Deep generative learning for medical data processing, analysis and modeling: application to cochlea ct imaging

Zihao Wang

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Apprentissage Génératif pour le Traitement, L'analyse et la Modélisation des Données Médicales:

application à l'imagerie CT de la cochlée

Zihao WANG

Inria, Équipe Epione

Dirigée par : Hervé Delingette

Présentée en vue de l'obtention du grade de Docteur en Automatique, Traitement du Signal et des Images de l'Université Côte d'Azur.

Soutenue le: le 17 septembre 2021

Devant le jury composé de :

Mauricio Reyes	University of Bern	Rapporteur
Rasmus R. Paulsen	Technical University of Denmark	Rapporteur
Nicolas Guevara	Centre Hospitalier Universitaire de Nice	Président
Hervé Delingette	Inria	Directeur de thèse
Alain Lalande	Université de Bourgogne	Invité
Maxime Sermesant	Inria	Invité
François Patou	Oticon Medical	Invité

Une thèse préparée à Inria

Inria

Page intentionally left blank.

DEEP GENERATIVE LEARNING FOR MEDICAL DATA PROCESSING, ANALYSIS AND MODELING

Application to cochlea CT imaging.

Zihao WANG

A thesis presented for the degree of
Doctor of Philosophy

L'Ecole Doctorale STIC
Inria-ComUE Université Côte d'Azur
Sophia Antipolis, France
2021

Page intentionally left blank.

大道至简

There is no spoon.

Page intentionally left blank.

Acknowledgments

I would like to thank Hervé Delingette, whom I have learned with for the past wonderful three years. The formulas on the whiteboard, the revision marks on the paper, and the 448 correspondences emails all recorded his efforts for helping me to improve in academics. What Hervé taught me is more than academic and professional skills, in addition to the rigorous attitude for research, the inspiration for philosophy, and the view of life. I would like to thank Nicholas Ayache, who created this excellent team and led the team to be a pioneer in the domain. I also appreciate his expectations and motivation for me to improve my French and my success in research. Thanks to Maxime Sermesant, who affirmed my academic process as a member of my comité de suivi in the past years, it is also my honor to have a chance to learn from him. Thank you, Marco, for delivering me your personal experiences in career development. Your encouragement inspired me a lot for both career planning and learning progress. I like to thank Xavier for his support for me to cope with the administration's procedures in EDSTIC, I admire his straightforwardness and determination as a mathematician. Thanks, Isabelle, who saved me from the vast array of administrative documents at Inria.

Many thanks to my dear colleagues: Pawel, Raphaël, Julian, Shuman, Wen, Yann T., Yann F., Nicolas G., Nicholas C., Clement, Jaume, Luigi, Thomas, Sara, Clair, Etrit, Irene, Hind, Paul, Buntheng, Benoit, Florent, Santiago, Yingyu, Gaëtan, Jairo, Hari, Haris, Tania, Qiao, Fanny, Hui, Chuan, Tingting, also Dai, Hao and Yaohui in STAR team, and many others for the great time we have shared together. I will always remember these beautiful moments: the beach volleyball when I came to Inria, the first time team dinner, the party at different colleagues' homes, learning skiing with everyone when we were in

Auron, and the dialogues with each other in the lab. Thank you all for making my Ph.D. career such a wonderful journey.

I like to thank my thesis reviewers, Dr. Rasmus R. Paulsen and Dr. Mauricio Reyes, for reviewing my thesis to help me to improve the chapters. I want to thank the other jury members and invited members: Dr. MD. Nicolas Guevara, Dr. Alain Lalande, Dr. Charles Raffaell, Dr. Wilhelm Wilhelm, and Dr. François Patou, for their time and presence on my Ph.D. defense. I like to thank Dr. Miguel A. González Ballester in UPF, who is also one of my comité de suivi members. I appreciate my mentors Dr. Xavier Hilaire and Dr. Yasmina Abdeddaïm in ESIEE Paris-ComUE Université Gustave Eiffel, thank them for their encouragement and care when I came to a foreign country. Thanks to my mentors: Dr. Sundermann Dietmar (GE), Xavier Boullier (GE), and Lear Liang (Philips), who brought me into the healthcare industry.

I would like to thank my wife Jia, thanks for her understanding and loves gave to me in the past years. This thesis cannot be finished without her supports. Also thanks to my son Hengheng, who made me learn responsibility in life. Thanks to my family to support me in my life, my growth is inseparable from your support. Each line of this thesis has a silent contribution of yours behind it.

Special thanks to the funding from the French government through the "Investments in the Future" project managed by the National Research Agency: ANR-15-IDEX-01, the funding of the regional council of Provence Alpes Cote d'Azur, and the funding from Oticon Medical.

Résumé

Cette thèse vise à exposer plusieurs applications de l'intelligence artificielle (IA) pour le traitement et la compréhension des données médicales. L'imagerie médicale est un domaine générant des données massives, qui nécessitent donc de plus en plus de temps aux cliniciencs pour être traitées et analysées. Dans ce manuscrit, nous montrons comment l'apprentissage génératif peut aider dans de nombreux aspects pour le traitement, la compréhension et la modélisation des images scanner de l'oreille interne.

Tout d'abord, nous développons un modèle génératif profond pour résoudre un problème couramment rencontré en imagerie CT : la présence d'artefacts métalliques. Ce modèle peut permettre aux cliniciens de mieux évaluer la qualité du positionnement des électrodes d'un implant cochléaire avec une présence réduite d'artefacts. Pour cela, un réseau de neurones antagoniste et génératif (GAN) est proposé intégrant une fonction de perte spécifique. Ce réseau a été entrainé sur un ensemble d'images volumiques scanner synthétiques résultant de l'application de simulations de la physique des rayons X.

Deuxièmement, étant donné que de nombreuses méthodes de segmentation d'apprentissage profond ne parviennent pas à gérer explicitement les modèles de forme, nous proposons un cadre génératif bayésien qui aborde les problèmes d'inférence de modèle de forme dans les images médicales 3D. Notre approche permet de faire un compromis entre les informations de forme et d'apparence issus de l'image à travers une approche d'espérance-maximisation (EM). Celle-ci est appliquée à la segmentation de plus de 200 volumes tomodensitométriques de patients. Les résultats indiquent des performances comparables aux méthodes supervisées et meilleures que les méthodes non supervisées proposées précédemment. En outre, nous montrons comment ce cadre méthodologique proposé peut estimer l'incertitude dans les paramètres de forme.

