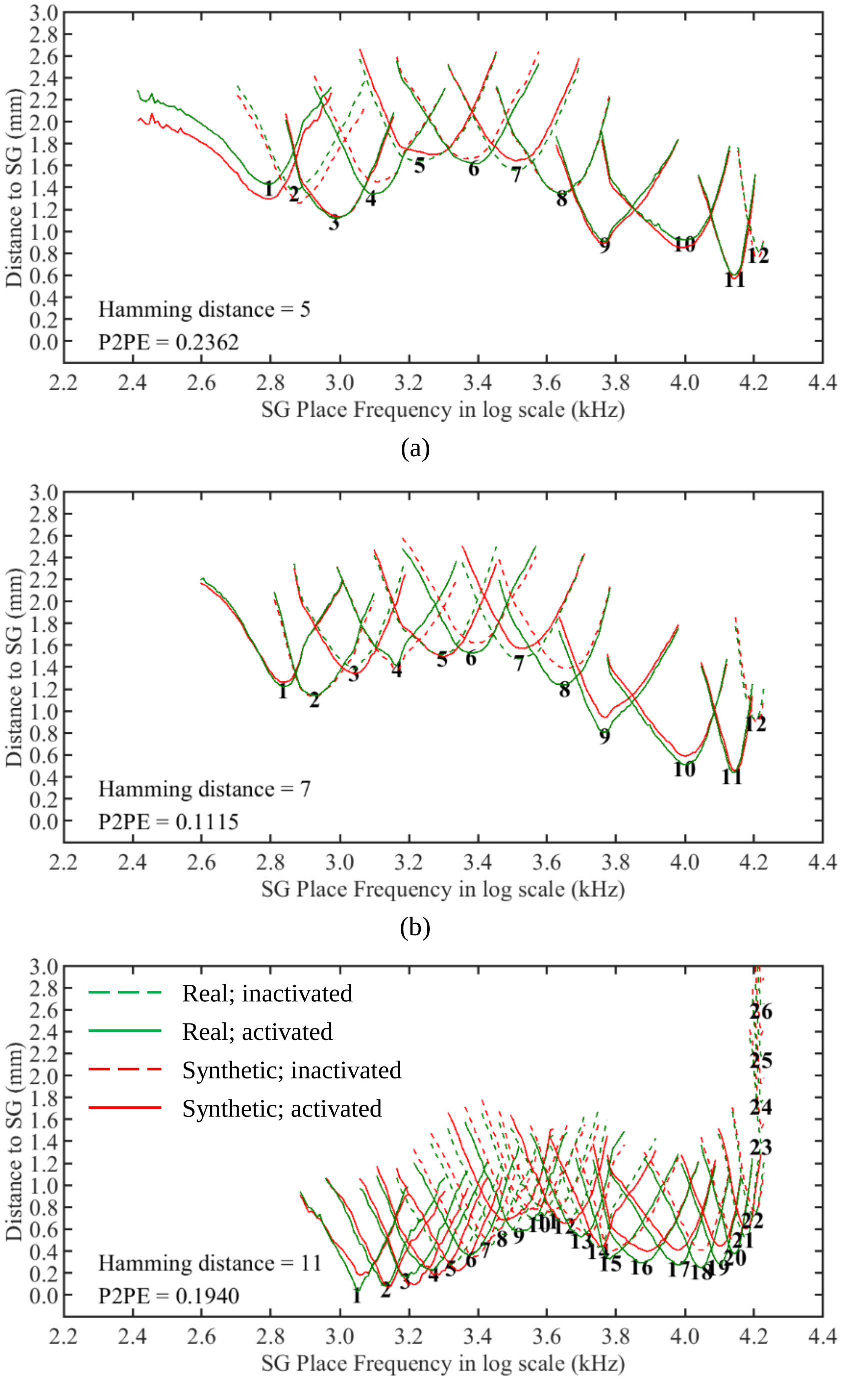


Figure 4.6: (a) A testing case for which the mean point-to-point segmentation error is 0.2362mm and the Hamming distance is 5. (b) A testing case for which the mean point-to-point segmentation error is 0.1115mm and the Hamming distance is 7. (c) A testing case for which the mean point-to-point segmentation error is 0.1940mm and the Hamming distance is 11. (P2PE – the mean point-to-point segmentation error in millimeter; SG – spiral ganglion)

4.3.3 Expert Evaluation Study

The results of our visual evaluation are summarized in the fourth row of Table 4.1. There are 69 cases with nonzero Hamming distance for cGANs-previous, and 65 cases of nonzero Hamming distance for cGANs-proposed, together, there are 72 unique cases that require visual validation. The numbers of good cases, fair cases, and poor cases for cGANs-previous are 55, 7, and 7. For cGANs-proposed, these numbers are 52, 11, and 2. For the control plans, the number of good cases is only 2, and the number of fair and poor cases is 70. The ratio of good cases to the visually validated cases is around 80% for both cGANs-previous (55 of 69) and cGANs-proposed (52 of 65). The ratio is only 2.8% for the control plans (2 of 72), which confirms that there is no evaluation bias in our visual study.

<table>
<tr><th></th><th colspan="2">cGANs-previous</th><th colspan="2">cGANs-proposed</th><th colspan="2">Control plan</th></tr>
<tr><td>Total testing cases</td><td colspan="2">87</td><td colspan="2">87</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Cases of zero Hamming distance</td><td colspan="2"><b>18</b></td><td colspan="2"><b>22</b></td></tr>
<tr><td rowspan="3">Cases that require visual validation</td><td rowspan="3">69</td><td>Good <b>55</b> (79.7%)</td><td rowspan="3">65</td><td>Good <b>52</b> (80%)</td><td rowspan="3">72</td><td>Good 2 (2.8%)</td></tr>
<tr><td>Fair 7</td><td>Fair 11</td><td rowspan="2">Fair/Poor 70</td></tr>
<tr><td>Poor 7</td><td>Poor 2</td></tr>
<tr><td>Total successful cases</td><td colspan="2"><b>73</b> (83.9%)</td><td colspan="2"><b>74</b> (85.1%)</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Total poor cases</td><td colspan="2">7 (8.0%)</td><td colspan="2">2 (2.3%)</td></tr>
</table>

Table 4.1: Number of the cases in each category in our evaluation study.

4.4 Conclusions

The cases with zero Hamming distance and the cases that are considered to be "good" by the expert in the visual evaluation study are together categorized as "successful" cases. As shown in the fifth row of Table 4.1, for cGANs-previous, the number of successful cases is 73, and the ratio of the successful cases to the total testing cases is 83.9%. For cGANs-proposed, the number of successful cases is 74, and the ratio is 85.1%. As shown in the sixth row of Table 4.1, the ratio of the poor cases to the total testing cases for cGANs-previous is 8.0%, and the ratio is 2.3% for cGANs-proposed. Note that, usually changing the state of only one electrode won't affect much on the final hearing outcomes of the cochlear implant recipients, so the "fair" cases could actually lead to good hearing outcomes.

However, this would require audiology tests to confirm their eligibility. Overall, cGANs-proposed and cGANs-previous have achieved a similar success rate, but cGANs-proposed has a lower rate of poor cases. This suggests that the new training loss proposed in this chapter can improve the performance of our cGANs in terms of the final output of our IGCIP system, although they cannot significantly reduce the segmentation errors when SegPre-ASM is applied to the synthetic CT images.

Acknowledgments

This work has been supported by NIH grants R01DC014037 and R01DC014462 and by the Advanced Computing Center for Research and Education (ACCRE) of Vanderbilt University. The content is solely the responsibility of the authors and does not necessarily represent the official views of these institutes. We are grateful for the support of the NVIDIA Corporation for providing us with a GPU used to support this work. The authors declare no conflict of interest.

CHAPTER V

Metal Artifact Reduction and Intracochlear Anatomy Segmentation in CT Images of the Ear with a Multi-Resolution Multi-task 3D Network

Jianing Wang, Jack H. Noble, and Benoit M. Dawant

Department of Electrical and Computer Engineering

Vanderbilt University

Nashville, TN, 37232, USA

This chapter is adapted from “Metal artifact reduction and intra cochlear anatomy segmentation in CT images of the ear with a multi-resolution multi-task 3D network” published in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI).

The published paper can be found at https://doi.org/10.1109/ISBI45749.2020.9098707

It has been reproduced with the permission of the publisher.

Abstract

In this chapter, we propose a semantic segmentation approach to segment the intracochlear anatomy in Post-implantation CT (Post-CT) images of the cochlear implant recipients. Compared to the two-step method proposed in **Chapter III**, the method proposed in this chapter can produce the voxel-wise segmentation labels of the intracochlear anatomy in one step. Our task is challenging due to the strong artifacts produced by the metallic electrodes, therefore we use a multi-task learning approach. We use a multi-resolution multi-task network, which consists of a shared feature extractor, an image synthesis branch, and an image segmentation branch, to synthesize Pre-CT images from Post-CT images and generate the segmentation mask of the intracochlear anatomy in the Post-CT images simultaneously. The task of image segmentation is our main task, and the task of image synthesis is used as an auxiliary task to assist the training of the main task. The output size of the image synthesis branch is 1/64 of that of the segmentation branch. This limits the memory usage for training while generating high-resolution segmentation labels. We use the segmentation results of an automatic method as the ground truth to provide supervision to train our model, and we achieve a median Dice index value of 0.792. Our experiments also confirm the usefulness of multi-task learning.

5.1 Introduction

Cochlear implants are surgically implanted neural prosthetic devices for treating severe-to-profound hearing loss. Cochlear implants are programmed postoperatively by audiologists who adjust a processor’s settings to send the appropriate signals to each of the implant’s electrodes. The efficacy of the programming is sensitive to the spatial relationship between the electrodes and the intracochlear anatomy. Accurately localizing the electrodes relative to the intracochlear anatomy in Post-implantation CT (Post-CT) images can help audiologists to fine-tune and customize the programming of the cochlear implants. This requires the accurate segmentation of the intracochlear anatomy including the scala tympani, the scala vestibuli, and the modiolus in these images. Segmenting the intracochlear anatomy in the Post-CT images is challenging due to the strong artifacts produced by the metallic electrodes. Over the years, we have developed algorithms that permit segmenting the intracochlear anatomy in the Post-CT images directly (Reda, McRackan, et al., 2014; Reda, Noble, et al., 2014) and indirectly (J. Wang et al., 2018, 2019). The direct methods, which segment the intracochlear anatomy in the Post-CT images directly, could at times lead to results that lacked accuracy. The indirect methods, which have been introduced in **Chapter III**, first synthesize an artifact-free image that looks like a Pre-implantation CT (Pre-CT) image of the cochlear implant recipient and then segment the intracochlear anatomy in the synthesized images by using a segmentation algorithm (which we refer to as SegPre-ASM) designed to operate on Pre-CT images (Noble et al., 2011). SegPre-ASM relies on a weighted active shape model created with high-resolution microCT scans of the cochlea acquired ex-vivo in which the intracochlear anatomy is visible. The model is fitted to the partial information available in the Pre-CT images and used to estimate the position of structures not visible in these images. We have shown that the indirect methods substantially and significantly improve the results obtained with the direct methods. However, the final results of the indirect method depend on SegPre-ASM, which could also lead to inaccuracy if the active shape model does not have enough degrees of freedom to capture small local shape variations, thus further improvement is difficult to achieve by using our current indirect approach.

Recently our group proposed a deep learning-based method (D. Zhang et al., 2019) that relies on a 3D U-Net to segment the intracochlear anatomy in Pre-CT images. The model is trained with a two-level approach in which we first train the model with the segmentations generated by SegPre-ASM on a large image set, then we fine-tune it with a small image set for which accurate manual segmentation is available. Results show that the deep learning-based method achieves higher accuracy than SegPre-ASM. Multi-task learning aims to leverage useful information contained in multiple related tasks to help improve the generalization performance of all the tasks, and it has been used successfully across a wide range of machine learning applications. Inspired by these, we explore ways to improve the segmentation accuracy of the intracochlear anatomy in the Post-CT images by using a multi-resolution two-branch deep network that synthesizes artifact-free images from the Post-CT images and segments the intracochlear anatomy simultaneously. The resolution of the output of the synthesis branch is 32×32×32 voxels with a voxel size of 0.4×0.4×0.4 mm^3, which is the same as the typical voxel size of the Post-CT images in our database. The resolution of the output of the segmentation branch is 128×128×128 voxels and the voxel size is 0.1×0.1×0.1 mm^3. This is done to limit the number of model parameters and the memory usage for training while generating high-resolution segmentation labels.

5.2 Materials and Methods

5.2.1 Dataset

Our dataset consists of Post-CT and Pre-CT image pairs of 180 ears. These ears are partitioned into a set of 110 ears for training, a set of 10 ears for validation, and a set of 60 ears for testing. The Pre-CT images are registered to the Post-CT images using intensity-based affine registration techniques. Then the images are resampled to 0.4×0.4×0.4 mm^3 so that all of the images have the same voxel size. Data augmentation is applied to the images for training and validation by rotating each image by 20 random angles in the range of -30 and 30 degrees about the x-, y-, and z-axis. Patches of 32×32×32 voxels that contain the cochlea are cropped from the Pre- and Post-CT images. In this study, we use the segmentation results obtained by applying SegPre-ASM to the real Pre-CT images as ground truth. The outputs of SegPre-ASM are meshes of scala tympani, scala vestibuli, and modiolus, these meshes are converted to segmentation masks with a voxel size of 0.1×0.1×0.1 mm^3, and the size of these label images is 128×128×128 voxels. Figure 5.1 shows a pair of the Post-CT image (Figures 5.1a) and the Pre-CT image (Figures 5.1b) and the ground-truth segmentation label image (Figure 5.1c).

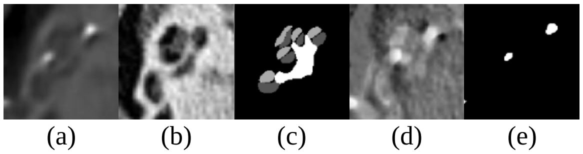


Figure 5.1: A pair of (a) Post-CT and (b) Pre-CT patches and (c) the segmentation label image, wherein the scala vestibuli, the scala tympani, and the modiolus are shown in light gray, dark gray, and white, respectively. (d) The weighting matrix for calculating the weighted L1 loss. (e) A mask of the artifact-affected region.

5.2.2 Network Architecture

As shown in Figure 5.2, the multi-task network consists of (1) a feature extractor composed of a common convolutional block, 2 down-sampling blocks, and 6 residual blocks, (2) a segmentation branch composed of 4 upsampling blocks and a Softmax block, and (3) a synthesis branch that consists of 2 up-sampling blocks and a Tanh block. The input of the multi-task network is a Post-CT patch, the outputs are the segmentation masks of the intracochlear anatomy and the synthetic Pre-CT patch. The synthetic and the real Pre-CT images are fed into the Sobel edge detector, which outputs their edge images. As done in **Chapter III**, (J. Wang et al., 2018, 2019), we also use a discriminator network to provide an adversarial loss to train the synthesis branch.

5.2.3 Training Objectives

5.2.3.1 Adversarial Loss

The adversarial loss can be expressed as:

$$L_{\mathrm{Advs}}(M, D) = \min_{M} \max_{D} \mathbb{E}_{pos,pre}\left[\log\left(D(pos, pre)\right)\right] + \mathbb{E}_{x}\left[\log\left(1 - D(pos, \widehat{pre})\right)\right] \tag{5.1}$$

wherein *M* is the multi-task networks, *D* is the discriminator, *pos* is a Post-CT patch, *pre* is the corresponding real Pre-CT patch, and $\widehat{pre}$ is the synthetic Pre-CT patch generated by the synthesis branch of *M*.

5.2.3.2 L1 Loss

As done in Section 4.2.3.3, we use an L1 loss weighted by the dissimilarity map between *pos* and *pre* (Figure 5.1d) to help the model to pay more attention to the artifact-affected regions. The weighted L1 loss $L_{\mathrm{WL1}}(M)$ is expressed as:

$$L_{\mathrm{WL1}}(M) = \mathbb{E}_{pos,pre}[\|W \circ (pre - \widehat{pre})\|_1] \tag{5.2}$$

in which *W* is the dissimilarity map and ∘ is the element-wise multiplication operation.