Troisièmement, nous abordons le problème de la représentation compacte des images scanner à travers un nouveau réseau génératif profond basé sur les flux. Les modèles génératifs peuvent créer une distribution implicite de l'ensemble de données d'imagerie à partir duquel on peut générer des échantillons. Pour une meilleure représentation, nous avons proposé un Autoencodeur Variationnel Quasi-

symplectique avec une dynamique de Langevin (Langevin-VAE) qui améliore les gradients actuels des modèles génératifs basés sur les flux.

Enfin, nous proposons une méthode pour la détection de points caractéristiques qui permet de s'affranchir de la difficulté de positionner manuellement ces points dans des images volumiques. Cette approche comprend une étape préalable d'apprentissage ne nécessitant qu'une seule image annotée pour l'entrainement. Elle est appliquée à l'annotation de centaines d'images scanner de l'oreille interne.

Mots-clés: Apprentissage génératif, Apprentissage Bayésien, Flux stochastique, Apprentissage Profond

Abstract

This thesis aims to expose several applications of artificial intelligence (AI) for medical data processing and understanding. Medical imaging is a domain generating massive data, which thus requires more and more time for clinicians to process and analyze them. In this manuscript, we show how generative learning can help in many aspects of the processing, understanding, and modeling of CT images of the inner ear.

First, we develop a deep generative model to solve a commonly encountered problem in CT imaging: the presence of metal artifacts. This model may allow clinicians to better assess the quality of cochlea implant (CI) positioning with a reduced presence of artifacts. To this end, a generative adversarial neural network (GAN) framework equipped with a specially designed loss function is proposed. That network was trained on a synthetic CT volume dataset resulting from the application of X-ray physics simulations.

Second, since many deep learning segmentation methods fail to cope with explicit shape representations, we propose a Bayesian generative framework that addresses the issues of shape model inference in 3D images. We focus on the balance between shape and appearance through an Expectation-Maximisation (EM) approach. The method is applied to the segmentation of more than 200 patient CT volumes. The results show performances that are comparable to supervised methods and better than previously proposed unsupervised ones. Besides, we show how the proposed framework can estimate the uncertainty in the shape parameters.

Third, we tackle the issue of the compact representation of CT images through a novel flow-based deep generative network. Generative models can create an implicit distribution of the imaging dataset from which one can generate samples. For a better representation, we proposed a Quasi-symplectic Langevin Variational Autoencoder (Langevin-VAE) that improves the current gradients, flow-based generative models.

Finally, we propose an online framework for medical landmarks detection that can cope with the difficulty to manually position landmarks in volumetric images. The one shot training framework includes an offline step that only requires a single annotated image for training and is applied to the annotation of hundreds

of images.

Keywords: Generative Learning, Bayesian Learning, Stochastic Flow, Deep Learning

List of Figures

List of Tables

Chapter 1

Introduction

The recent success of deep learning in computer vision promotes also the blooming of research on artificial intelligence (AI) in healthcare. Performances that exceed the level of human expertise are constantly being reported in public AI competitions. Yet there is still a lot of discussion about the safety and robustness of medical AI applications in medicine. This thesis is not focused on the comparison of performances between medical doctors and algorithms, but on helping the cooperation between humans and machines around clinical data. From a collection of human auditory system CT imaging datasets, we explored the feasibility and effectiveness of AI's practical applications in different aspects. First, we introduce the clinical background of the thesis, the cochlea, and CT imaging. We then present some machine learning concepts about generative learning models which is the main theoretical basis of the thesis. The application of those methods in otology is present in all of this thesis. The organization of the work and contributions are given in the last section of this chapter.

1.1 Clinical context

1.1.1 Auditory system

If life is a magnificent symphony, then the auditory system is a beautiful movement created by nature. Auditory perception is one of the main senses of humans for environmental interaction. Hearing-impaired illnesses are commonly attributed to the abnormal function of the hearing system, a small part is related to nervous system problems or psychological disorders. The structure of human auditory system can be roughly classified as three different parts: outer ear, middle ear and inner ear. Fig. 1.1 shows a sketch of human auditory system. The sound vibrations are amplified by the outer ear section between $3000Hz$ and $12000Hz$. The vibrations signals are collected by the eardrum and passed to the inner ear through a series of structured bones. In the middle ear, the sound waves are transferred as mechanical vibrations through the sophisticated bones combination. Further, the mechanical vibrations are passed to the inner ear and converted to nervous electric signals for for further processing inside the brain.

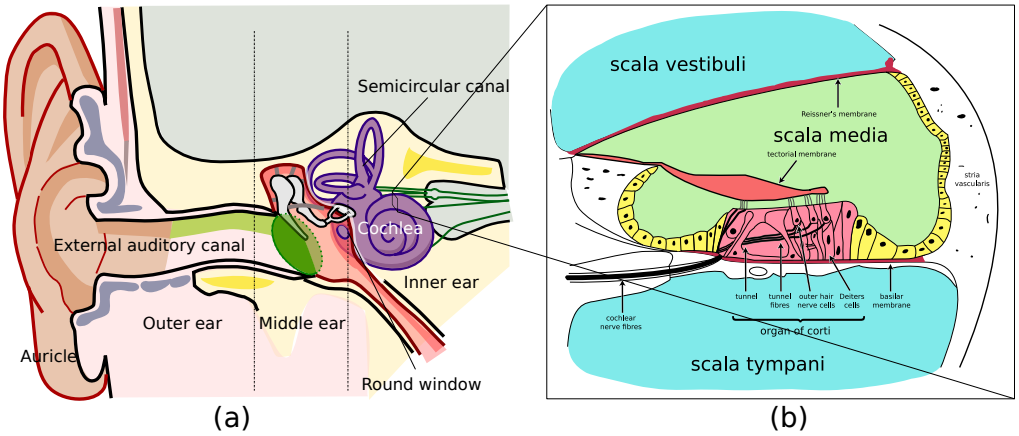

Figure 1.1: (a) Human ear anatomy (adapted from [Chittka and Brockmann, 2005]); (b) Cross section of the organ of Corti inside the Cochlea (adapted from [Chittka and Brockmann, 2005])

1.1.2 Cochlea and cochlear implant

The cochlea is an organ located in the inner ear with a spiral structure. The cochlea has three different spiraling substructures: scala vestibuli, scala media, and scala tympani. The cochlea plays a pivotal role in the hearing system that

converts the mechanical vibrations to neuron electrical impulse. This occurs at the organ of Corti inside the cochlea. The organ of Corti is a part of the cochlea organ which is composed of different neuron cells and biological structures (see (b) of Fig. 1.1). The hair cells can detect the vibrations caused by the flow of liquid and transfer the movement to electric signals.

As shown in sub-figure (a) of Fig. 1.1 any malfunction of the structures from the outer ear to the inner ear path will lead to varying degrees of hearing loss. A group of hearing disease can be relieved by a cochlear implant (CI) surgery. The CI surgery will first open a tunnel to the cochlea round window through interventional surgery on the skull. Then, an electrodes array (see Fig. 1.2 (a)) will be inserted through that tunnel and accurately positioned in the cochlea scala tympani (ST). The positioning of the electrodes array is the key factor that influences the quality of prognosis. Yet, the CI process is a high risk surgery as many important neurons or nerves are distributed along the insertion path. Any damage to those structures may lead to the degeneration of those cells or may result in the CI complete failure. The Fig. 1.2 (b) shows an example that a failed CI that the electrode array is folded. Currently, the insertion process of the electrode array still relies on experienced surgeons. During the surgery, the expert will carefully inserts the electrode array with a very subtle feedback. The absence of visual navigation entails a high requirement on the surgeon experiences and the quality of the preoperative planning.

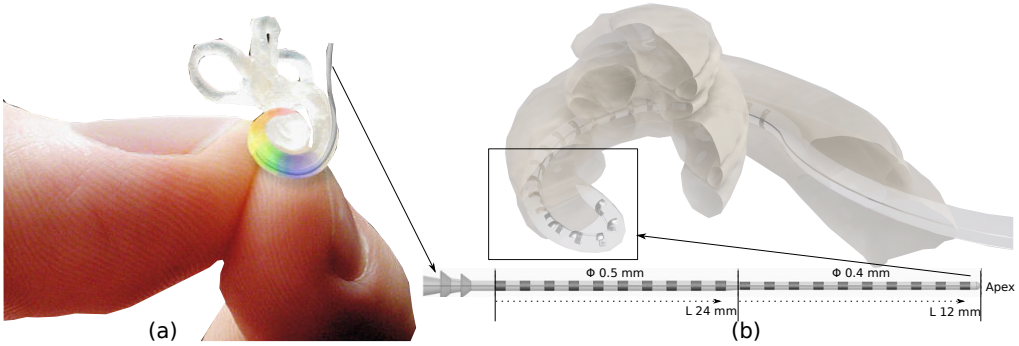

Figure 1.2: (a) Cochlea phantom with an electrode array inserted. (adapted from [AC, 2010]); (b) CI electrode array and fold-over of CI electrode array in 3D view. (adapted from [Bento et al., 2016, Dhanasingh and Jolly, 2019])

1.1.3 Cochlea imaging

Micro-CT (μ-CT) can offer a detailed description of the anatomical structure of the cochlea with high voxel resolution ($5-50\mu m$). This modality is not amenable to clinical usage for human as the μ-CT is mainly used for scanning animals with a limited FOV (field of view usually less than $100mm$) and extreme high radiation dose in comparison with conventional CT scanner. However, the μ-CT can offer a fine imaging of the cochlea details which allows the expert to identify the scala tympani and scala vestibuli (see 1.3(d)). The μ-CT can be used for scanning cadaveric samples to get fine images for expert to segment the cochlea scala structures. The segmentation result of the μ-CT images can be considered as the 'gold standard' as the segmentation uncertainty of the boundary is small in comparison to the low resolution modality (e.g. conventional CT (c) or cone-beam CT (b) shown in Fig. 1.3).

Magnetic resonance imaging (MRI) is an effective modality for identifying the soft-tissues. As the cochlea scala structures are fully filled with lymph this makes MRI very helpful for identifying the structures of the cochlea. Fig. 1.3 (a) shows an MRI imaging of the temporal bone structures.

Conventional computed tomography (CT) systems are widely used for inner ear imaging as one of the most widely used screening system. Cochlea and neural diseases usual need CT for clinical diagnosis. Yet, the human cochlea is a small organ (width: 6.53 $\pm 0.35mm$, height: 3.26 $\pm 0.24mm$ [Zahara et al., 2019]) which is challenging for conventional CT system imaging. Thus, the problem of extracting clinical information for conventional CT imaging becomes difficult but essential and significant for CI.

1.2 Machine learning

Machine learning is the focus of extensive research attention in the computer science field, especially after entering the 21th century with the rapid rise of AI technology. This general term describes a set of methods or algorithms which try to extract knowledge from training dataset for decision making or prediction analysis. Although there is still debates about in the historiography of science, the general consensus is that formulation of Bayes formula marks the beginning of human formally extracting knowledge from data [Wikipedia contributors, 2021].