As done in (Isola et al., 2017), a conventional L1 loss term that is expressed as:

$$L_{\mathrm{L1}}(M) = \mathbb{E}_{pos,pre}[\|pre - \widehat{pre}\|_1] \tag{5.3}$$

is also used in our experiments.

5.2.3.3 Edge-aware Loss

Inspired by the edge-aware GAN (Yu et al., 2019), we use a loss term that penalizes the dissimilarity between the edges in the real and the synthesized images. The edge-aware loss is defined as:

$$L_{\mathrm{Edge}}(M) = \mathbb{E}_{pos,pre}[\|\mathrm{Sobel}(pre) - \mathrm{Sobel}(\widehat{pre})\|_2] \tag{5.4}$$

wherein $\text{Sobel}(pre)$ and $\text{Sobel}(\widehat{pre})$ are the edge images of the real and the synthetic Pre-CT patches, respectively.

5.2.3.4 Region-aware Loss

By thresholding *pos* at the 99-th percentile of its voxel values, we can obtain a mask that roughly captures the artifact-affected voxels in the image (Figure 5.1e). Assume the index of an affected voxel is (i, j, k), in which $i, j, k \in [1, 128]$, then an affected region-aware loss can be expressed as:

$$L_{\text{Region}}(M) = \mathbb{E}_{pos,pre}\left[\sum_{i,j,k}\|pre(i,j,k) - \widehat{pre}(i,j,k)\|_2\right] \tag{5.5}$$

5.2.3.5 Generalized Dice

The output of the segmentation branch has four channels corresponding to the scala tympani, the scala vestibuli, the modiolus, and the background. Suppose $S \in \boldsymbol{B}^{128\times128\times128\times4}$ ($\boldsymbol{B} = \{0, 1\}$) is the binary mask of the ground-truth segmentation, $S_{i,j,k,n} = 1$ if the (i, j, k)-th voxel belongs to the n-th class, wherein $n = \{1, 2, 3, 4\}$ correspond to the scala tympani, the scala vestibuli, the modiolus, and the background, respectively. $P \in \boldsymbol{R}^{128\times128\times128\times4}$ ($\boldsymbol{R} \in [0, 1]$) is the output of the Softmax block of the segmentation branch of M, wherein $P_{i,j,k,n}$ is the probability of that the (i, j, k)-th voxel of the Post-CT patch belongs to the n-th class. As defined in (Sudre et al., 2017), the generalized Dice that measure the disagreement between the network's prediction and the ground truth can be expressed as:

$$L_{\text{Dice}}(M) = 1 - \frac{2\sum_{n=1}^{4} w_n \sum_{i=1}^{128}\sum_{j=1}^{128}\sum_{k=1}^{128}(S_{i,j,k,n} \times P_{i,j,k,n})}{\sum_{n=1}^{4} w_n \sum_{i=1}^{128}\sum_{j=1}^{128}\sum_{k=1}^{128}(S_{i,j,k,n} + P_{i,j,k,n})} \tag{5.6}$$

wherein $w_n = \frac{1}{\sum_{i=1}^{128}\sum_{j=1}^{128}\sum_{k=1}^{128}(S_{i,j,k,n})}$ is the weighting factor of the n-th class.

5.2.3.6 Weighted Binary Cross-entropy

The weighted binary cross-entropy loss is computed as:

$$L_{\text{BCE}}(M) = \frac{1}{4\times128^3}\sum_{n=1}^{4} w_n \sum_{i=1}^{128}\sum_{j=1}^{128}\sum_{k=1}^{128}\left[-S_{i,j,k,n}\log\left(P_{i,j,k,n}\right) + \left(1 - S_{i,j,k,n}\right)\log\left(1 - P_{i,j,k,n}\right)\right] \tag{5.7}$$

5.2.3.7 Total Loss

The total loss for training our model is the weighted sum of the 7 loss terms defined above.

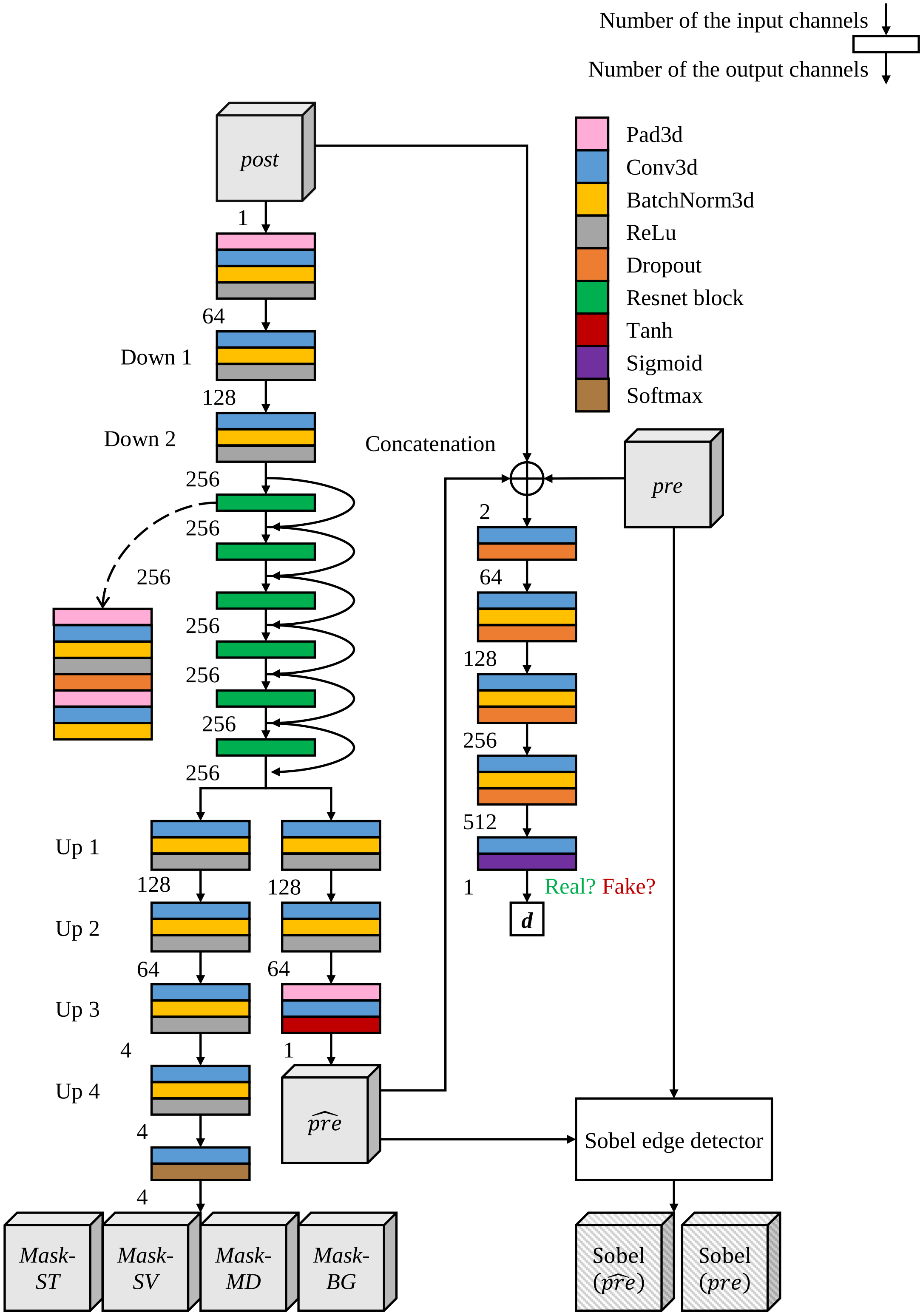


Figure 5.2: An illustration of the network architecture, in which *pos* is a Post-CT image, *pre* is the corresponding real Pre-CT image, and $\widehat{pre}$ is the synthetic Pre-CT image. *Mask-ST*, *Mask-SV*, *Mask-MD*, and *Mask-BG* are segmentation

masks of the scala tympani, the scala vestibuli, the modiolus, and the background respectively. $\mathrm{Sobel}(\widehat{pre})$ and $\mathrm{Sobel}(pre)$ are the edge images of the synthetic Pre-CT image and the real Pre-CT image.

5.3 Experiments

We train our model alternatively between one stochastic gradient descent step on the discriminator, then one step on the multi-task network, using a minibatch size of 1. The model is trained for 200 epochs in which a fixed learning rate of 0.0002 is applied in the first 100 epochs and a learning rate that is linearly reduced to zero in the second 100 epochs. The quality of the synthetic images is measured by the commonly used Mean Structural SIMilarity (MSSIM) index (Z. Wang et al., 2004), and the accuracy of the segmentation is measured by the Dice similarity coefficients between the predicted and the ground-truth segmentation masks of the scala tympani, the scala vestibuli, and the modiolus, which are denoted as DiceST, DiceSV, and DiceMD, respectively. The mean of the Dice values of the intracochlear anatomy is calculated as:

$$\mathrm{MDice} = \frac{1}{3}(\mathrm{DiceST} + \mathrm{DiceSV} + \mathrm{DiceMD}) \tag{5.8}$$

We run inference on the validation set every 5 epochs, and the epoch where it achieves the highest median MDice is selected as the optimal epoch. Note that we have observed a statistically significant positive correlation between MSSIM and MDice on our validation set (Pearson's linear correlation coefficient ~ 0.8, p-value < 10^{-5}). We also experiment with different parameter settings to investigate how each loss term affects the performance of the model. Table 5.1 lists the weighting parameter settings of our 11 experiments.

Cfg	L_{L1}	L_{WL1}	L_{Edge}	L_{Region}	L_{Adv}	L_{Dice}	L_{BCE}
1	1	10	10	10	1	100	50
2	10	10	10	10	1	50	50
3	100	0	0	10	1	100	0
4	0	100	0	10	1	100	0
5	0	100	0	0	1	100	0
6	0	0	10	10	1	100	0
7	0	0	10	10	1	0	100
8	0	0	10	10	1	50	50
9	100	0	0	0	1	100	0
10	0	0	10	0	1	100	0
11	0	0	0	0	0	100	50

Table 5.1: Configurations of the parameters in each experiment.

5.4 Results

The median and standard deviation of MDice and MSSIM for our 11 experiments are listed in Table 5.2. Models trained with parameter setting Cfg1 achieve the best result in terms of the median of MDice. Figure 5.3 shows examples of a good case (MDice = 0.875, MSSIM = 0.826), an average case (MDice = 0.799, MSSIM = 0.815), and a bad case (MDice = 0.156, MSSIM = 0.390). Models trained with Cfg9, Cfg10, and Cfg11 failed to segment the intracochlear anatomy in all testing images. Figure 5.4 shows a representative example in which instead of segmenting the intracochlear anatomy, the model seems to pick up the artifact-affected areas. Figure 5.5 shows boxplots of Dice and MSSIM values of models trained with Cfg1 to Cfg8. For all these 8 experiments, MSSIM and MDice are statistically significantly correlated (Pearson's linear correlation coefficient > 0.65 and p-value < 10^{-5}). There are visible differences between the distributions of MDice and MISSIM among the models (Figure 5.5b, 5.5c), however, these differences are not statistically significant. As shown in Figure 5.5a, the models tend to perform better in segmenting modiolus than scala tympani and scala vestibuli.

Cfg	Median(MDice)	STD(MDice)	Median(MSSIM)	STD(MSSIM)
1	0.792	0.159	0.770	0.117
2	0.790	0.152	0.759	0.106
3	0.789	0.145	0.778	0.117
4	0.785	0.141	0.756	0.098
5	0.783	0.122	0.767	0.096
6	0.779	0.155	0.763	0.116
7	0.776	0.152	0.757	0.105
8	0.774	0.147	0.779	0.110

Table 5.2: The median and STandard Deviation (STD) of MDice and MSSIM in each experiment.

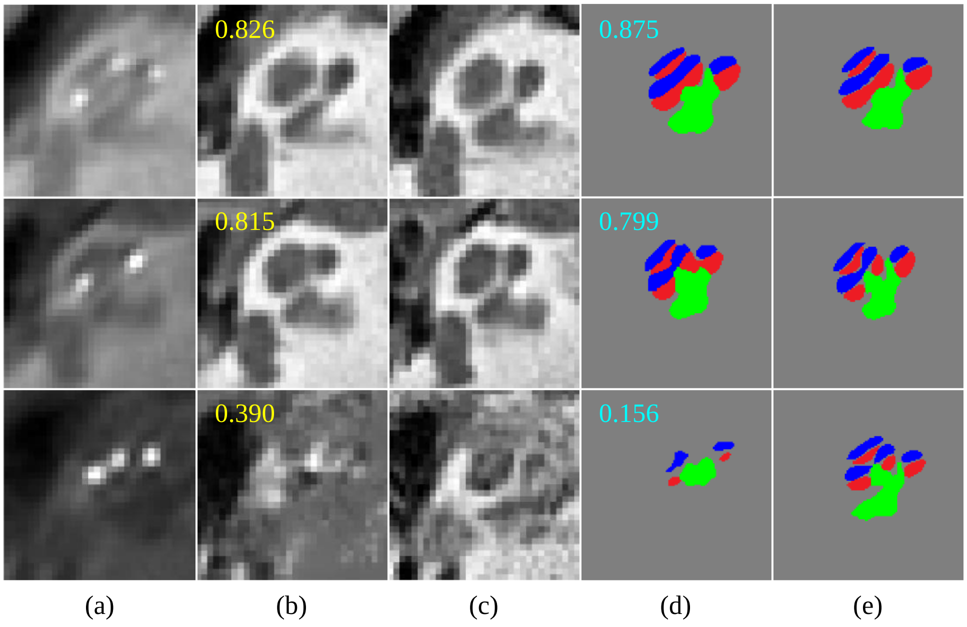


Figure 5.3: A good case (first row), an average case (second row), and a bad case (last row) of a model trained with Cfg1. In each row, (a) is the input Post-CT patches, (b) and (c) are the synthetic and real Pre-CT patches, and (d) and (e) are the predicted and the ground-truth segmentation masks of the intracochlear anatomy. The yellow texts in (b) are the values of MSSIM, the cyan texts in (d) are the MDice values, the red, blue, and green regions in (d), and (e) are the segmentation masks of scala tympani (red), scala vestibuli (blue), and modiolus (green), respectively.

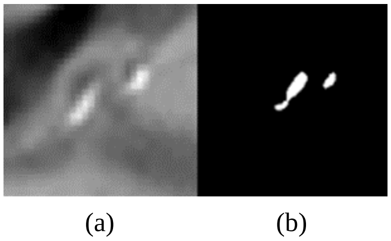


Figure: 5.4: A failed case. (a) The input Post-CT patch. (b) The segmentation mask produced by the model.

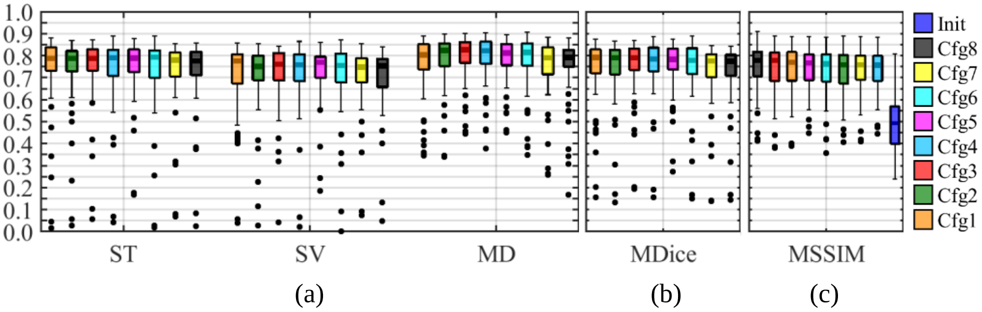


Figure 5.5: Boxplots of (a) DiceST, DiceSV, and DiceMD, (b) MDice, and (c) MSSIM.