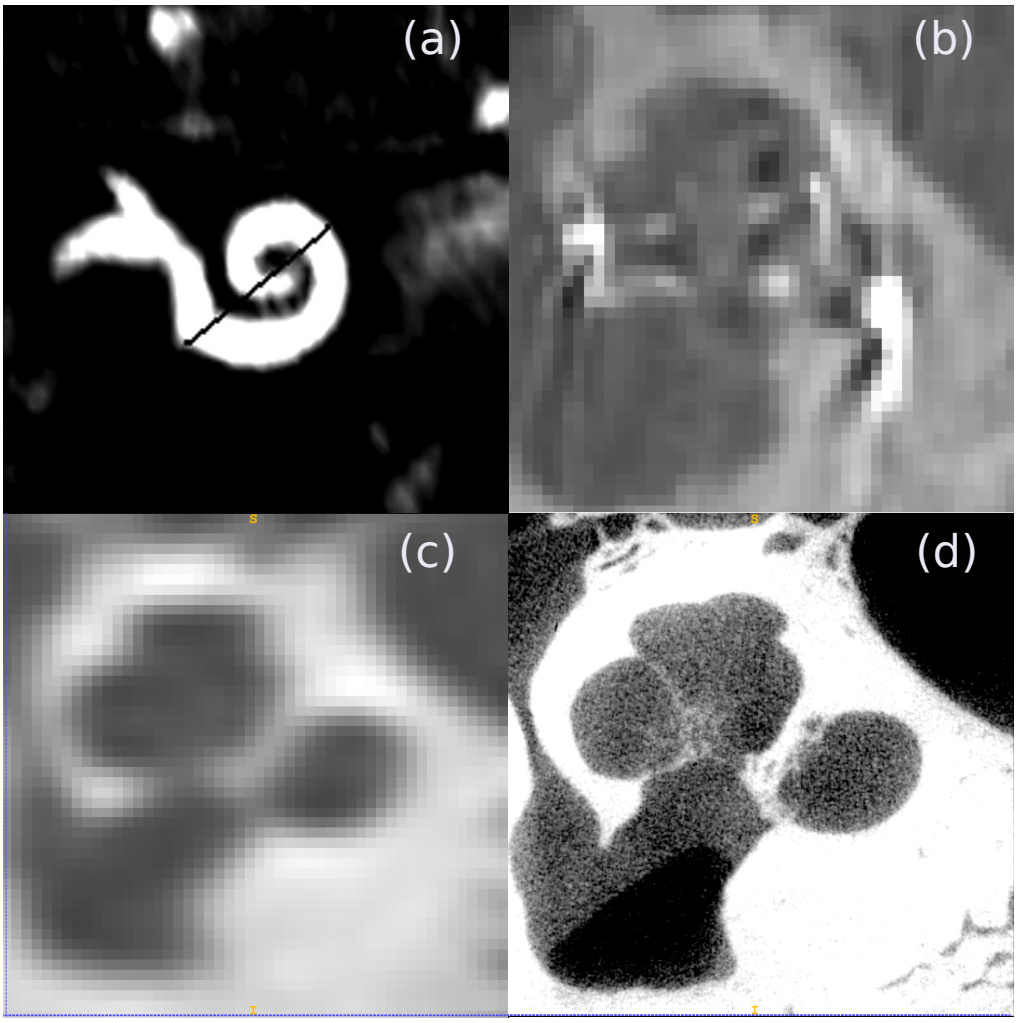

Figure 1.3: (a) Cochlea MR imaging reformatted from 3D-DRIVE MR sequence . (adapted from [Connor et al., 2009]); (b) Cochlea Cone beam CT imaging. (c) Cochlea conventional CT imaging. (d) Cochlea micro-CT imaging.

The development of learning methods based on Bayesian theory eventually formed the ***Bayesian learning***, one of the main branches of machine learning.

1.2.1 Bayesian learning

With a given set of dataset $X = \{x_i | x_i \in \mathbb{R}^d; d, i \in \mathbb{N}\}$, one wants to describe the data with a probability distribution model $p(x|z)$ which driven by the random variables z. We call the probability distribution of the variables $p(z)$ as 'prior distribution' which means an educated knowledge of the model that is suitable for that dataset. Correspondingly, we call the parameters distribution of the variables being fitted on the dataset X as 'posterior distribution', that is $p(z|x)$. The Bayesian inference is based on the Bayesian formula that reflects the relationships between the prior and the posterior distribution:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x|z)p(z)dz} \tag{1.1}$$

The posterior distribution $p(z|x)$ can be used for generating new data points by sampling the distribution of the parameters z. In a very common problem setting, the integration part $\int p(x|z)p(z)dz$ of Eq. 1.1 is intractable due to many reasons such as curse of dimensionality or lack of analytical forms etc. A group of methods introduces some other distributions $q(z)$ as the approximations of the true posterior $p(z|x)$ through maximizing a similarity metric between $q(z)$ and $p(z|x)$. This family of methods is the so called: ***Variational Inference*** (VI) which is the theoretical root of chapters 3 and 4 (parametric shape inference) in this thesis.

1.2.2 Variational inference

The variational inference framework aims to find the suitable replacement model of $p(z|x)$. To this end, a metric needs to be selected for measuring this similarity. The usual metric selected for measuring the similarity between two distributions is the Kullback-Leibler (KL) divergence:

$$D_{KL}(q(z)||p(z|x)) = \int q(z) log(\frac{q(z)}{p(z|x)})dz \tag{1.2}$$

We can then maximize the similarity metrics 1.2 through the optimization of the θ of variational distribution $q(z|\theta)$ which is parameterized by θ to get a variational replacement for the posterior distribution:

$$\underset{\theta}{\operatorname{argmax}} D_{KL}(q(z|\theta)||p(z|x)) \tag{1.3}$$

To cope with the intractable part $p(z)$ we need to convert this problem into a minimization problem since the maximization of Eq. 1.3 is equivalent to minimizing its lower bound with given log evidence $\log p(x)$:

$$\underset{\theta}{\operatorname{argmin}} \int_z q(z|\theta)(\log p(x,z) - \log q(z|\theta))dz \tag{1.4}$$

The problem becomes tractable as we get rid of the posterior term $p(z|x)$. The above equation is the well known ELBO (Evidence lower bound) which can help us to approximate the posterior distribution. However, in case the posterior distribution is very complex, it is difficult to use an explicit distribution to represent the posterior, the approximation quality of the model will decrease.

Kingma and Welling [2014] proposed to use a neural network to make the inference of the distribution of the latent parameters. The ***Variational Auto-Encoder*** (VAE) is the basis of chapter 4. We introduce an adaptation of VAE for cochlea CT dataset generation.

1.2.3 Variational auto-encoder and deep generative model

The core idea of VAE is that the latent variables distribution is modeled through the posterior approximation of a neural network $g_\theta^{neural}(x)$ and using reparametrization trick $z = q_\theta(g_\theta^{neural}(x), \epsilon); \epsilon \sim \mathcal{N}(0,1)$ to make the computational graph differentiable. The approximate posterior distribution $q(z|x)$ and the marginal likelihood $p(x|z)$ are modeled through the inference neural network (encoder) and the generator neural network (decoder), respectively (see Fig. 1.4).

From a high-level perspective, the previously introduced approaches are all falling into the branch of ***Generative Learning***. Correspondingly, another branch of learning technique is the ***Discriminative Learning*** which identifies the features that differentiates data-points into various categories. A very successive practical application of the generative learning and discriminative learning is

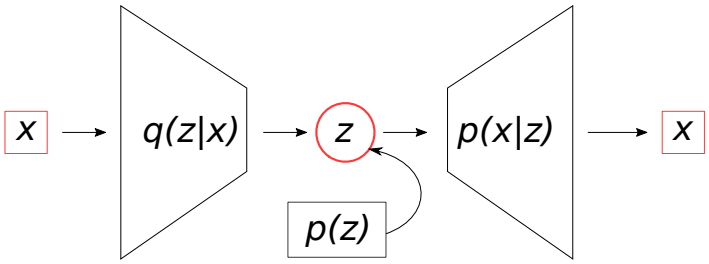

Figure 1.4: The VAE structure.

the ***Generative Adversarial Network*** (GAN) [Goodfellow et al., 2014]. The GAN consists of two deep neural networks: the generator neural network G_ϕ and the discriminator neural network D_θ. The generator network tries to fool the discriminator by generating fake samples and the discriminator tries to identify the input instances that have been collected from the real dataset or from the ones created by the generator (fake data). This objective can be achieved by optimizing the target function of the GAN:

$$\underset{\phi}{\text{argmax}}\ \underset{\theta}{\text{argmin}}\ \mathbb{E}_x[log(D_\theta(x))] + \mathbb{E}_z[log(1 - D(G_\phi(z)))] \tag{1.5}$$

where z is a noise vector sampled from a parametric distribution (often Gaussian). The practical application of GAN is introduced in chapter 2.