5.5 Discussion and Conclusions

The failure of Cfg11 suggests that a single-task network is not feasible for our problem, this also confirms the usefulness of multi-task learning. As opposed to other successful configurations, Cfg9 and Cfg10 do not provide any supervision to help the models to focus on the artifact-affected regions. This suggests that using an attention-related loss is helpful. Though the difference is not significant, Cfg1 and Cfg2 that use all of the loss terms have achieved the best results in terms of MDice. The highest MDice value we have achieved is 0.79, which is good. The parameters of Cfg1 to Cfg12 are selected arbitrarily, and parameter sensitivity analysis has not been done in this study.

As manual ground truth is unavailable for the images used in this study, we only train and evaluate our network using the segmentation results obtained with SegPre-ASM, which is itself imperfect. It has been reported that SegPre-ASM achieves Dice values equal to 0.76, 0.75, and 0.6 for the scala tympani, the scala vestibuli, and the modiolus, respectively, when evaluated on a dataset for which the manual ground truth is known (D. Zhang et al., 2019). It has also been reported that deep learning-based methods that are developed to segment Pre-CT images can produce better results than SegPre-ASM on the dataset for which manually delineated ground truth is available (D. Zhang et al., 2019). It is thus possible that our new approach designed to operate on Post-CT images also produces better results than those achieved previously with the method proposed in **Chapter III**.

We haven't conducted further studies to evaluate our proposed method by going through the fine-tuning and validation steps using another dataset that has manual ground truth, because our current IGCIP system requires surface meshes of the intracochlear anatomy that have a predefined number of vertices, such that the proposed method is not useful for our specific application. However, it can be useful in other situations when a segmentation mask is required.

Acknowledgments

This work has been supported by NIH grants R01DC014037 and R01DC014462 and by the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN. We are grateful for the support of the NVIDIA Corporation for providing us with a GPU used to support this work. The content is solely the responsibility of the authors and does not necessarily represent the official views of these institutes. The authors declare no conflict of interest.

CHAPTER VI

Atlas-Based Segmentation of Intracochlear Anatomy in Metal Artifact Affected CT Images of the Ear with Co-trained Deep Neural Networks

Jianing Wang, Dingjie Su, Yubo Fan, Srijata Chakravorti, Jack H. Noble, and Benoit M. Dawant

Department of Electrical and Computer Engineering

Vanderbilt University

Nashville, TN, 37232, USA

This chapter is adapted from “Atlas-based segmentation of intracochlear anatomy in metal artifact affected CT images of the ear with co-trained deep neural networks”, which has been accepted by MICCAI 2021, the 24th International Conference on Medical Image Computing and Computer Assisted Intervention.

The original paper can be found at https://arxiv.org/abs/2107.03987

Abstract

In **Chapter V** we have developed a one-step method to segment the intracochlear anatomy in the Post-implantation CT (Post-CT) images of cochlear implant recipients. However, it cannot be applied to our Image-guided Cochlear Implant Programming (IGCIP) system directly. As mentioned in earlier chapters, the outputs of SegPre-ASM, SegPost-UL, and SegPost-BL are surface meshes of the scala tympani, the scala vestibuli, and the modiolus that have a predefined number of vertices. Importantly, each vertex corresponds to a specific anatomical location on the surface of the structures and the meshes are encoded with the information needed for the programming of the implant. However, such point-to-point correspondence is not preserved by the method that we have developed in **Chapter V**, because the method generates pixel-wise segmentation masks of the intracochlear anatomical structures.

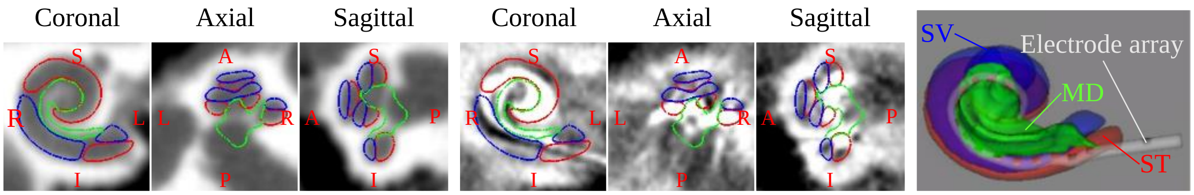
In this chapter, we propose an atlas-based method to segment the intracochlear anatomy in the Post-CT images that preserves the point-to-point correspondence between the meshes in the atlas and the segmented volumes. To solve this problem, which is challenging because of the strong artifacts produced by the implant, we use a pair of co-trained deep networks that generate Dense Deformation Fields (DDFs) in opposite directions. One network is tasked with registering an atlas image to the Post-CT images and the other network is tasked with registering the Post-CT images to the atlas image. The networks are trained using loss functions based on voxel-wise labels, image content, fiducial registration error, and cycle-consistency constraint. The segmentation of the intracochlear anatomy in the Post-CT images is subsequently obtained by transferring the predefined segmentation meshes of the intracochlear anatomy in the atlas image to the Post-CT images using the corresponding DDFs generated by the trained registration networks. Our model can learn the underlying geometric features of the intracochlear anatomy even though they are obscured by the metal artifacts. We show that our end-to-end network produces results that are comparable to the current state of the art for segmenting the Post-CT images, which is the two-step method proposed in **Chapter III**, but it requires a fraction of the time needed by the state-of-the-art method, which is important for end-user acceptance.

6.1 Introduction

The cochlea (Figure 6.1c) is a spiral-shaped structure that is part of the inner ear involved in hearing. It contains two main cavities: the scala tympani and the scala vestibuli. The modiolus is a porous bone around which the cochlea is wrapped that hosts the auditory nerves. A cochlear implant is an implanted neuroprosthetic device that is designed to produce hearing sensations in a person with severe to profound deafness by electrically stimulating the auditory nerves (National Institute on Deafness and Other Communication Disorders, 2016). Cochlear implants are programmed postoperatively in a process that involves activating all or a subset of the electrodes and adjusting the stimulus level for each of these to a level that is beneficial to the recipient (Noble, 2017). Programming parameters adjustment is influenced by the intracochlear position of the cochlear implant's electrodes, which requires the accurate localization of the electrodes relative to the intracochlear anatomy in the Post-implantation CT (Post-CT) images of the recipients of cochlear implants. This, in turn, requires the accurate segmentation of the intracochlear anatomy in the Post-CT images. Segmenting the intracochlear anatomy in the Post-CT images is challenging due to the strong artifacts produced by the metallic electrodes (Figure 6.1b) that can obscure these structures, often severely. For patients who have been scanned before implantation, the segmentation of the intracochlear anatomy can be obtained by segmenting their Pre-implantation CT (Pre-CT) image (Figure 6.1a) using an active shape model-based method (which we refer to as SegPre-ASM) (Noble et al., 2011). The outputs of SegPre-ASM are surface meshes of the scala tympani, the scala vestibuli, and the modiolus that have a predefined number of vertices. Importantly, each vertex corresponds to a specific anatomical location on the surface of the structures and the meshes are encoded with the information needed for the programming of the implant. Preserving point-to-point correspondence when registering the images is thus of critical importance in our application. The intracochlear anatomy in the Post-CT image of the patients can be obtained by registering their Pre-CT image to the Post-CT image and then transferring the segmentations of the intracochlear anatomy in the Pre-CT image to the Post-CT image using that transformation. This approach does not extend to the recipients for whom a Pre-CT image is unavailable. To overcome this issue, we have proposed a two-step method in **Chapter III** (J. Wang et al., 2019, 2018), which we refer to as "cGANs+ASM". The method first uses conditional Generative Adversarial Networks (cGANs) (Mirza & Osindero, 2014; Isola et al., 2017) to synthesize artifact-free Pre-CT images from the Post-CT images and then uses SegPre-ASM (Noble et al., 2011) to segment the intracochlear anatomy in the synthetic images. To the best of our knowledge, cGANs+ASM is the most accurate published automatic method for intracochlear anatomy segmentation in Post-CT images.

Here, we propose an atlas-based method to segment the intracochlear anatomy in the Post-CT images in one step: we first generate a Dense Deformation Field (DDF) between an artifact-free atlas image and a Post-CT image. The segmentation of the intracochlear anatomy in the Post-CT image can then be obtained by transferring the predefined segmentation meshes of the intracochlear anatomy in the atlas image to the Post-CT image using that DDF.

We note that the inter-subject non-rigid registration between the atlas image and the Post-CT image is a difficult task because (1) considerable variation in cochlear anatomy across individuals has been documented (Pelosi et al., 2013), and (2) the artifacts in the Post-CT image change, often severely, the appearance of the anatomy, which has a significant influence on the accuracy of registration methods guided by intensity-based similarity metrics. To overcome the challenges, we propose a method to perform registrations between an atlas image and the Post-CT images that rely on deep networks. Following the idea of consistent image registration obtained by jointly estimating the forward and reverse transformations between two images that is proposed by Christensen et al. (Christensen & Johnson, 2001), we use a pair of co-trained networks that generate DDFs in opposite directions. One network is tasked with registering the atlas image to the Post-CT image and the other network is tasked with registering the Post-CT image to the atlas image. The networks are trained using loss functions that include voxel-wise labels, image content, Fiducial Registration Error (FRE), and cycle-consistency constraint. We show that our model can segment the intracochlear anatomy and preserve point-to-point correspondence between the atlas and the Post-CT meshes, even when the intracochlear anatomy is difficult to localize visually.



(a) (b) (c)

Figure 6.1: A pair of registered (a) Pre-CT and (b) Post-CT images of an ear of a recipient of cochlear implants. (c) An illustration of the intracochlear anatomy with an implanted electrode array. The meshes of the scala tympani (ST, shown in red), the scala vestibuli (SV, shown in blue), and the modiolus (MD, shown in green) are obtained by applying SegPre-ASM to the Pre-CT image.

6.2 Methods

6.2.1 Data

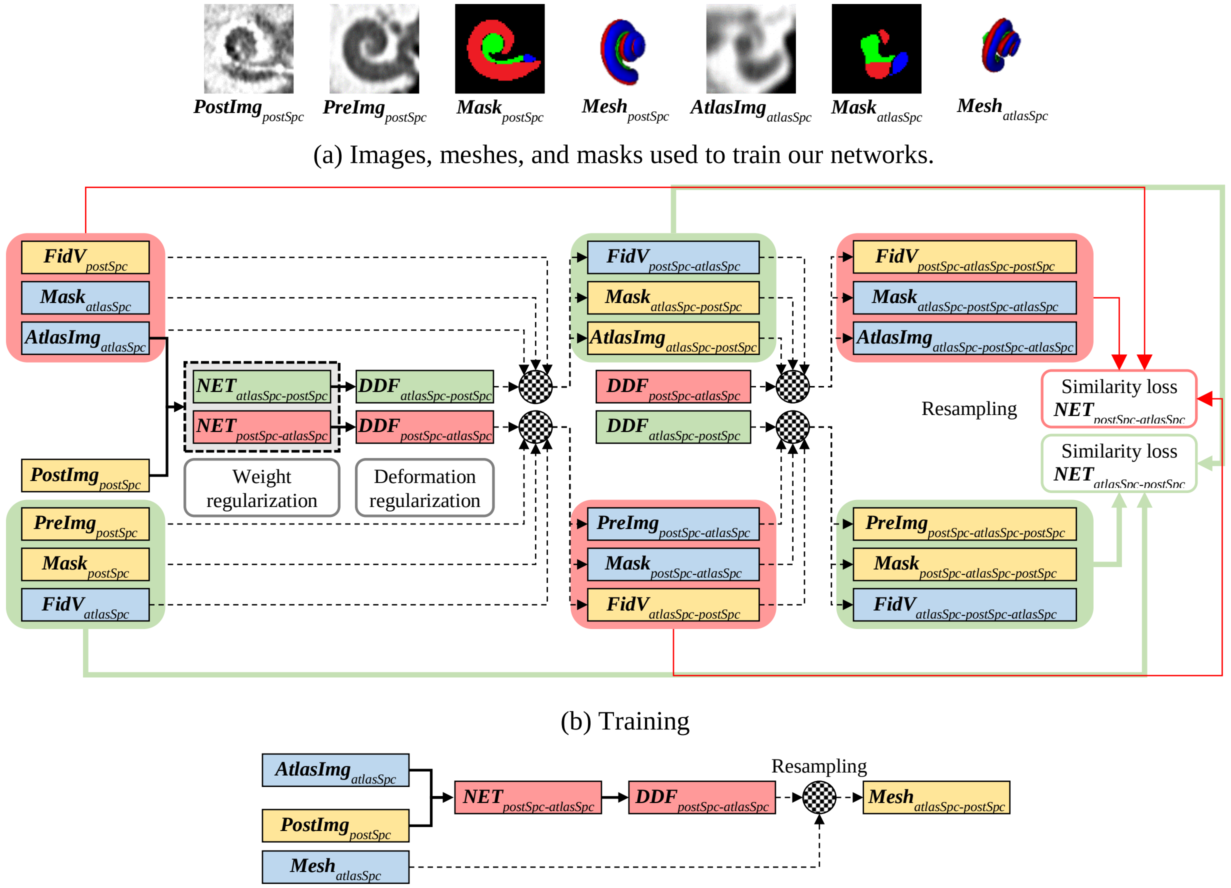
Our dataset consists of Pre-CT and Post-CT image pairs of 624 ears. The atlas image is a Pre-CT image of an ear that is not in the 624 ears. The Pre-CT images are acquired with several conventional scanners (GE BrightSpeed, LightSpeed Ultra; Siemens Sensation 16; and Philips Mx8000 IDT, iCT 128, and Brilliance 64) and the Post-CT images are acquired with a low-dose flat-panel volumetric scanner (Xoran Technologies xCAT® ENT). The typical voxel size is 0.25×0.25×0.3mm^3 for the Pre-CT images and 0.4×0.4×0.4mm^3 for the Post-CT images. For each ear, the Pre-CT image is rigidly registered to the Post-CT image. The registration is accurate because the surgery, which consists of threading an electrode array through a small hole into the bony cavity, does not induce non-rigid deformation of the cochlea. The registered Pre-CT and Post-CT image pairs are then aligned to the atlas image so that the ears are roughly in the same spatial location and orientation. All of the images are resampled to an isotropic voxel size of 0.2mm. Images of 64×64×64 voxels that contain the cochleae are cropped from the full-sized images, and our networks are trained to process such cropped images.

6.2.2 Learning to Register the Artifact-affected Images and the Atlas Image with Assistance of the Paired Artifact-free Images

Figure 6.2a shows a list of objects, i.e., images, meshes, and masks used to train our networks. For simplicity, we use $\boldsymbol{O}_{xSpc}$ to denote an object ***O*** in the *x* space. For example, $\boldsymbol{AtlasImg}_{atlasSpc}$ is our atlas image in the atlas space. Similarly, $\boldsymbol{PostImg}_{postSpc}$ is a Post-CT image in the Post-CT space. $\boldsymbol{Mesh}_{atlasSpc}$ is the segmentation mesh of the intracochlear anatomy in $\boldsymbol{AtlasImg}_{atlasSpc}$ generated by applying SegPre-ASM to $\boldsymbol{AtlasImg}_{atlasSpc}$. $\boldsymbol{PreImg}_{postSpc}$ is the paired Pre-CT image of $\boldsymbol{PostImg}_{postSpc}$ registered to the original Post-CT. $\boldsymbol{Mesh}_{postSpc}$ is the segmentation mesh of the intracochlear anatomy in $\boldsymbol{PostImg}_{postSpc}$. It has been generated by applying SegPre-ASM to $\boldsymbol{PreImg}_{postSpc}$ and then transferring the meshes to $\boldsymbol{PostImg}_{postSpc}$. $\boldsymbol{Mask}_{atlasSpc}$ and $\boldsymbol{Mask}_{postSpc}$ are segmentation masks of the scala tympani, the scala vestibuli, and the modiolus. They are generated by converting $\boldsymbol{Mesh}_{atlasSpc}$ and $\boldsymbol{Mesh}_{postSpc}$ to masks.

As shown in Figure 6.2b, the input of our networks is the concatenation of $\boldsymbol{AtlasImg}_{atlasSpc}$ and $\boldsymbol{PostImg}_{postSpc}$. The networks consist of a first network ($\boldsymbol{NET}_{atlasSpc\text{-}postSpc}$) that generates a DDF from the atlas space to the Post-CT space ($\boldsymbol{DDF}_{atlasSpc\text{-}postSpc}$) and a second network ($\boldsymbol{NET}_{postSpc\text{-}atlasSpc}$) that generates a DDF from the Post-CT space to the atlas space ($\boldsymbol{DDF}_{postSpc\text{-}atlasSpc}$). $\boldsymbol{FidV}_{atlasSpc}$ and $\boldsymbol{FidV}_{postSpc}$ are fiducial vertices randomly sampled from $\boldsymbol{Mesh}_{atlasSpc}$ and $\boldsymbol{Mesh}_{postSpc}$ on the fly for calculating FRE during training.

Assuming that *sSpc* is the source space and *tSpc* is the target space. The Pre-CT image, the segmentation masks, and the fiducial points in *sSpc* are warped to *tSpc* by using the corresponding DDFs (note that one DDF is used for the images and masks and the other for the fiducial points), and the results are denoted as $\boldsymbol{PreImg}_{sSpc\text{-}tSpc}$, $\boldsymbol{Mask}_{sSpc\text{-}tSpc}$, and $\boldsymbol{FidV}_{sSpc\text{-}tSpc}$. Then, $\boldsymbol{PreImg}_{sSpc\text{-}tSpc}$, $\boldsymbol{Mask}_{sSpc\text{-}tSpc}$, and $\boldsymbol{FidV}_{sSpc\text{-}tSpc}$ are transferred back to *sSpc* using the corresponding DDF, and the results are denoted as $\boldsymbol{PreImg}_{sSpc\text{-}tSpc\text{-}sSpc}$, $\boldsymbol{Mask}_{sSpc\text{-}tSpc\text{-}sSpc}$, and $\boldsymbol{FidV}_{sSpc\text{-}tSpc\text{-}sSpc}$, respectively. The training objective for $\boldsymbol{NET}_{sSpc\text{-}tSpc}$ can be constructed by using similarity measurements between the target object in *tSpc* (denoted as $\boldsymbol{O}_{tSpc}$) and the source object that has been transferred to *tSpc* from *sSpc* (denoted as $\boldsymbol{O}_{sSpc\text{-}tSpc}$). Specifically, we use the Multiscale Soft Probabilistic Dice (MSPDice) (Milletari et al., 2016) between $\boldsymbol{Mask}_{tSpc}$ and $\boldsymbol{Mask}_{sSpc\text{-}tSpc}$, which is denoted as MSPDice($\boldsymbol{Mask}_{tSpc}$, $\boldsymbol{Mask}_{sSpc\text{-}tSpc}$), to measure the similarity of the segmentation masks. The multiscale soft probabilistic Dice is less sensitive to the class imbalance in the segmentation tasks and is more appropriate for measuring label similarity in the context of image registration (Hu, Modat, Gibson, Li, et al., 2018). The similarity between $\boldsymbol{FidV}_{tSpc}$ and $\boldsymbol{FidV}_{sSpc\text{-}tSpc}$ is measured by the mean fiducial registration error $\overline{\mathrm{FRE}}$($\boldsymbol{FidV}_{tSpc}$, $\boldsymbol{FidV}_{sSpc\text{-}tSpc}$), which is calculated as the average Euclidean distance between the vertices in $\boldsymbol{FidV}_{tSpc}$ and the corresponding vertices in $\boldsymbol{FidV}_{sSpc\text{-}tSpc}$. The Post-CT images cannot be used for calculating intensity-based loss due to the artifacts, thus we use the Normalized Cross-Correlation (NCC) between $\boldsymbol{PreImg}_{tSpc}$ and $\boldsymbol{PreImg}_{sSpc\text{-}tSpc}$, which is denoted as NCC($\boldsymbol{PreImg}_{tSpc}$, $\boldsymbol{PreImg}_{sSpc\text{-}tSpc}$), to measure the similarity between the warped source image and the target image. A cycle-consistency loss is used for regularizing the transformations. It imposes inverse consistency between the objects in the two spaces and has been shown to reduce folding problems (Kim et al., 2019). Our cycle-consistency loss $\boldsymbol{CycConsis}_{sSpc\text{-}tSpc}$ measures the similarity between the original source objects in the source space and the source objects that have been transferred from the source space to the target space and then transferred back to the source space, it is calculated as MSPDice($\boldsymbol{Mask}_{sSpc}$, $\boldsymbol{Mask}_{sSpc\text{-}tSpc\text{-}sSpc}$) + 2× $\overline{\mathrm{FRE}}$ ($\boldsymbol{FidV}_{sSpc}$, $\boldsymbol{FidV}_{sSpc\text{-}tSpc\text{-}sSpc}$) + 0.5×NCC($\boldsymbol{PreImg}_{sSpc}$, $\boldsymbol{PreImg}_{sSpc\text{-}tSpc\text{-}sSpc}$). Furthermore, the DDF from the source space to the target space $\boldsymbol{DDF}_{sSpc\text{-}tSpc}$ is regularized using bending energy (Rueckert et al., 1999), which is denoted as BendE($\boldsymbol{DDF}_{sSpc\text{-}tSpc}$). The learnable parameters of the registration network $\boldsymbol{NET}_{sSpc\text{-}tSpc}$ (except for the biases) are regularized by an L2 term, which is denoted as L2($\boldsymbol{NET}_{sSpc\text{-}tSpc}$). To summarize, the training objective for our networks is the weighted sum of the loss terms listed in Table 6.1; wherein the weights have been selected empirically by looking at training performance on a small number of epochs.



(a) Images, meshes, and masks used to train our networks.

(b) Training

(c) Inference

Figure 6.2: The framework of our method. (a) Objects used for training the networks. (b) Training phase. (c) Inference phase.

Loss	Definition	Weight
MSPDice	MSPDice($\boldsymbol{Mask}_{postSpc}$, $\boldsymbol{Mask}_{atlasSpc\text{-}postSpc}$) + MSPDice($\boldsymbol{Mask}_{atlasSpc}$, $\boldsymbol{Mask}_{postSpc\text{-}atlasSpc}$)	1
Mean FRE	$\overline{\text{FRE}}$($\boldsymbol{FidV}_{postSpc}$, $\boldsymbol{FidV}_{atlasSpc\text{-}postSpc}$) + $\overline{\text{FRE}}$($\boldsymbol{FidV}_{atlasSpc}$, $\boldsymbol{FidV}_{postSpc\text{-}atlasSpc}$)	2
NCC	NCC($\boldsymbol{PreImg}_{postSpc}$, $\boldsymbol{AtlasImg}_{atlasSpc\text{-}postSpc}$) + NCC($\boldsymbol{AtlasImg}_{atlasSpc}$, $\boldsymbol{PreImg}_{postSpc\text{-}atlasSpc}$)	0.5
Cycle-consistency	$\boldsymbol{CycConsis}_{atlasSpc\text{-}postSpc}$ + $\boldsymbol{CycConsis}_{postSpc\text{-}atlasSpc}$	0.5
BendE	BendE($\boldsymbol{DDF}_{atlasSpc\text{-}postSpc}$) + BendE($\boldsymbol{DDF}_{postSpc\text{-}atlasSpc}$)	0.5
L2	L2($\boldsymbol{NET}_{atlasSpc\text{-}postSpc}$) + L2($\boldsymbol{NET}_{postSpc\text{-}atlasSpc}$)	0.0001

Table 6.1: Loss terms that are used to train our model.

6.2.3 Network Architecture

The registration networks in our model are adapted from the network architecture proposed by Hu et al. (Hu, Modat, Gibson, Ghavami, et al., 2018) and Ghavami et al. (Ghavami et al., 2018). As shown in Figure 6.3, $\boldsymbol{NET}_{sSpc\text{-}tSpc}$, which is tasked with generating a DDF for warping the source image *S* to the target image *T*, is composed of a Global-net and a Local-net. After receiving the concatenation of *S* and *T*, the Global-net generates an affine transformation matrix. *S* is warped to *T* by using this affine transformation and the resulting image is denoted as *S'*. Then, the Local-net takes the concatenation of *S'* and *T* to generate a non-rigid local DDF. The affine transformation and the local DDF are composed to produce the output DDF. The details about the Global-net and Local-net can be found in (Hu, Modat, Gibson, Ghavami, et al., 2018).

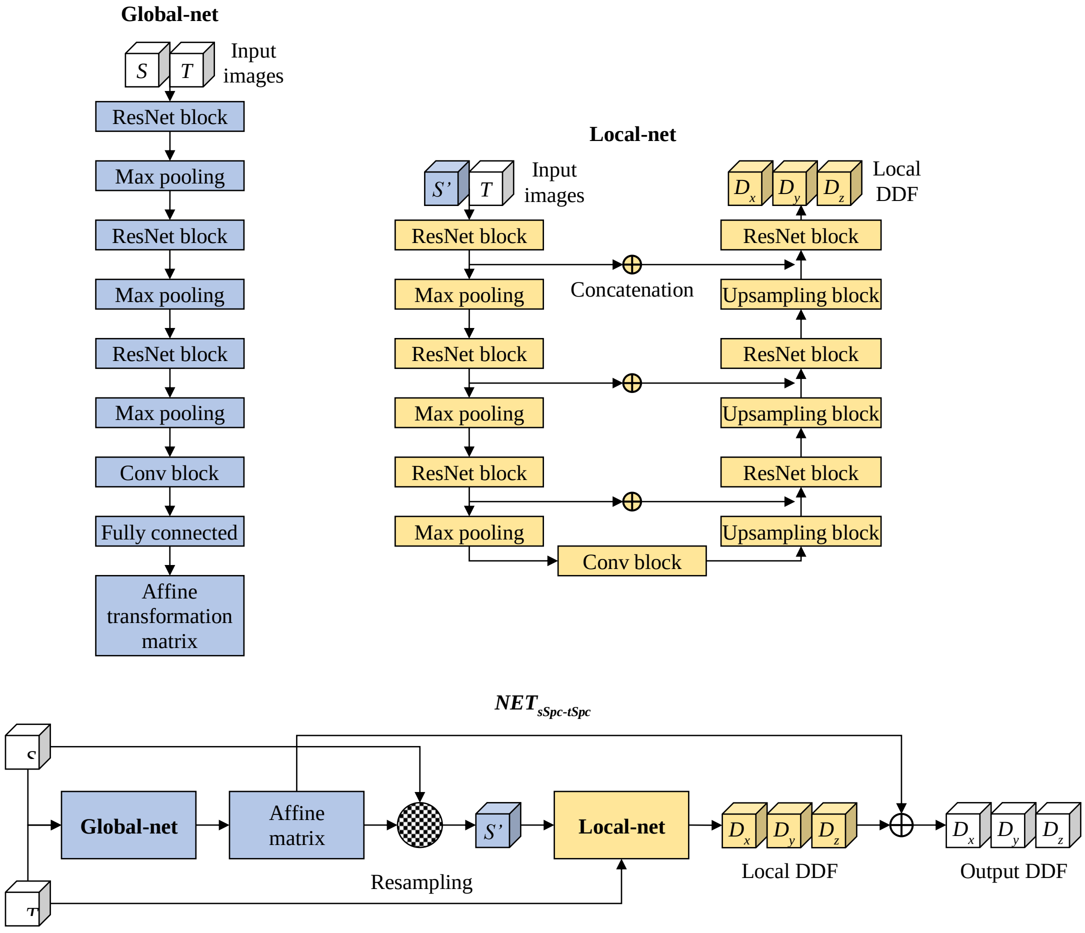


Figure 6.3: Illustration of a registration network $\boldsymbol{NET}_{sSpc\text{-}tSpc}$ that is tasked to generate a DDF from the source space to the target space.

6.2.4 Evaluation

As shown in Figure 6.2c, at the inference phase, given a new Post-CT image $\boldsymbol{PostImg}_{postSpc}$, the intracochlear anatomy in $\boldsymbol{PostImg}_{postSpc}$ can be segmented by warping $\boldsymbol{Mesh}_{atlasSpc}$ to $\boldsymbol{PostImg}_{postSpc}$ using the DDF generated by the trained network. The resulting segmentation mesh of the intracochlear anatomy is denoted as $\boldsymbol{Mesh}_{atlasSpc\text{-}postSpc}$. $\boldsymbol{Mesh}_{postSpc}$, which has been described in Section 6.2.2, is used as the ground truth for comparison. As $\boldsymbol{Mesh}_{atlasSpc}$ and $\boldsymbol{Mesh}_{postSpc}$ are the outputs of SegPre-ASM, both of them have a predefined number of vertices, and the vertices of $\boldsymbol{Mesh}_{atlasSpc}$ and $\boldsymbol{Mesh}_{postSpc}$ have a one-to-one correspondence. There are 3344, 3132, and 2852 vertices on the scala tympani, the scala vestibuli, and the modiolus mesh surfaces, respectively, for a total of 9328 vertices. Point-to-Point Error (P2PE), computed as the Euclidean distance in millimeter, between the corresponding vertices on $\boldsymbol{Mesh}_{atlasSpc\text{-}postSpc}$ and $\boldsymbol{Mesh}_{postSpc}$ are used to quantify the accuracy of the segmentation and registration.

The P2PEs between the corresponding vertices on $\boldsymbol{Mesh}_{postSpc}$ and the meshes generated by cGANs+ASM are calculated and serve as values that are used to compare the proposed method with the state of the art.

The method proposed in (Hu, Modat, Gibson, Ghavami, et al., 2018), which uses a unidirectional registration network trained with the MSPDice loss and the regularization loss, is used as a baseline for comparison. In addition to the MSPDice loss and the regularization loss, our training objective also includes the FRE loss, NCC loss, and the cycle-consistency loss. An ablation study is conducted to analyze how these loss terms affect the performance of our networks.

6.3 Experiments

The 624 ears are partitioned into 465 ears for training, 66 ears for validation, and 93 ears for testing. The partition is random, with the constraint that ears of the same object should not be used in both training and testing. We apply augmentation to the training set by rotating each image by 6 random angles in the range of -25 and 25 degrees about the x-, y-, and z-axis, such that 18 additional images are created from each original image. The training images are blurred by applying a Gaussian filter with a kernel size selected randomly from {0, 0.5, 1.0, 1.5} with equal probability. This results in a training set that is expanded to 8835 images. Each image is clipped between its 5-th and 95-th intensity percentiles, and the intensity values are subsequently rescaled to -1 to 1. We use a batch size of 1, at each training step, 30% of the vertices on the intracochlear anatomy meshes are randomly sampled and used as the fiducial points for calculating the FRE loss. The registration networks used in our proposed method, the cGANs used in the state-of-the-art approach and the registration network used in the baseline method are trained and validated using the same set of images.

6.4 Results

Figure 6.4 shows three example cases of our testing results. The first column shows the Post-CT images to be segmented. These images illustrate the severity of the artifact introduced by the implant. In Case 3, the cochlea is barely visible. The second to the fourth columns show the segmentation meshes generated by cGAN+ASM, our proposed method, and the baseline method, respectively, and these meshes are color-coded with the P2PE at each vertex on the meshes. The last two columns show the paired Pre-CT images of the Post-CT images and the ground truth segmentation meshes. As shown in the figure, the proposed method and cGAN+ASM achieve similar accuracy for Case 1. For Case 2, the proposed method achieves a better outcome. For Case 3, the cGAN+ASM method performs better. In general, the proposed method and cGANs+ASM achieve comparable good results that have small segmentation errors. In contrast, the baseline method has significantly larger errors.