1.3 Objective of the thesis

We present many applications of artificial intelligence on cochlea CT images processing and analysis in the remaining chapters. In summary, we study the listed research questions:

- Metal artifacts are commonly found in CT imaging, which can trouble clinicians to perform image examinations. Especially, in inner ear CT imaging, postoperative images are often polluted by the serve metal artifacts introduced by the insertion of an electrode array. Is it possible to learn a representation model between the metal artifacts free images and the metal artifacts spoiled images? How can we employ the learned model to address the metal artifacts reduction task? (Chapter 2)

- Anatomical shape features are important information in medical image analysis, which can be key reference information for many illness diagnoses. In another aspect, based image segmentation are a usual prior step for object shape analysis, which vastly relies on machine learning in recent. Can we employ machine learning no only to segment images but also to perform shape inference for understanding the shape attributions of the objects at the same time? (Chapter 3)
- Generative models are effective tools for data modeling that can be used for learning VAE for learning the dataset which can be modeled with a group of latent variables. The modeling quality of the VAE is constrained by the tightness of ELBO. Can we improve the VAE performance further by tightening the ELBO more? (Chapter 4)
- Landmarks in medical images are often expensive to get as the annotation in 3D volume is very time-consuming and the accuracy is hard to guarantee. Can an algorithm learns how to detect the landmarks automatically with a only one training sample is available? (Chapter 5)

1.4 Organization of the thesis

The structure of the thesis is organized as following:

Chapter 2 presents an adapted Generative Adversarial Network for metallic artifacts reduction and predicting the presence of electrode array of CI in postoperative CT images. The work is adapted from [Wang et al., 2019e].

Chapter 3 highlights the Bayesian inference of a shape model for object segmentation which incorporates a parametric shape information into an expectation-maximization algorithm. This chapter is adapted from [Wang and *et al.*, 2020, Wang et al., 2021a].

Chapter 4 shows the use of a gradients informed variational autoencoder for medical volume dataset modeling. The framework allows us to generate samples from a simple distribution. This chapter is adapted from [Wang and Delingette, 2021b].

Chapter 5 introduces an one-shot learning based landmarks detection approach for 3D volume landmarks detection. The proposed approach requires only one

volume for training and is able to detect hundreds of volumes. This chapter is adapted from [Wang et al., 2020e].

Chapter 6 summarizes the content of the thesis with contributions and perspective. This chapter is partially adapted from [Wang and Delingette, 2021a].

Chapter 2

Metallic Artifact Reduction based on Generative Learning

Metal Artifacts pose a common difficulty for post-operative quality assessment in computed tomography (CT). A vast body of methods have been proposed to tackle this issue for CT imaging. Yet, these methods were designed for regular CT scans and their performance is usually insufficient for fine imaging of tiny implants. For the clinical requirements of high-resolution detailed CT imaging, we

propose a 3D volume metal artifact reduction algorithm based on a 3D generative adversarial neural network. Depending on the method of data collection, our approach can be either supervised or unsupervised, and applied to 3D CT volume artifact reduction. We show quantitatively and qualitatively that the proposed method outperforms other general metal artifact reduction approaches. This chapter is based on a augmented work [Wang et al., 2021b] of our conference paper presented at MICCAI 2019 [Wang et al., 2019d].

2.1 Introduction

Computed Tomography (CT) is one of the most widely used imaging techniques in clinical practice. The physical principles of spiral CT lead to the unavoidable creation of artifacts in the reconstructed images in the presence of dense materials, *i.e.,* those composed of atoms with high atomic numbers. Several physical phenomena contribute to the creation of such artifacts, including X-ray beam hardening, X-ray scatter, electronic noise, edge effects and also the geometrical characteristics of metal parts. The artifacts are commonly found in routine clinical postoperative imaging, for instance due to fixation plates in orthopaedics, cochlear electrode implants in otology, contrast agents, *etc.* These spurious signals in CT images may impair postoperative analysis. For instance, during cochlear implant surgery, an electrode array inserted along the cochlear scala tympani is usually comprised of a metal alloy, for its high electrical conductivity. The existence of metal artifacts in postoperative CT makes the evaluation of the position of the electrodes along the scala difficult. The knowledge of the relative position of the cochlear implant is one of the main determinants for assessing the success of the surgery and leads to appropriate and more personalized patient care.

2.1.1 Augmented MARGAN approach

Metal artifact reduction (MAR) methods aim to decrease the extent of such artifacts (see: 2.1). Classical non-learning-based MAR algorithms are divided into two groups: corrupted projection recovery and iterative image reconstruction-based methods [Mehranian et al., 2013]. In the former case, projections corrupted by the presence of metal absorbing the X-rays are detected and then replaced by predicted or interpolated values, based on prior knowledge [Kalender et al.,

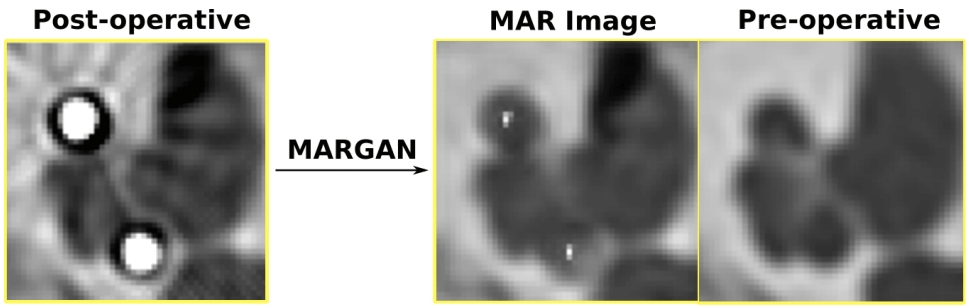

Figure 2.1: Sketch of the MARGAN applied on postoperative images

1987a]. The efficiency of the approach is related to the ability to recover the projected signals in the absence of metal parts [Mehranian et al., 2013]. In the case of iterative methods, the missing data in image or projection space is estimated on the basis of statistical principles, possibly including prior knowledge. Aside from Filtered Back Projection (FBP) based methods, Naranjo et al. [2011] introduced mathematical morphology algorithms for MAR by converting the image to polar coordinates centered on the metal artifact.