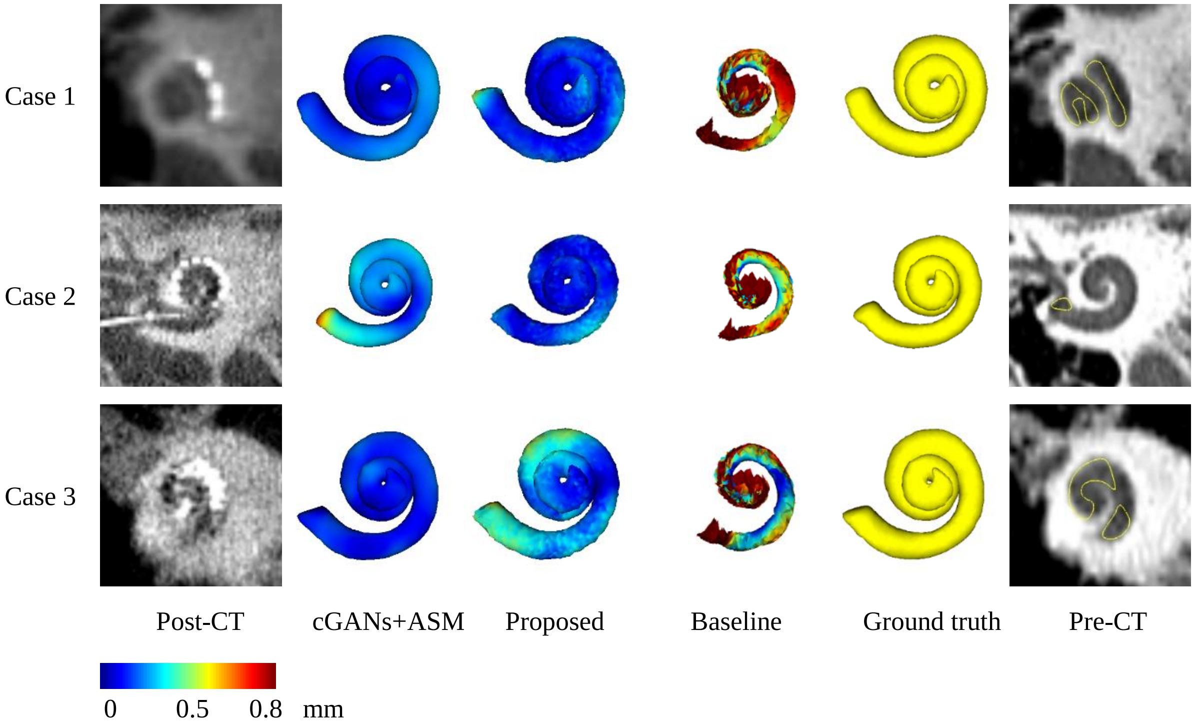
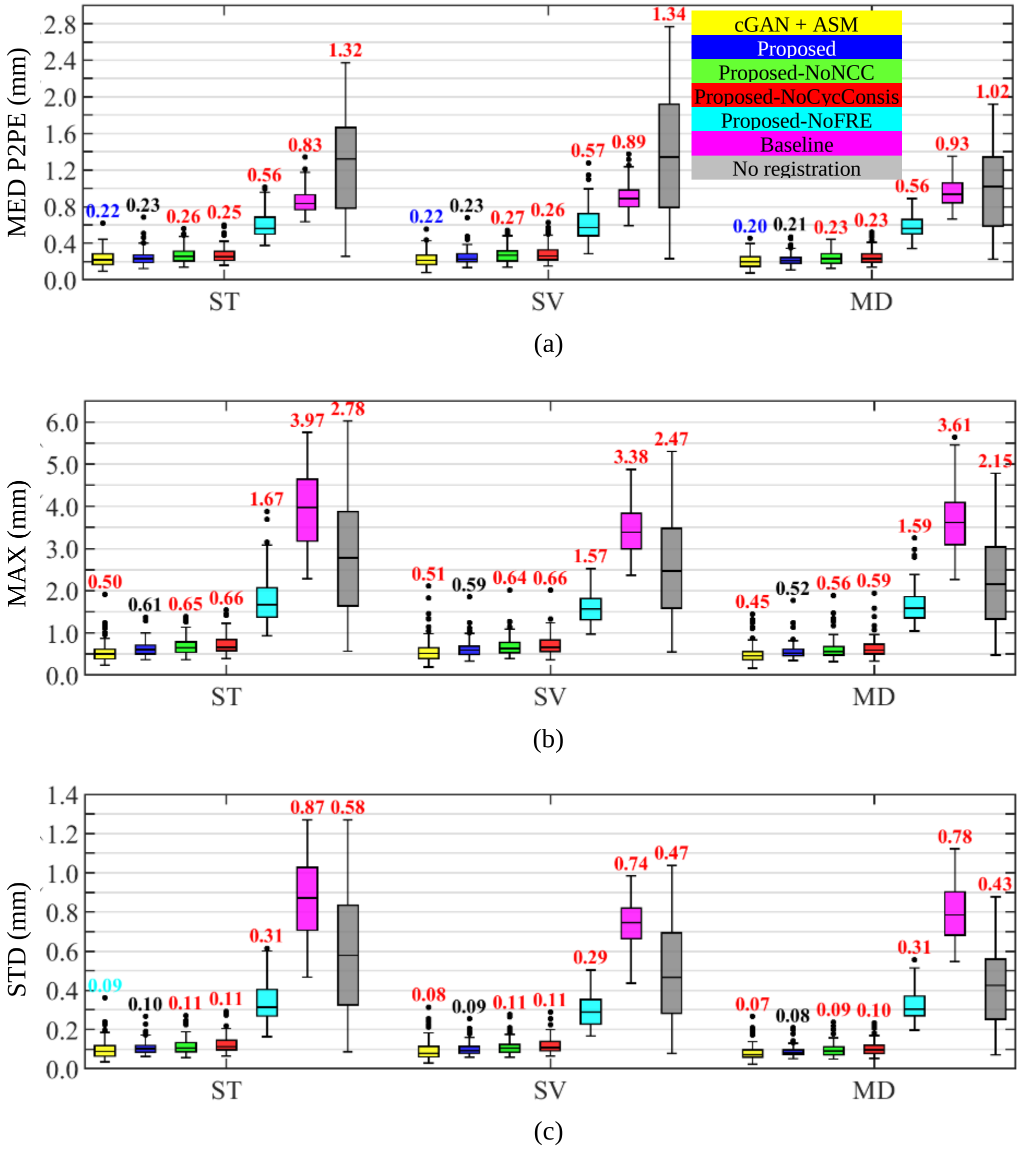


Figure 6.4: Three example cases of our testing results. The meshes shown in the second to the fourth columns are the segmentation meshes of the scala vestibuli generated by cGAN+ASM, our proposed method, and the baseline method, respectively, and these meshes are color-coded with the P2PE at each vertex on the meshes. The meshes in the fifth column are the ground-truth meshes of the scala vestibuli generated by applying Seg-Pre-ASM to the Pre-CT image.

For each testing ear, we calculate the P2PEs of the vertices on the mesh surfaces of the scala tympani, the scala vestibuli, and the modiolus, respectively. We calculate the MAXimum (MAX), MEDian (MED), and STandard Deviation (STD) of the P2PEs. Figure 6.5 shows the boxplots of these statistics for the 93 testing ears. “cGAN+ASM” denotes the results of the state of the art. “Proposed” denotes the results of our method. “Proposed-NoNCC”, “Proposed-NoCycConsis”, and “Proposed-NoFRE” denote the results of our proposed networks trained without using the NCC loss, the cycle-consistency loss, and the FRE loss, respectively. “Baseline” denotes the results of the baseline method. “No registration” denotes the P2PEs between the vertices on the mesh surfaces in the original atlas space and the Post-CT space. We perform two-sided and one-sided Wilcoxon signed-rank tests between the “Proposed” group and the other groups. The p-values have been corrected using the Holm-Bonferroni method (Holm, 1979). The median values for each group are shown on top of the boxplots, in which red denotes that both the two-sided and the one-sided tests are significant, cyan denotes that only the two-sided test is significant, and blue denotes that the two-sided test is not significant. The results show that our networks trained using all of the proposed loss terms achieve a significantly lower segmentation error compared to the baseline method and the networks that are not trained using all of the loss terms. Our method produces results that are similar to those obtained with the state of the art in terms of the MEDs of the segmentation error. The MAXs of the segmentation error and the STDs of the segmentation error for the scala vestibuli and modiolus remain slightly superior to those obtained with the state of the art.



(a)

(b)

(c)

Figure 6.5: Boxplots of (a) the MED, (b) the MAX, and (c) the STD of the P2PEs. A description of the numerical value color legend can be found in the text. (ST – scala tympani, SV – Scala Vestibuli, MD – MoDiolus)

Our proposed approach produces results in a fraction of the time needed by the state of the art. As mentioned earlier, the state of the art is a two-step process: (1) generate a synthetic Pre-CT image from a Post-CT image with cGANs trained for this purpose and (2) apply SegPre-ASM to the synthetic image. Step 2 requires the very accurate registration of an atlas to the image to be segmented to initialize the active shape model. This is achieved through an affine and then a non-rigid intensity-based registration in a volume-of-interest that includes the inner ear. Step 1 takes about 0.3s while step 2 takes on average 75s. The proposed method only requires providing a volume-of-interest that includes the inner ear to the networks and inference time is also about 0.3s. Segmentation is thus essentially instantaneous with the proposed method while it takes over a minute with the state of the art. This is of importance for clinical deployment and end-user acceptance.

6.5 Summary

We have developed networks capable of performing image registration between artifact-affected CT images and an artifact-free atlas image, which is a very challenging task because of the severity of the artifact introduced by the implant. Because we need to maintain point-to-point correspondence between meshes in the atlas and meshes in the segmented Post-CT images, we have introduced a point-to-point loss, which, to the best of our knowledge, has not yet been proposed. Our experiments have shown that this loss is critical to achieving results that are comparable to those obtained with the state of the art that relies on an active shape model fitted to a preoperative image synthesized from a post-operative image. By design, active shape model-based methods always produce plausible shapes. We have observed that with the point-to-point loss, our network also produces plausible shapes even when the images are of very poor quality. We hypothesize that, thanks to the point-to-point loss, the network has been able to learn the shape of the cochlea and can fit this shape to partial information in the post-operative image. More experiments are required to verify this hypothesis.

Acknowledgments

This work has been supported by NIH grants R01DC014037 and R01DC014462 and by the Advanced Computing Center for Research and Education (ACCRE) of Vanderbilt University. The content is solely the responsibility of the authors and does not necessarily represent the official views of these institutes.

CHAPTER VII

Retrospective Evaluation of a Technique for Patient-Customized Placement of Pre-curved Cochlear Implant Electrode Arrays

Jianing Wang[1], Benoit M. Dawant[1], Robert F. Labadie[2], Jack H. Noble[1]

[1] Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, 37232, USA

[2]Department of Otolaryngology-Head and Neck Surgery, Vanderbilt University Medical Center, 7209 Medical Center East, South Tower, Nashville, TN 37232, USA

This chapter is adapted from "Retrospective evaluation of a technique for patient-customized placement of precurved cochlear implant electrode arrays" published in Otolaryngology-Head and Neck Surgery, 2017, 157(1): 107-112.

The published paper can be found at https://doi.org/10.1177/0194599817697298

It has been reproduced with the permission of the publisher.

Abstract

Pre-curved electrode arrays are commonly used in cochlear implants. Modiolar placement of such arrays has been shown to lead to better hearing outcomes. In this chapter, we retrospectively evaluated the modiolar positioning of electrode arrays within a large cochlear implant imaging database. We aimed to discover the rate at which perimodiolar placement is successfully achieved and to evaluate a new technique we propose to pre-operatively plan patient-customized electrode array insertion depths to improve perimodiolar placement at the time of surgery. Ninety-seven ears with cochlear implants were evaluated. Perimodiolar positioning of electrodes was quantified using pre- and post-implantation CT scans and automated image analysis techniques. The average perimodiolar distance was 0.59 ± 0.18mm. Disagreement between the actual and our recommended insertion depth was found to be positively correlated with perimodiolar distance ($r = 0.49$, $p < 0.0001$). These results show that the average cochlear implant recipient with a pre-curved electrode array has a number of electrodes distant to the modiolus where they are not most effective. Our results also indicate the approach we propose for selecting patient-customized electrode array insertion depth would lead to better perimodiolar placement of pre-curved electrode arrays.

7.1 Introduction

With over 324,000 cochlear implants performed worldwide, cochlear implant surgery has become the preferred treatment for severe to profound hearing loss (National Institute on Deafness and Other Communication Disorders, 2016). In cochlear implant surgery, an electrode array, which directly stimulates auditory nerves to induce the sensation of hearing, is implanted ideally into the scala tympani of the cochlea. The auditory nerves have a finely tuned tonotopic spatial arrangement where stimulation of nerves near the cochlear entrance, also known as the basal turn, induces the sensation of high-frequency sound, and stimulation of nerves deeper within the cochlea, near the apex, induces the sensation of lower frequency sounds (Greenwood, 1990). Cochlear implants take advantage of this tonotopic with electrode arrays designed to position electrodes along the length of the cochlea so that deeper electrodes can be activated to stimulate low frequency-specific neurons and more shallow electrodes stimulating high-frequency specific neurons. Current commercially available electrode arrays can be broadly divided into two classes – straight and pre-curved. Historically, pre-curved electrode arrays are more widely used and are designed such that the resting state shape of the array roughly matches the coiled shape of the "average" human cochlea noting that wide variation in cochlear shape has been reported (Pelosi et al., 2013). The shape of the cochlea can be seen in the surface of Figure 7.1a where the green curve traces along the inner wall of the cochlea shows where pre-curved electrode arrays are intended to be positioned. Such perimodiolar positioning places the electrodes closer to the spiral ganglion cells – the intended sites of stimulation leading to activation of the auditory nerve and hearing. Positioning of electrodes closer to the spiral ganglion has been hypothesized to improve hearing outcomes potentially by reducing current spread allowing a cleaner signal to be delivered to the neurons (Rubinstein, 2004). While it remains inconclusive whether pre-curved or straight electrode arrays lead to better outcomes, it has been shown that when pre-curved electrode arrays are used, better outcomes are achieved with better perimodiolar positioning of the array in a study by Holden et al. (Holden et al., 2013). In that study, a correlation of r = 0.402 was found between Consonant-Nucleus-Consonant (CNC) word recognition scores (Peterson & Lehiste, 1962) and their "wrapping factor" metric proposed to indirectly measure modiolar positioning of the array. The CNC word recognition score measures the percentage of words in a CNC word list correctly recognized when presented to the subject in a sound booth. An increase in CNC word scores represents an increase in a subject's speech recognition, which can have a significant impact on hearing quality of life. Thus, the study of Holden et al. suggests that any advance in the cochlear implant procedure that leads to better perimodiolar positioning of pre-curved electrode arrays could lead to an increase in average speech recognition, which would have a significant impact on the cochlear implant community.

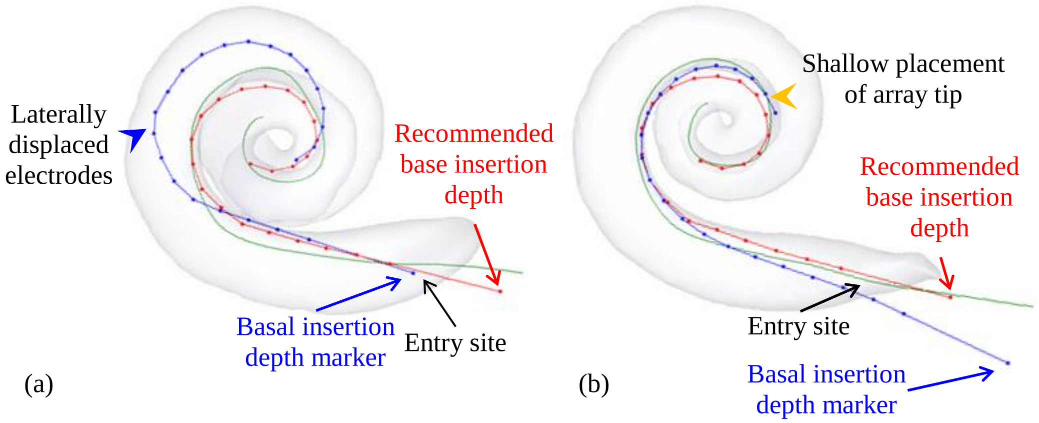

Laterally displaced electrodes
Recommended base insertion depth
Basal insertion depth marker
Entry site
(a)
Shallow placement of array tip
Recommended base insertion depth
Entry site
Basal insertion depth marker
(b)

Figure 7.1: (a) An electrode insertion that is too deep (blue dotted curve). (b) An electrode insertion that is too shallow (blue dotted curve).

Currently, both straight and pre-curved arrays are inserted using “soft” surgical techniques in which the array is threaded at a tangential angle into the scala tympani through either the existing round window membrane or a separate cochleostomy site while attempting to inflict as little trauma as humanly possible on the soft tissue contained within the cochlea (Roland & Wright, 2006). For pre-curved electrode arrays, the array is advanced straightened using either an internal or external stylet. This allows insertion tangential to the basal turn – an essentially straight entry – following which it is threaded off the stylet curving around the cochlea until a predetermined “average” depth is reached. This is a one-size-fits-all approach as the patient-specific cochlear size and shape are not considered. It may have merit if the variations in anatomy are clinically insignificant but it is well-known that considerable variation in cochlear anatomy exists across individuals (Hardy, 1938; Pelosi et al., 2013). Because the array inside the scala tympani is invisible to the surgeon, how close it is placed to the modiolar wall of the scala tympani is generally unknown during and after the surgery. Recently developed automated image analysis techniques permit using post-implantation CT images to detect the intracochlear position of the cochlear implant’s electrodes relative to intracochlear anatomy (Noble et al., 2011; Zhao et al., 2014; Noble & Dawant, 2015). These make it feasible, for the first time, to evaluate the electrode positioning on a large scale. The focus of the work presented in this chapter is to use these techniques to retrospectively evaluate the position of pre-curved cochlear implant electrode arrays on a large cohort. Further, we retrospectively evaluate whether customized insertion depths would improve perimodiolar placement.

7.2 Methods

With IRB approval, we collected pre- and post-implantation CT scans (denoted as Pre-CT scans and Post-CT scans) of 97 ears with cochlear implants. Eighty-two of the ears were implanted with a Cochlear (Sydney, Australia) Contour Advance array (A1), and 15 were implanted with an Advanced Bionics (Valencia, CA, USA) Mid-Scala array (A2).

Segmentation of intracochlear anatomy in the Pre-CT scan was performed using an automatic active shape model-based method (Noble et al., 2011; Cootes et al., 1995). The electrodes were identified in Post-CT images using automatic or semi-automatic electrode extraction techniques (Noble & Dawant, 2015; Zhao et al., 2014). The location of the electrodes relative to intracochlear anatomy was quantified by rigidly registering the pre- and post-implantation datasets using mutual information-based methods (Viola & Wells, 1995; Maes et al., 1997). Intracochlear anatomy is identified in the Pre-CT image rather than the Post-CT image because cochlear anatomy is typically more difficult to identify in Post-CT image where the hardware of the cochlear implant causes significant beam hardening artifacts.

7.2.1 Estimating Perimodiolar Position

To evaluate how often a perimodiolar positioning of the electrodes was achieved in our cohort, we defined the modiolar position in the scala tympani of our active shape model that we use to localize intracochlear anatomy (Noble et al., 2011) by manually defining a 3D modiolar hugging curve as a point sequence within the scala tympani using in-house developed 3D object editing software. A cubic spline was then fitted to the manually selected points to generate a dense and smooth 3-D curve (green curve in Figure 7.1a) (Press et al., 1992). This modiolar curve is automatically transformed to each patient's ear using a Thin Plate Spline (TPS) registration (Goshtasby, 1988) of the model scala tympani to the patient scala tympani. TPS is a landmark-based non-rigid registration method and requires a one-to-one point correspondence between the landmarks used to define the transformation – in this case, the landmarks being the surface points of the patient and model scala tympani surfaces. The modiolar curve computed using the TPS was visually verified to be correct in each case.

The average distance between the electrodes and the modiolar curve was used as a measure of array positioning and was calculated as the average distance between each electrode and its closest point on the modiolar curve. This distance, $\overline{M}$, is calculated as:

$$\overline{M} = \frac{1}{K}\sum_{i=1}^{K}|E_i - E_i^*| \tag{7.1}$$

where K is the number of electrodes, and $|E_i - E_i^*|$ is the distance between the electrode E_i and its closest point on the modiolar curve E_i^*. Figures 7.2a, 7.2b, and 7.2c show 3 cases where the values of $\overline{M}$ are 0.173mm, 0.497mm, and 1.062mm. Using this metric, large $\overline{M}$ indicates the array is further from the modiolus, and $\overline{M}$ close to zero indicates optimal perimodiolar positioning.

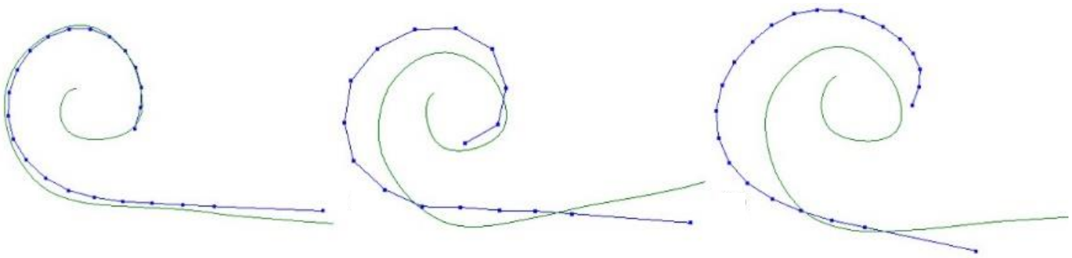


Figure 7.2: Shown are the modiolar curves (green), and the actual electrode positions for each of the 3 subjects (blue dotted curves) with an increasing average distance between electrodes and modiolus.

7.2.2 Patient-customized Electrode Placement Method

We hypothesize that one approach for increasing the probability of achieving optimal perimodiolar positioning is through a patient-customized selection of the electrode array base insertion depth. Currently, the generically recommended insertion depth is specified by the manufacturer to be the depth at which the surgeon should stop inserting. The generic insertion depth is reached when a visual marker on the base of the electrode array reaches the cochlea entry site. To implement a patient-customized base depth, the surgeon would alter this insertion depth to a pre-operatively defined, patient-customized depth conveyed in reference to the depth marker (e.g., “Insert the array until the base depth marker is 1mm outside the entry site.”). Calculating this customized depth required two steps. First, we created electrode array shape models for each array, A1 and A2, by localizing the electrodes from the CT images of the electrode arrays in the air under no load. Second, the electrode array shape model was rigidly registered to the patient’s modiolar curve using an iterative closest point registration technique (Besl & McKay, 1992). Figures 7.1a and 7.1b show the output of this process for two cases with the red dotted curves representing the electrode array shape models registered to the modiolar curves – we will refer to this registered electrode array shape model curve as the “optimal electrode position curve” for that individual. We then determine the patient-customized insertion base depth as the depth of the base of the optimal electrode position curve.

To reiterate, this approach outputs a recommended insertion depth at which the resting state shape of the array is in best agreement with the shape of the patient’s modiolar curve, i.e. $\bar{M}$ is minimized, no matter where the depth marker ends up relative to the cochleostomy. Because the array has likely returned to its resting shape once the base of the array has been inserted to this depth, deeper insertion of the electrode array would not result in further advancement of the tip of the array but push the middle portion of the array away from the modiolus. Figure 7.1a shows such a clinical case of over-insertion where the blue dotted curve represents the position of the electrodes. The basal depth marker was inserted ~2mm past our recommended depth. This additional advancement likely resulted in the lateral displacement of the electrode array away from the modiolus. On the other hand, inserting the base of the array shallower than the recommended depth would lead to a shallower depth of the tip of the array, which also could be detrimental to audiological outcomes as the range of nerves stimulated by the array may be reduced. A representative shallow insertion case is shown as the blue dotted curve in Figure 7.1b. The basal depth marker was placed ~2mm further from the entry site than recommended by our model. While the electrodes are perimodiolar, the tip of the array is much shallower than the expected depth demonstrated by the red curve.

7.2.3 Evaluation of Patient-customized Insertion Depth Technique

To evaluate whether our patient-customized approach would lead to more perimodiolar placement of the array, we perform a retrospective analysis to determine whether an agreement between the actual base insertion depth achieved for patients and our recommended insertion depth is predictive of modiolar hugging placement of the electrode array. For each subject in our dataset, we measure the distance between the position of the depth marker on the surgically placed electrode array, E_d, and the position where the marker is recommended to be by the techniques described above, $\hat{E}_d$ (see Figure 7.3). To ensure the distance measured corresponds to a difference in insertion depth, we define $G(E_d, \hat{E}_d)$ as the distance in the direction tangential to the insertion by projecting E_d and $\hat{E}_d$ onto the modiolar curve and then measuring the geodesic distance between those two points along the modiolar curve. $G(E_d, \hat{E}_d)$ is a signed function with positive values indicating that E_d is deeper than $\hat{E}_d$ and negative values indicating $\hat{E}_d$ is deeper than E_d.

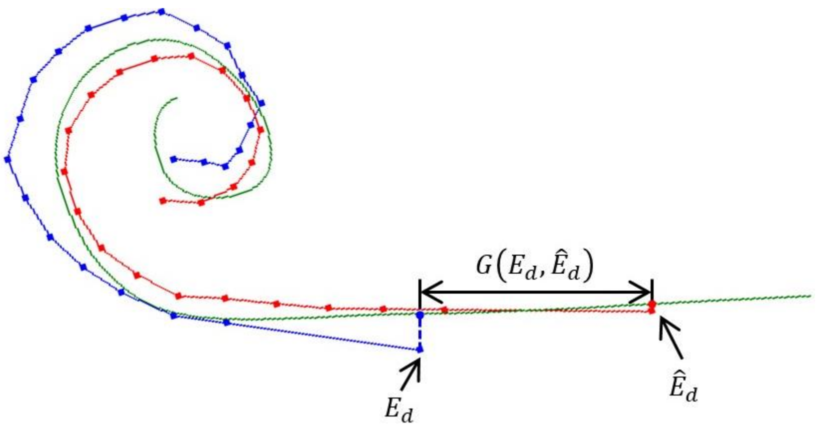


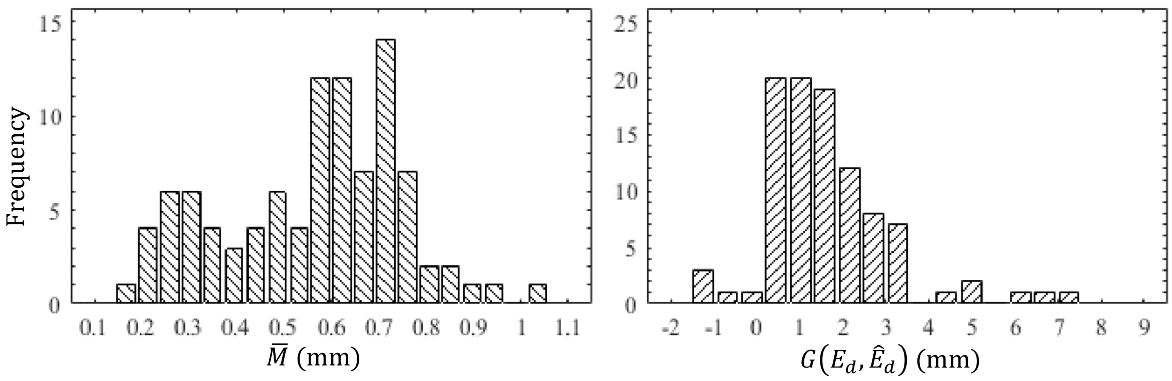
Figure 7.3: The geodesic distance between those two projected points along the modiolar curve is shown as $G(E_d, \hat{E}_d)$.

To evaluate whether the agreement with the recommended insertion depth was associated with more modiolar hugging placement of the electrode array, Pearson correlation coefficients were calculated between $\bar{M}$ and the absolute value of $G(E_d, \hat{E}_d)$, $|G(E_d, \hat{E}_d)|$. A significant positive relationship between $\bar{M}$ and $|G(E_d, \hat{E}_d)|$ indicates that our patient-customized insertion depth would be better at achieving perimodiolar array positioning as compared to the generic insertion depth.

7.3 Results

The mean and standard deviation of $\bar{M}$ were 0.59 ± 0.18mm. A histogram of these data is shown in Figure 7.4a. These data suggest that more tightly modiolar hugging placement of the arrays ($\bar{M}$ < 0.5mm) is not the norm and that the average individual has several electrodes that are substantially distant to the modiolus. For a typical cochlea, the distance an electrode can vary between the modiolar and lateral walls is approximately 1.5mm. Overall, arrays with $\bar{M}$ < 0.5mm can be considered mostly perimodiolar.

Figure 7.4b shows the histogram of $G(E_d, \hat{E}_d)$, which has a mean and standard deviation of 1.44 ± 1.46mm, suggesting that for the majority of individuals (91 out of 97 cases), our technique recommends inserting the array to a depth shallower than what was realized intraoperatively although there is substantial variation around the mean.



(a) (b)

Figure 7.4: (a) Histogram of our modiolar distance results $\{\bar{M}\}$. (b) Histogram of the geodesic distance between recommended and actual insertion depths $\{G(E_d, \hat{E}_d)\}$.

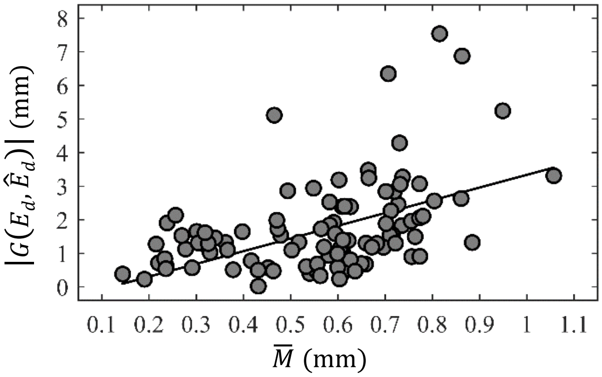


Figure 7.5: Scatter plot of the relationship between $\bar{M}$ and $|G(E_d, \hat{E}_d)|$ with linear regression line shown in black.

Figure 7.5 shows the scatter plot of $|G(E_d, \hat{E}_d)|$ versus $\bar{M}$. The black line shows the linear regression relationship between them. $|G(E_d, \hat{E}_d)|$ was positively correlated with $\bar{M}$ (r = 0.49, p < 0.0001), suggesting that cases with smaller values of $|G(E_d, \hat{E}_d)|$, i.e., cases where the insertion depth matched our recommended depth, had better perimodiolar placement of the array. Figure 7.5 also suggests that achieving mostly modiolar hugging placement ($\bar{M}$ < 0.5mm) only occurs when $|G(E_d, \hat{E}_d)|$ is relatively small although $|G(E_d, \hat{E}_d)|$ is small in many cases where $\bar{M}$ is large. We interpret this finding as follows: an insertion depth close to the patient-customized insertion depth is necessary but not sufficient for perimodiolar placement to occur.

7.4 Conclusions

Although cochlear implants are remarkably successful, a significant number of individuals experience poor outcomes, and restoration to normal fidelity is rare even among the best performers. This is likely due to many variables beyond the control of the surgeon or audiologist (e.g. neural survival), but at least one variable within surgical control is intracochlear placement. As their name suggests, perimodiolar electrodes are designed to be placed close to the modiolus to decrease channel interaction. Our results show that perimodiolar electrode arrays, more often than not, do not sit adjacent to the modiolus where they are likely most effective.