Recently, the field of MAR has been revived by the development of deep learning methods that provide supervised mechanisms for extracting relevant image features. A number of 2D Convolutional Neural Network (CNN)-based MAR methods have been proposed that are summarized in Table 2.1.

Zhang and Yu [2018] introduced CNNs as prior information in the sinogram (projection) space for the inpainting or sinogram completion task using a simulated dataset in the training stage. However, this method needs either the original CT sinograms (usually unavailable to the typical user) or to project back the input image in order to fill-in the missing traces. This limits its application in our dataset, and the sinogram-based MAR algorithms tend to generate over-smoothed images due to their filtering effect.

Huang et al. [2018] developed a deep learning network, *RL-ARCNN*, in image space to predict residual images (the difference between the images with and without artifacts) to remove metal artifacts in cervical CT images.

The 2D network *DestreakNet* was proposed in [Gjesteby et al., 2017] for streak artifact reduction as a post-processing step in order to recover the details lost

after the application of the interpolation-based normalized MAR [Meyer, 2010] algorithm.

Lyu et al. [2020] proposed Dudonet++ for 2D CT metal artifact reduction. Their approach relies on processing the image with artifacts (henceforth referred to as artifact image) in both sinogram and image spaces in order to restore fine details in the image. Their quantitative evaluation shows that the Dudonet++ is effective for artifact reduction on simulated CT images but it lacks a quantitative evaluation on a clinical dataset. Furthermore, the method uses a beam hardening correction [Verburg and Seco, 2012], which is not always optimal, for instance in the case of Cochlear Implant MAR (see 2.4).

Recently, generative adversarial networks (GAN) [Chen et al., 2021, Wang et al., 2020a] arouse widespread interest in many research communities. The GAN was also devised for solving MAR problems instead of CNN classification or regression networks, owing to their ability to generate high quality images. Wang et al. [2019a] proposed a conditional GAN (cGAN) approach for CT images with cochlear implants (CI), using a collection of paired and registered post- and preoperative cochlear implant volumes to train 2D/3D cGANs for inner ear MAR. A difficulty in this approach is to collect and, most importantly, to register (preoperative) artifact-free and (postoperative) artifact images. This registration problem must be able to cope with the presence of outliers due to the presence of artifacts.

[Nakao et al., 2020] also proposed a MAR method based on CycleGANs for artifact reduction in dental filling and neck CT images. The approach is unsupervised and aims to achieve a cross-domain (artifact and artifact-free dataset) style transformation through feature swapping. This approach does not require training on paired datasets, *i.e.,* with and without artifacts, but CycleGAN performance significantly worsens when unpaired data is used [Zhu et al., 2017] for training instead of paired data. This approach was qualitatively compared with the manual corrections available in commercial CT scans and quantitatively assessed on synthetic datasets. While the output of the CycleGANs seems effective, this method may not be useful for the reduction of tiny artifacts like cochlear implants, due to the difficult separability of artifacts in feature space.

In this chapter, we propose a GAN-based MAR method that relies on simulated

training data and is suitable for pre- and postoperative images. To the best of our knowledge, our approach is the first MAR algorithm that combines the physical simulation of metal artifacts with 3D GAN networks. While classical GAN-based methods such as [Wang et al., 2019a] rely on the existence of paired images with and without artifacts for training, our approach has several advantages. First, only preoperative images (without artifacts) are required for the training stage, because the generation of the corresponding artifact image is based on physical simulation. This allows a large set of training images (800 images) to be used, without the need for registering the pre- and postoperative images. Second, the nature of artifacts can be easily modulated by controlling the complexity of the artifact simulation model complexity. Third, we introduce the concept of *augmented metal artifact reduction* by optionally adding landmarks in the corrected image that indicate the central location of metal parts. More precisely, in this chapter, we show that for the postoperative cochlear implant CT images, the location of each electrode center can be identified in the corrected image such that ENT (ear, nose and throat) surgeons can assess the quality of the implantation surgery. Compared to CycleGANs [Nakao et al., 2020], the MARGAN approach allows artifacts to be easily disentangled from the background. This is why we believe this approach is probably more appropriate to attenuate artifacts created by tiny implants. Fourth, MARGAN was evaluated on postoperative, cone beam CT images. Finally, MARGAN was developed as a 3D GAN since metal artifacts usually vary continuously between slices. A summary of current studies of MAR is shown in Tab. 2.1.

Table 2.1: Summary of major MAR approaches. (In the data collection column: BH, SC and EN indicate Beam Hardening, Scattering and Electronic Noise, respectively)

	Processing Domain	Inner Ear MAR	2D/3D	Dataset Collection	Quantitative evaluation on clinical data
marBHC	Sinogram	NO	2D	Non-Learning	YES
marLI	Sinogram	NO	2D	Non-Learning	YES
NMAR	Sinogram	NO	2D	Non-Learning	YES
CNN Prior Zhang and Yu [2018]	Sinogram	NO	2D	Simulation (BH)	YES
RL-ARCNN Huang et al. [2018]	Image	NO	2D	Simulation (BH)	YES
DestreakNet Gjesteby et al. [2017]	Image	NO	2D	Simulation (BH)	YES
DudoNet++ Lyu et al. [2020]	Sinogram+Image	NO	2D	Simulation (BH;SC;EN)	NO
CycleGAN Nakao et al. [2020]	Image	NO	3D	Unsupervised	NO
cGAN Wang et al. [2019a]	Image	YES	2D	Paired Data	YES
Augmented MARGAN (Proposed)	Image	3D	YES	Simulation (BH;SC;EN)	YES

The MARGAN method is based on two main stages (see Fig.2.2 - 2.2). In the first stage (Fig. 2.5), given a preoperative image from the training set, one or several images with metal artifacts are generated. This requires a rough segmentation of the structures of interest, the position of metal parts (e.g., electrode arrays) and the simulation of artifacts based on a CT image formation model. Furthermore, the location of the electrode arrays is added to the generated images. In the second stage (Fig. 2.2), a 3D GAN is trained using preoperative and corresponding simulated artifact images. The GAN loss is improved by adding a term based on Retinex theory to decrease the image blur in generated images. After training, the GAN is applied on a postoperative image without any segmentation or other preprocessing. It results in images with attenuated metal artifacts but also with landmarks corresponding to electrode centers.

The MARGAN method was applied to a set of inner ear CT images to reduce the artifacts created by cochlear implants. Qualitative and quantitative results are provided for 33 paired pre- and postoperative CT images, including a comparison with two classical open source MAR algorithms. Qualitative evaluation of cone beam CT (CBCT) postoperative images is also provided.

This chapter extends the initial work published in [Wang et al., 2019d] in several ways. The artifact simulation model is more sophisticated, including scattering effects and electronic noise of the CT system detectors. The algorithm evaluation is more comprehensive, with the addition of paired CT images, CBCT images and a study of the impact of the Retinex loss. The postoperative electrode position is assessed in a few cases with postmortem photographic views of the cochlea.

The chapter is organized as follows: In section 2.2, we introduce the CI and CI metal artifact simulation procedures (the gray box in Fig. 2.2). In section 2.3, the network implementation is described (the green box in Fig. 2.2). Results of the MARGAN algorithm are presented in section 2.4. Sections 2.5 and 2.6 discuss the contributions and limitations of the proposed approach.

2.2 Simulation of metal artifacts in CT images

The difference between traditional algorithms and learning-based algorithms is the learning method need data to fit. However, the data is quite difficult to get for medical images. Moreover, in our problem, the CT images with metal

artifacts and correspond metal-free images are impossible to get in clinical. In order to collect training pairs, we propose to use a simulation based artifacts generation approach for generating metal artifacts images.

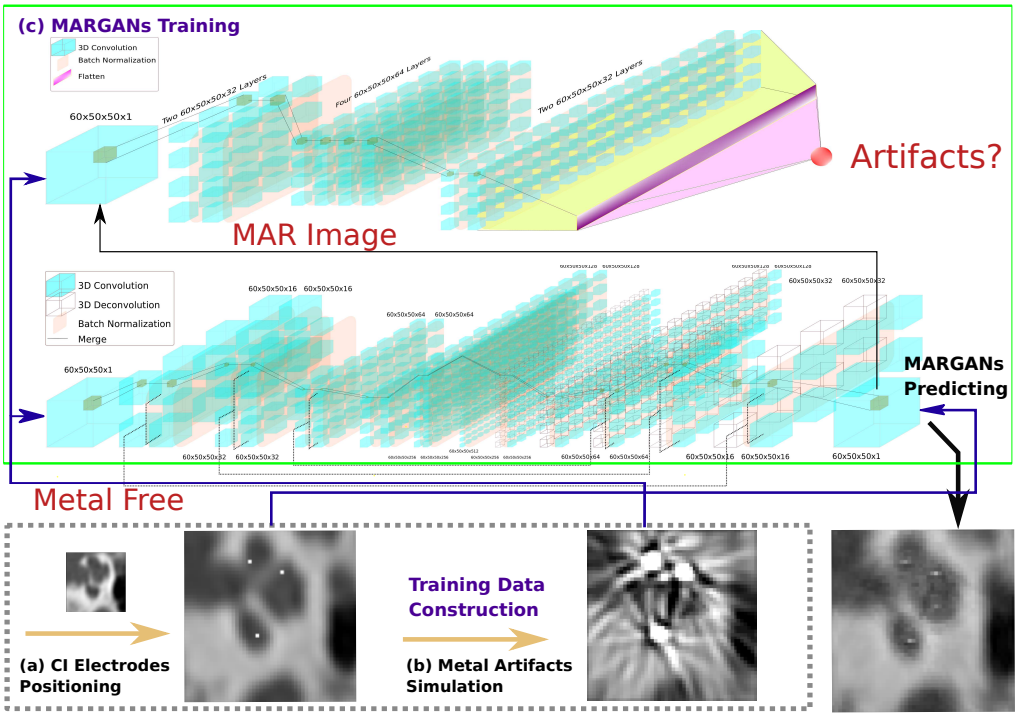

Figure 2.2: The framework of MARGAN for metal artifact reduction. (a) The cochlear implant positioning simulation; (b) CI metal artifact physical simulation. (c) A 3D GAN is trained with simulated and preoperative datasets: The discriminator network aims to identify whether or not the input image is one polluted by artifacts. The generator network accepts an input artifact image and generates a MAR image.

The simulation of metal artifacts in CT images from artifact-free images entails i) simulating the presence of metal parts in the images as shown in Fig: 2.3 and ii) simulating the creation of artifacts caused by those metal parts as shown in Fig: 2.4. The former algorithm is completely dependent on the organ or implant considered, while the latter is far more generic, based on the physics of image formation.

2.2.1 Simulating the presence of metal parts

The processing pipeline to generate the training set for the MARGANs is displayed in Fig. 2.3. In this section, we consider the case of preoperative CT images of the inner ear prior to cochlear implant surgery. The objective is therefore to simulate, in these preoperative images, the addition of metal electrode arrays associated with the implant.

The 3D CT volumes of the inner ear, written as $I(\mathbf{x})$, are first rigidly registered on a template image by a block matching algorithm [Ourselin et al., 2000a]. The template is a sample CT image that has been manually cropped around the temporal bone. The registration is necessary to cope with the variations of field of view and pose in the input image dataset.

A region of interest (ROI) is then cropped to get a cochlear volume suitable for further processing. We then fit a parametric cochlear shape model [Demarcy et al., 2017] to automatically reconstruct the shape of the cochlea (step (II) of Fig: 2.3). The accuracy required for the registration and segmentation steps is limited.

The signed distance map [Wang and *et al.*, 2020] from the fitted triangular mesh of the parametric shape model is generated as shown in step (III). It is then thresholded (step (IV)) to create a 3D tubular binary mask near the center-line of the scala tympani of the cochlea. This mask corresponds to the probable location of the electrodes after a CI intervention. Finally, in step (V), the voxel values in Hounsfield units (HU) of the mask region are then set to $3071HU$ which is the maximum detectable HU of the CI metal artifacts. This creates the image $I^{\text{train}}(\mathbf{x})$ used for training the GAN network.