To achieve more optimal perimodiolar positioning, we proposed a simple, image-guided approach to selecting a patient-customized electrode insertion depth based upon the final resting shape of the electrode array and the patient's anatomy as assessed on pre-operative CT scanning. Our findings show a correlation between perimodiolar placement of the array, $\overline{M}$, and the customized insertion depth as assessed by the absolute value of the difference between actual and recommended insertion depth, $\left|G\left(E_d, \hat{E}_d\right)\right|$, suggesting that if our patient-customized electrode insertion depth technique is implemented, it will increase the fraction of cases where perimodiolar placement of the electrode array is achieved. Our results indicate that over-insertion of electrode arrays is the norm and under-insertion uncommon. Furthermore, it is rare that the generically recommended insertion depth is the best one for any individual suggesting that the one size fits all approach is rather a "one size fits few" approach. Even when arrays are inserted precisely to the manufacturer's recommended depth, the over-insertion artifact of a laterally displaced mid-portion of the array is common as shown in the example case of Figure 7.1a. While we have shown that over-insertion of the electrode arrays explains many cases where perimodiolar positioning is not achieved, our data indicate that an insertion depth close to the patient's customized insertion depth is necessary but not sufficient for perimodiolar placement to occur, and thus clearly other factors contribute to achieving perimodiolar placement.

Clinically, our customized insertion depth approach recommends a base insertion depth that is typically shallower than the manufacturer's specified depth. This shallower depth is more likely to position the most basal electrodes closer to the cochlear entrance where they are farther away from neural stimulation sites typically most in need of stimulation. While this may be a potential drawback of our approach, it can be overcome by deactivating these most basal electrodes. Recent studies support this idea as they have shown the benefit of the deactivation of electrodes that are determined to be suboptimally positioned (Noble et al., 2014). With the deactivation of electrodes placed near the entrance of the cochlea, we anticipate that the benefits of better perimodiolar positioning will outweigh this potential drawback, but this hypothesis will need to be verified in future studies. Another potential solution is redesigned electrode arrays with electrodes concentrated in those regions where periomodiolar positioning can be achieved.

Summarizing, we have found that (1) pre-curved electrode arrays tend to be over inserted leading to displacement away from the modiolus, which has been shown by Holden et al. (Holden et al., 2013) to be correlated with poorer hearing outcomes, and (2) we can achieve better perimodiolar positioning using patient-customized insertion depths.

Acknowledgments

This work was supported in part by grants R01DC008408 and R01DC014462 from the National Institute on Deafness and Other Communication Disorders. The content is solely the responsibility of the authors and does not necessarily represent the official views of these institutes.

CHAPTER VIII

Summary and Future Work

This dissertation introduces several innovative machine learning-based techniques for medical image registration and segmentation and a technique for patient-customized placement of cochlear implant electrode arrays. We have made three major contributions: (1) As described in **Chapter II**, we developed a method to initialize non-rigid registration algorithms. (2) As described in **Chapters III**, **IV**, **V**, and **VI**, we developed three methods to segment the intracochlear anatomy in the Post-CT images of cochlear implant recipients. (3) As described in **Chapter VII**, for the first time, we retrospectively evaluated the modiolar positioning of electrode arrays within a large cochlear implant imaging database. We found the rate at which perimodiolar placement is successfully achieved and evaluated a new technique we propose to pre-operatively plan patient-customized electrode array insertion depths to improve perimodiolar placement at the time of surgery.

We believe that our work has made valuable contributions and that it might prove interesting areas for future research. The contributions and potential future research topics related to each chapter are summarized below:

In **Chapter II**, we present a learning-based method to automatically find a set of robust landmarks in T1-weighted MR images of the head to initialize non-rigid registration algorithms. Preregistration initialization can be performed using our two-step approach: First, a set of robust landmarks are localized in the source image and the target image. Second, the initialization transformation between the two images is calculated by registering the two images using the TPS-based transformations with the robust landmarks as control points. We show that, by using our method to compute preregistration initialization, higher registration accuracy can be achieved for 5 well-established deformable registration algorithms, than when a standard preregistration initialization approach is used.

The technique is generic and could be used to initialize non-rigid registration algorithms for other applications. Because this approach operates on principles that are different from most non-rigid registration methods in routine use, it could also be used as an error detection mechanism. In this context, it could be run in parallel with existing processing pipelines and differences observed between methods in either deformation fields or landmark positions could trigger alerts (J. Wang, Liu, et al., 2017), this can be explored in future studies.

In **Chapter III**, we propose an approach based on the cGANs for the reduction of metal artifacts in CT ear images of cochlear implant recipients. To the best of our knowledge, we were the first to use GANs to eliminate or reduce metallic artifacts in CT images. Compared to the current leading traditional methods for reducing artifacts in CT images, which generally require the raw data from CT scanners, our method is a post-reconstruction processing method for which the raw data is not required. Compared to other published machine learning-based methods, which either depend on existing traditional methods or require post-processing of the outputs produced by the machine learning models, ours is unique in being able to synthesize directly an artifact-free image from an image in which artifacts are present (J. Wang et al., 2019). The segmentation of the intracochlear anatomy in Post-CT images can be acquired by using our method with a two-step approach: (1) generate a synthetic Pre-CT image from a Post-CT image with cGANs trained for this purpose and (2) apply SegPre-ASM to the synthetic image. This approach significantly reduces the segmentation error for intracochlear anatomy in Post-CT images when compared to the previous state of the art. The effect of our method on the final results of our IGCIP system has been evaluated in **Chapter IV**. Results show that more than 80% of the implant configurations generated by our IGCIP system using our approach to produce the segmentation of the intracochlear anatomy can be used for the programming of the cochlear implants, and these configurations are likely to lead to hearing outcomes that are comparable to those achieved using the best possible configurations. The success rate we have achieved is good, which confirms that our techniques could benefit patients for whom Pre-CT images are unavailable. Our method for the reduction of metal artifacts in CT images has been integrated into our IGCIP system and is in routine use at our institution.

In **Chapter V**, we propose an end-to-end semantic segmentation approach for segmenting the intracochlear anatomy in Post-CT images. We use a multi-resolution two-branch deep network to perform artifact reduction and produce segmentation masks of the intracochlear anatomy simultaneously. Our method has achieved good outcomes. Compared to the two-step method in **Chapter III**, the proposed method can segment the Post-CT images in one step. Our multi-resolution approach limits the memory usage for training while generating segmentation masks at a high resolution. Our results also confirm the usefulness of multi-task learning (J. Wang et al., 2020). The ideas of multi-resolution networks and multi-task learning are generic, they can be applied to other research tasks in the future.

In **Chapter VI**, we propose an atlas-based method to segment the intracochlear anatomy in the Post-CT images. We first generate a DDF between an artifact-free atlas image and a Post-CT image by using registration networks trained for this purpose. The segmentation of the intracochlear anatomy in the Post-CT image can then be obtained by transferring the predefined segmentation meshes of the intracochlear anatomy in the atlas image to the Post-CT image using that DDF. This method preserves the point-to-point correspondence between the meshes in the atlas image and the segmented volumes that is useful for the programming of the implant, and it achieves results that are comparable to those obtained with the current state-of-the-art method in terms of segmentation accuracy. The current state of the art, as described earlier, comprises two steps: (1) generate a synthetic Pre-CT image from a Post-CT image with cGANs and (2) segment the synthetic image by using SegPre-ASM. The second step requires accurate registration of an atlas to the image to be segmented to initialize the active shape model. The first step takes about 0.3s while the second step takes on average 75s. Our atlas-based method only requires providing a volume that includes the inner ear to the trained registration networks, and the inference time is also about 0.3s. Segmentation is thus essentially instantaneous with our atlas-based method while it takes over a minute with the current state of the art. This is important for clinical deployment and end-user acceptance (J. Wang et al., 2021).

The labyrinth is the innermost part of the ear. It externally bounds the intracochlear anatomy and it includes three semicircular canals, which are a complex set of fluid-filled channels that contributes to the sense of balance. Our current atlas-based method can only segment the cochlea, it would be interesting to extend the method to segment the full labyrinth because understanding changes and variations of the labyrinth can help diagnose and predict a number of conditions such as inflammatory and neoplastic processes (Vaidyanathan et al., 2021).

The models proposed in **Chapters III**, **V**, and **VI** are based on the conventional convolutional neural networks, recently attention-based models are proposed and have achieved success in image processing tasks including medical image segmentation (Oktay et al., 2018; Sinha & Dolz, 2019). It might be interesting to explore whether using attention-based models can improve the segmentation accuracy of the intracochlear anatomy. Prior knowledge such as shape priors and neighborhood relation between vertices on the segmentation meshes have not been provided to our models during the training. It might be helpful to incorporate such prior knowledge into the training objective functions in the future. Besides, all of our models are trained through supervised learning, although accurate, paired Pre-CT and Post-CT images are always required, which may limit the generalizability of our methods. It might be interesting to explore unsupervised approaches such as methods based on the Cycle-Consistent Adversarial Networks (Zhu et al., 2017) or the Artifact Disentanglement Network (Liao et al., 2020) to solve our problems.

In **Chapter VII**, we retrospectively evaluate the positions of pre-curved cochlear implant electrodes on a large cohort. We also develop a preoperatively planning technique to determine an optimal patient-customized insertion depth for the pre-curved cochlear implant electrode array. First, a geometric model of the electrode array is aligned with the scala tympani of the patient, such that the array positioning best agrees with the inner wall of the scala tympani. We then determine the patient-customized insertion base depth as the depth of the base of the optimal electrode position curve. We have found that (1) pre-curved electrode arrays tend to be over inserted leading to displacement away from the modiolus in the first turn of the cochlea, and that (2) this lateral displacement leads to worse hearing outcomes, and that (3) we can achieve better perimodiolar positioning using the patient-customized insertion depths. We recognize that our report is retrospective in nature and that prospective studies will be necessary to substantiate these findings (J. Wang, Dawant, et al., 2017).

REFERENCES

Ardekani, B. A., Guckemus, S., Bachman, A., Hoptman, M. J., Wojtaszek, M. & Nierenberg, J. (2005). Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans. *Journal of Neuroscience Methods*, *142*(1), 67–76. https://doi.org/10.1016/j.jneumeth.2004.07.014

Aschendorff, A., Kubalek, R., Turowski, B., Zanella, F., Hochmuth, A., Schumacher, M., Klenzner, T. & Laszig, R. (2005). Quality control after cochlear implant surgery by means of rotational tomography. *Otology & Neurotology*, *26*(1), 34–37. https://doi.org/10.1097/00129492-200501000-00007

Avants, B. B., Epstein, C. L., Grossman, M. C. & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Banalagay, R. A., Labadie, R. F., Chakravorti, S. & Noble, J. H. (2020). Insertion depth for optimized positioning of precurved cochlear implant electrodes. *Otology & Neurotology*, *41*(8), 1066–1071. https://doi.org/10.1097/MAO.0000000000002726

Besl, P. J. & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 239–256. https://doi.org/10.1109/34.121791

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buss, E., Pillsbury, H. C., Buchman, C. A., Pillsbury, C. H., Clark, M. S., Haynes, D. S., Labadie, R. F., Amberg, S., Roland, P. S., Kruger, P., Novak, M. A., Wirth, J. A., Black, J. M., Peters, R., Lake, J., Wackym, P. A., Firszt, J. B., Wilson, B. S., Lawson, D. T., … Barco, A. L. (2008). Multicenter U.S. bilateral MED-EL cochlear implantation study: Speech perception over the first year of use. *Ear and Hearing*, *29*(1), 20–32. https://doi.org/10.1097/AUD.0b013e31815d7467

Christensen, G. E. & Johnson, H. J. (2001). Consistent image registration. *IEEE Transactions on Medical Imaging*, *20*(7), 568–582. https://doi.org/10.1109/42.932742

Cootes, T. F., Taylor, C. J., Cooper, D. H. & Graham, J. R. (1995). Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, *61*(1), 38–59. https://doi.org/10.1006/cviu.1995.1004

D'Haese, P.-F., Cetinkaya, E., Konrad, P. E., Kao, C. & Dawant, B. M. (2005). Computer-aided placement of deep brain stimulators: From planning to intraoperative guidance. *IEEE Transactions on Medical Imaging*, *24*(11), 1469–1478. https://doi.org/10.1109/TMI.2005.856752

D'Haese, P.-F., Pallavaram, S., Li, R., Remple, M. S., Kao, C., Neimat, J. S., Konrad, P. E. & Dawant, B. M. (2012). CranialVault and its CRAVE tools: A clinical computer assistance system for deep brain stimulation (DBS) therapy. *Medical Image Analysis*, *16*(3), 744–753. https://doi.org/10.1016/j.media.2010.07.009

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements (1st ed.).* Springer-Verlag New York. https://doi.org/10.1007/b97287

Dorman, M. F., Yost, W. A., Wilson, B. S. & Gifford, R. H. (2011). Speech perception and sound localization by adults with bilateral cochlear implants. *Seminars in Hearing*, *32*(2), 212–214. https://doi.org/10.1055/s-0031-1278417

Fan, Y., Zhang, D., Wang, J., Noble, J. H. & Dawant, B. M. (2020). Combining model- and deep-learning-based methods for the accurate and robust segmentation of the intra-cochlear anatomy in clinical head CT images. *Proc. SPIE 11313, Medical Imaging 2020: Image Processing*, 113131D. https://doi.org/10.1117/12.2549390

Finley, C. C., Holden, T. A., Holden, L. K., Whiting, B. R., Chole, R. A., Neely, G. J., Hullar, T. E. & Skinner, M. W. (2008). Role of electrode placement as a contributor to variability in cochlear implant outcomes. *Otology & Neurotology*, *29*(7), 920–928. https://doi.org/10.1097/MAO.0b013e318184f492

Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395. https://doi.org/10.1145/358669.358692

Ghavami, N., Hu, Y., Bonmati, E., Rodell, R., Gibson, E., Moore, C. & Barratt, D. (2018). Automatic slice segmentation of intraoperative transrectal ultrasound images using convolutional neural networks. *Proc. SPIE 10576, Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, *1057603*. https://doi.org/10.1117/12.2293300

Gifford, R. H., Dorman, M. F., Sheffield, S. W., Teece, K. & Olund, A. P. (2014). Availability of binaural cues for bilateral implant recipients and bimodal listeners with and without preserved hearing in the implanted ear. *Audiology and Neurotology*, *19*(1), 57–71. https://doi.org/10.1159/000355700

Gifford, R. H., Shallop, J. K. & Peterson, A. M. (2008). Speech recognition materials and ceiling effects:

Considerations for cochlear implant programs. *Audiology and Neurotology*, *13*(3), 193–205. https://doi.org/10.1159/000113510

Gjesteby, L., De Man, B., Jin, Y., Paganetti, H., Verburg, J., Giantsoudi, D. & Wang, G. (2016). Metal artifact reduction in CT: Where are we after four decades? *IEEE Access*, *4*, 5826–5849. https://doi.org/10.1109/ACCESS.2016.2608621

Gjesteby, L., Yang, Q., Xi, Y., Claus, B., Jin, Y., Man, B. De & Wang, G. (2017). Reducing Metal Streak Artifacts in CT Images via Deep Learning: Pilot Results. *Fully3D 2017 Proceedings*. https://doi.org/10.12059/Fully3D.2017-11-3202009

Glocker, B., Feulner, J., Criminisi, A., Haynor, D. R. & Konukoglu, E. (2012). Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012. MICCAI 2012. Lecture Notes in Computer Science*, *7512*, 590–598. https://doi.org/10.1007/978-3-642-33454-2_73

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Goshtasby, A. (1988). Registration of images with geometric distortions. *IEEE Transactions on Geoscience and Remote Sensing*, *26*(1), 60–64. https://doi.org/10.1109/36.3000

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–2605. https://doi.org/10.1121/1.399052

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, *29*(2), 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Han, D., Gao, Y., Wu, G., Yap, P.-T. & Shen, D. (2015). Robust anatomical landmark detection with application to MR brain image registration. *Computerized Medical Imaging and Graphics*, *46*, 277–290. https://doi.org/10.1016/j.compmedimag.2015.09.002

Hardy, M. (1938). The length of the organ of Corti in man. *American Journal of Anatomy*, *62*(2), 291–311. https://doi.org/10.1002/aja.1000620204

He, K., Zhang, X., Ren, S. & Sun, J. (2015). Deep residual learning for image recognition. Retrieved from https://arxiv.org/abs/1512.03385

Holden, L. K., Finley, C. C., Firszt, J. B., Holden, T. A., Brenner, C., Potts, L. G., Gotter, B. D., Vanderhoof, S. S., Mispagel, K., Heydebrand, G. & Skinner, M. W. (2013). Factors affecting open-set word recognition in adults with cochlear implants. *Ear and Hearing*, *34*(3), 342–360. https://doi.org/10.1097/AUD.0b013e3182741aa7

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C. M., Emberton, M., Noble, J. A., Barratt, D. C. & Vercauteren, T. (2018). Label-driven weakly-supervised learning for multimodal deformable image registration. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 1070–1074. https://doi.org/10.1109/ISBI.2018.8363756

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M., Ourselin, S., Noble, J. A., Barratt, D. C. & Vercauteren, T. (2018). Weakly-supervised convolutional neural networks for multimodal image registration. *Medical Image Analysis*, *49*, 1–13. https://doi.org/10.1016/j.media.2018.07.002

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Kahn, E., D’Haese, P.-F., Dawant, B., Allen, L., Kao, C., Charles, P. D. & Konrad, P. (2012). Deep brain stimulation in early stage Parkinson’s disease: Operative experience from a prospective randomised clinical trial. *Journal of Neurology, Neurosurgery & Psychiatry*, *83*(2), 164–170. https://doi.org/10.1136/jnnp-2011-300008

Kim, B., Kim, J., Lee, J. G., Kim, D. H., Park, S. H. & Ye, J. C. (2019). Unsupervised deformable image registration using cycle-consistent CNN. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science*, *11769*, 166–174. https://doi.org/10.1007/978-3-030-32226-7_19

Kingma, D. P. & Ba, J. L. (2014). Adam: A Method for stochastic optimization. Retrieved from https://arxiv.org/abs/1412.6980

Liao, H., Lin, W. A., Zhou, S. K. & Luo, J. (2020). ADN: Artifact disentanglement network for unsupervised metal artifact reduction. *IEEE Transactions on Medical Imaging*, *39*(3), 634–643.

https://doi.org/10.1109/TMI.2019.2933425

Litovsky, R., Parkinson, A., Arcaroli, J. & Sammeth, C. (2006). Simultaneous bilateral cochlear implantation in adults: A multicenter clinical study. *Ear and Hearing, 27*(6), 714–731. https://doi.org/10.1097/01.aud.0000246816.50820.42

Liu, Y., D’Haese, P.-F. & Dawant, B. M. (2014). Effects of deformable registration algorithms on the creation of statistical maps for preoperative targeting in deep brain stimulation procedures. *Proc. SPIE 9036, Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, 90362B. https://doi.org/10.1117/12.2043529

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G. & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, *16*(2), 187–198. https://doi.org/10.1109/42.563664

Maintz, J. B. A. & Viergever, M. A. (1998). A survey of medical image registration. *Medical Image Analysis*, *2*(1), 1–36. https://doi.org/10.1.1.39.4417

Mertes, J. & Chinnici, J. (2006). *Cochlear implants — considerations in programming for the pediatric population*. https://www.audiologyonline.com/articles/cochlear-implants-considerations-in-programming-1011

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571. https://doi.org/10.1109/3DV.2016.79

Mirza, M. & Osindero, S. (2014). Conditional generative adversarial nets. Retrieved from https://arxiv.org/abs/1411.1784

Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C. & Ourselin, S. (2010). Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine, 98*(3), 278–284. https://doi.org/10.1016/j.cmpb.2009.09.002

National Institute on Deafness and Other Communication Disorders. (2016). *Cochlear implants*. NIH Publication No. 00-4798. https://www.nidcd.nih.gov/sites/default/files/Documents/cochlear-implants.pdf

Noble, J. H. (2017). *Image-guided Cochlear Implant Programming (IGCIP)*. https://clinicaltrials.gov/ct2/show/NCT03306082

Noble, J. H. & Dawant, B. M. (2015). Automatic graph-based localization of cochlear implant electrodes in CT.

Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, *9350*, 152–159. https://doi.org/10.1007/978-3-319-24571-3_19

Noble, J. H., Gifford, R. H., Hedley-Williams, A. J., Dawant, B. M. & Labadie, R. F. (2014). Clinical evaluation of an image-guided cochlear implant programming strategy. *Audiology and Neurotology*, *19*(6), 400–411. https://doi.org/10.1159/000365273

Noble, J. H., Hedley-Williams, A. J., Sunderhaus, L., Dawant, B. M., Labadie, R. F., Camarata, S. M. & Gifford, R. H. (2016). Initial results with image-guided cochlear implant programming in children. *Otology & Neurotology*, *37*(2), e63–e69. https://doi.org/10.1097/MAO.0000000000000909

Noble, J. H., Labadie, R. F., Gifford, R. H. & Dawant, B. M. (2013). Image-guidance enables new methods for customizing cochlear implant stimulation strategies. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *21*(5), 820–829. https://doi.org/10.1109/TNSRE.2013.2253333

Noble, J. H., Labadie, R. F., Majdani, O. & Dawant, B. M. (2011). Automatic segmentation of intracochlear anatomy in conventional CT. *IEEE Transactions on Biomedical Engineering*, *58*(9), 2625–2632. https://doi.org/10.1109/TBME.2011.2160262

Nordfalk, K. F., Rasmussen, K., Hopp, E., Greisiger, R. & Jablonski, & G. E. (2014). Scalar position in cochlear implant surgery and outcome in residual hearing and the vestibular system. *International Journal of Audiology*, *53*(2), 121–127. https://doi.org/10.3109/14992027.2013.854413

Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B. & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. Retrieved from https://arxiv.org/abs/1804.03999

Park, H. S., Lee, S. M., Kim, H. P. & Seo, J. K. (2017). Machine-learning-based nonlinear decomposition of CT images for metal artifact reduction. Retrieved from https://arxiv.org/abs/1708.00244

Pauly, O., Glocker, B., Criminisi, A., Mateus, D., Möller, A. M., Nekolla, S. & Navab, N. (2011). Fast multiple organ detection and localization in whole-body MR Dixon sequences. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011. MICCAI 2011. Lecture Notes in Computer Science*, *6893*, 239–247. https://doi.org/10.1007/978-3-642-23626-6_30

Pelosi, S., Noble, J. H., Dawant, B. M. & Labadie, R. F. (2013). Analysis of intersubject variations in intracochlear and middle ear surface anatomy for cochlear implantation. *Otology & Neurotology*, *34*(9), 1675–1680.

https://doi.org/10.1097/MAO.0b013e3182a1a7e6

Peterson, G. E. & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders*, *27*(1), 62–70. https://doi.org/10.1044/jshd.2701.62

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing (2nd ed.)*. Cambridge University Press.

Reda, F. A., McRackan, T. R., Labadie, R. F., Dawant, B. M. & Noble, J. H. (2014). Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients. *Medical Image Analysis*, *18*(3), 605–615. https://doi.org/10.1016/j.media.2014.02.001

Reda, F. A., Noble, J. H., Labadie, R. F. & Dawant, B. M. (2014). An artifact-robust, shape library-based algorithm for automatic segmentation of inner ear anatomy in post-cochlear-implantation CT. *Proc. SPIE 9034, Medical Imaging 2014: Image Processing*, 90342V. https://doi.org/10.1117/12.2043260

Rohde, G. K., Aldroubi, A. & Dawant, B. M. (2003). The adaptive bases algorithm for intensity-based nonrigid image registration. *IEEE Transactions on Medical Imaging*, *22*(11), 1470–1479. https://doi.org/10.1109/TMI.2003.819299

Rohr, K., Stiehl, H. S., Sprengel, R., Buzug, T. M., Weese, J. & Kuhn, M. H. (2001). Landmark-based elastic registration using approximating thin-plate splines. *IEEE Transactions on Medical Imaging*, *20*(6), 526–534. https://doi.org/10.1109/42.929618

Roland, P. S. & Wright, C. G. (2006). Surgical aspects of cochlear implantation: Mechanisms of insertional trauma. *Advances in Oto-Rhino-Laryngology*, *64*, 11–30. https://doi.org/10.1159/000094642

Rubinstein, J. T. (2004). How cochlear implants encode speech. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *12*(5), 444–448. https://doi.org/10.1097/01.moo.0000134452.24819.c0

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O. & Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, *18*(8), 712–721. https://doi.org/10.1109/42.796284

Sinha, A. & Dolz, J. (2019). Multi-scale self-guided attention for medical image segmentation. Retrieved from https://arxiv.org/abs/1906.02849

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Sørensen, T. J. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter*, *5*, 1–34.

Sotiras, A., Davatzikos, C. & Paragios, N. (2013). Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, *32*(7), 1153–1190. https://doi.org/10.1109/TMI.2013.2265603

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2017, ML-CDS 2017. Lecture Notes in Computer Science*, *10553*, 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

Vaerenberg, B., Smits, C., De Ceulaer, G., Zir, E., Harman, S., Jaspers, N., Tam, Y., Dillon, M., Wesarg, T., Martin-Bonniot, D., Gärtner, L., Cozma, S., Kosaner, J., Prentiss, S., Sasidharan, P., Briaire, J. J., Bradley, J., Debruyne, J., Hollow, R., … Govaerts, P. J. (2014). Cochlear implant programming: A global survey on the state of the art. *The Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/501738

Vaidyanathan, A., van der Lubbe, M. F. J. A., Leijenaar, R. T. H., van Hoof, M., Zerka, F., Miraglio, B., Primakov, S., Postma, A. A., Bruintjes, T. D., Bilderbeek, M. A. L., Sebastiaan, H., Dammeijer, P. F. M., van Rompaey, V., Woodruff, H. C., Vos, W., Walsh, S., van de Berg, R. & Lambin, P. (2021). Deep learning for the fully automated segmentation of the inner ear on MRI. *Scientific Reports*, *11*(1), 1–14. https://doi.org/10.1038/s41598-021-82289-y

Vercauteren, T., Pennec, X., Perchant, A. & Ayache, N. (2009). Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, *45*(1), S61–S72. https://doi.org/10.1016/j.neuroimage.2008.10.040

Viola, P. & Wells, W. M. (1995). Alignment by maximization of mutual information. *Proceedings of IEEE International Conference on Computer Vision*, 16–23. https://doi.org/10.1109/ICCV.1995.466930

Wang, J., Dawant, B. M., Labadie, R. F. & Noble, J. H. (2017). Retrospective evaluation of a technique for patient-customized placement of precurved cochlear implant electrode arrays. *Otolaryngology–Head and Neck Surgery*, *157*(1), 107–112. https://doi.org/10.1177/0194599817697298

Wang, J., Liu, Y., Noble, J. H. & Dawant, B. M. (2017). Automatic selection of landmarks in T1-weighted head MRI with regression forests for image registration initialization. *Journal of Medical Imaging*, *4*(4), 044005. https://doi.org/10.1117/1.JMI.4.4.044005

Wang, J., Noble, J. H. & Dawant, B. M. (2019). Metal artifact reduction for the segmentation of the intra cochlear anatomy in CT images of the ear with 3D-conditional GANs. *Medical Image Analysis*, *58*, 101553. https://doi.org/10.1016/j.media.2019.101553

Wang, J., Noble, J. H. & Dawant, B. M. (2020). Metal artifact reduction and intra cochlear anatomy segmentation in CT images of the ear with a multi-resolution multi-task 3D network. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI),* 596–599. https://doi.org/10.1109/ISBI45749.2020.9098707

Wang, J., Su, D., Fan, Y., Chakravorti, S., Noble, J. H. & Dawant, B. M. (2021). Atlas-based segmentation of intracochlear anatomy in metal artifact affected CT images of the ear with co-trained deep neural networks. Retrieved from https://arxiv.org/abs/2107.03987

Wang, J., Zhao, Y., Noble, J. H. & Dawant, B. M. (2018). Conditional generative adversarial networks for metal artifact reduction in CT images of the ear. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science*, *11070*, 3–11. https://doi.org/10.1007/978-3-030-00928-1_1

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wanna, G. B., Noble, J. H., Carlson, M. L., Gifford, R. H., Dietrich, M. S., Haynes, D. S., Dawant, B. M. & Labadie, R. F. (2014). Impact of electrode design and surgical approach on scalar location and cochlear implant outcomes. *Laryngoscope*, *124*(S6), S1–S7. https://doi.org/10.1002/lary.24728

Wanna, G. B., Noble, J. H., Gifford, R. H., Dietrich, M. S., Sweeney, A. D., Zhang, D., Dawant, B. M., Rivas, A. & Labadie, R. F. (2015). Impact of intrascalar electrode location, electrode type, and angular insertion depth on residual hearing in cochlear implant patients: Preliminary results. *Otology & Neurotology*, *36*(8), 1343–1348. https://doi.org/10.1097/MAO.0000000000000829

Welch, B. L. (1947). The generalisation of student's problems when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

Wells, W. M., Viola, P., Atsumi, H., Nakajima, S. & Kikinis, R. (1996). Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, *1*(1), 35–51. https://doi.org/10.1016/S1361-8415(01)80004-9

Yi, X., Walia, E. & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, *58*, 101552. https://doi.org/https://doi.org/10.1016/j.media.2019.101552

Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J. & Bourgeat, P. (2019). Ea-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis. *IEEE Transactions on Medical Imaging*, *38*(7), 1750–1762. https://doi.org/10.1109/TMI.2019.2895894

Zhang, D., Banalagay, R., Wang, J., Zhao, Y., Noble, J. H. & Dawant, B. M. (2019). Two-level training of a 3D U-Net for accurate segmentation of the intra-cochlear anatomy in head CTs with limited ground truth training data. *Proc. SPIE 10949, Medical Imaging 2019: Image Processing*, 1094907. https://doi.org/10.1117/12.2512529

Zhang, Yanbo & Yu, H. (2018). Convolutional neural network based metal artifact reduction in X-ray computed tomography. *IEEE Transactions on Medical Imaging*, *37*(6), 1370–1381. https://doi.org/10.1109/TMI.2018.2823083

Zhang, Yongyue, Brady, M. & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. https://doi.org/10.1109/42.906424

Zhao, Y., Chakravorti, S., Labadie, R. F., Dawant, B. M. & Noble, J. H. (2019). Automatic graph-based method for localization of cochlear implant electrode arrays in clinical CT with sub-voxel accuracy. *Medical Image Analysis*, *52*, 1–12. https://doi.org/10.1016/j.media.2018.11.005

Zhao, Y., Dawant, B. M., Labadie, R. F. & Noble, J. H. (2018). Automatic localization of closely spaced cochlear implant electrode arrays in clinical CTs. *Medical Physics*, *45*(11), 5030–5040. https://doi.org/10.1002/mp.13185

Zhao, Y., Dawant, B. M., Labadie, R. F. & Noble, J. H. (2014). Automatic localization of cochlear implant electrodes in CT. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. MICCAI 2014. Lecture Notes in Computer Science*, *8673*, 331–338. https://doi.org/10.1007/978-3-319-10404-1_42

Zhao, Y., Dawant, B. M. & Noble, J. H. (2016). Automatic selection of the active electrode set for image-guided cochlear implant programming. *Journal of Medical Imaging*, *3*(3), 035001. https://doi.org/10.1117/1.JMI.3.3.035001

Zhao, Y., Dawant, B. M. & Noble, J. H. (2017). Automatic localization of cochlear implant electrodes in CTs with a limited intensity range. *Proc. SPIE 10133, Medical Imaging 2017: Image Processing*, 101330T. https://doi.org/10.1117/12.2254569

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244