2.2.2 Simulation of beam hardening, scattering and electronic noise due to metal parts

The metal parts have large absorption ratios of X-ray energy which is the cause of of the visible artifacts in CT images. It impacts the whole image formation process through several physical effects. Our previous work [Wang et al., 2019d] only considered the simulation of the beam hardening effect, inspired by the work of [Zhang and Yu, 2018]. In this chapter, we improve the realism of the simulated artifacts by also including the X-ray scatter effect through Monte Carlo

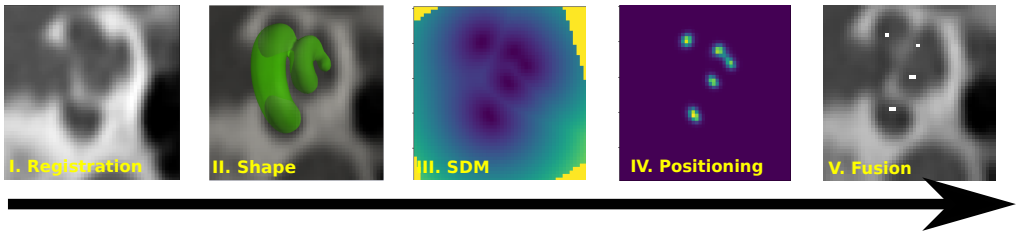

Figure 2.3: Cochlear implant electrode positioning simulation; (I) Registration of CT image on a template image; (II) Cochlear shape fitting; (III) Signed distance map generation; (IV) Electrode positioning; (V) Image fusion with electrodes.

simulation and the detector electronic noise. The three main physical effects governing the generation of metal artifacts are described below, along with the processing pipeline.

Beam hardening effect For a monoenergetic X-ray source entering a material of thickness δz along direction z at position x, y, the number of photons $L(x, y, \delta z)$ is given by the Beer-Lambert law : $L(x, y, \delta z) = L_0 e^{-\mu(x,y)\delta z}$ where L_0 is the initial photon number and $\mu(x, y)$ is the linear attenuation coefficient of the material. The attenuation coefficient depends on the energy of the input photon $\mu(E_v)$, and therefore for a polychromatic X-ray beam having the energy distribution (or spectrum), $\phi(E_v)$, the number of photons received by the entire detector surface is then:

$$L = \int_{E_0}^{E_n} (\phi(E_v) e^{-\iiint \mu(x,y,z,E_v) dx dy dz} + S(E_v)) dE_v \tag{2.1}$$

where E_0 and E_n are the minimum and maximum energies for a fixed tube peak voltage, and $S(E_v)$ is an additive offset that captures X-ray scattering.

Scattering effect The Compton effect applies to incoming X-ray photons that interact with the free electrons in the traversed materials. This effect results in random changes (scatter) in the directions of the photons, which may still reach the detector plate despite collimator devices. The Compton scatter is enhanced in the presence of metal parts, thus resulting in an offset in the number of photons $S(E_v)$ and leading to a reduction in the image contrast. Computing

this additional scatter is very complex as it depends on the projected plane and the material and geometry of the tissue surrounding the metal parts. To this end, we use Monte Carlo simulation to estimate the offset value $S(E_v)$ for different detector positions and orientations. The governing equation for the simulation provides the emission energy $E_p(\beta)$ of a polychromatic ray deviating by an angle β from its initial trajectory :

$$E_p(\beta) = \int \frac{E_v}{(1 + E_v/m_e c^2)(1 - cos(\beta))} dE_v \tag{2.2}$$

where m_e is the electron mass and c the speed of light. To estimate the scatter effect inside the cochlea on X-ray detectors, we use the Zubal [Zubal et al., 1994] head phantom where metal parts are roughly positioned inside the temporal bone. Based on the MCGPU software [Badal and Badano, 2009] performing GPU Monte Carlo simulations of photon transport in voxelized geometry, we simulate thousands of X-ray photon trajectories at different energies, positions and orientations through the head and produce both the scatter-free sinogram $F(E_v)$ and the scatter sinogram offset $\tilde{S}(E_v)$. The scatter sinogram offset is corrected by a scale factor such that the resulting *scatter to primary ratio* $\alpha = \frac{\text{mean}(\tilde{S}(E_v))}{\text{mean}(F(E_v))}$, falls within the range of 0.1% to 2%, which was experimentally found by Glover *et al.*[GH, 1982]. This is simply done by randomly picking a ratio α_r within 0.1% to 2% and computing

$$S(E_v) = \frac{\text{mean}(F(E_v))}{\text{mean}(\tilde{S}(E_v))} \alpha_r \ \tilde{S}(E_v) \tag{2.3}$$

The same ratio α_r is used for simulating all sinograms of the same image to obtain spatially consistent artifacts.

The computation of the scatter offset is dependent on the X-ray energy, position, and orientation but is independent of the input image as it relies on the digital head phantom augmented with metal parts next to the temporal bone. Only the scatter to primary ratio varies between different volumes. This implies that the scatter sinograms can be precomputed, thus alleviating the computational load when generating images with metal artifacts.

Detector Noise Once photons hit the x-ray detector, the scintillator transforms the deposited energy into visible light, while a photomultiplier translates

this light into an electric signal. In this process, some electronic noise is introduced which can be modeled by a zero mean Gaussian distribution [Benson and Man, 2010] with standard deviation σ_e: $N(0, \sigma^2)$. The signal measured in each sinogram L_{final} can then be written as: $L_{\text{final}} = L + \mathcal{N}(0, \sigma_e^2)$ where L is the energy deposited as described Eq. 2.1 and $\sigma_e^2 = 0.04$.

Simulation pipeline The overall metal artifact simulation pipeline is described in Fig. 2.5. In the first step, we use an X-ray energy spectrum $\phi(E_v)$ extracted from a CT manufacturer dedicated site [1] for a tungsten anode tube at 140 kVp. The spectrum is sampled at five sample energies from which attenuation maps $\mu(x, y, z, E_{v_i})$ are generated. This computation is based on the Hounsfield unit formula and the water absorption coefficients as a function of energy. We then perform fan-beam projection (Step III) of the five attenuation maps to produce sinogram-like images representing absorbed energy on the CT detectors. The scattering and attenuation sinograms are precomputed on a head phantom for various orientations and positions of the source. The projection of the ROI of the head where metal parts have been inserted creates a sine trace on the scattering and attenuation sinograms. This trace is randomly sampled, then normalized as in Eq. 2.3 to obtain a plausible scatter to primary ratio. It is then added to the electronic noise and to the weighted sum of the five sinograms (Step IV) and a discretization of Eq. 2.1. Finally, inverse fan-beam projection produces the output image with metallic artifacts (Step V).

The difference between simulated images with and without scattering noise is shown in Fig. 2.6 with a subtraction map. We see that scattering and electronic noise can introduce significant new artifacts.

2.3 GAN-based Metal Artifact Reduction

Given pairs of preoperative and simulated postoperative images, we aim to train a network that generates the former given the latter as a way to reduce metal artifacts. The use of a GAN to tackle the MAR issue is motivated by the successful use of 2D and 3D GANs such as SRGAN [Ledig et al., 2017, Sanchez and Vilaplana, 2018] to solve imaging Super-Resolution (SR) problems.

[1] https://www.oem-xray-components.siemens.com/x-ray-spectra-simulation

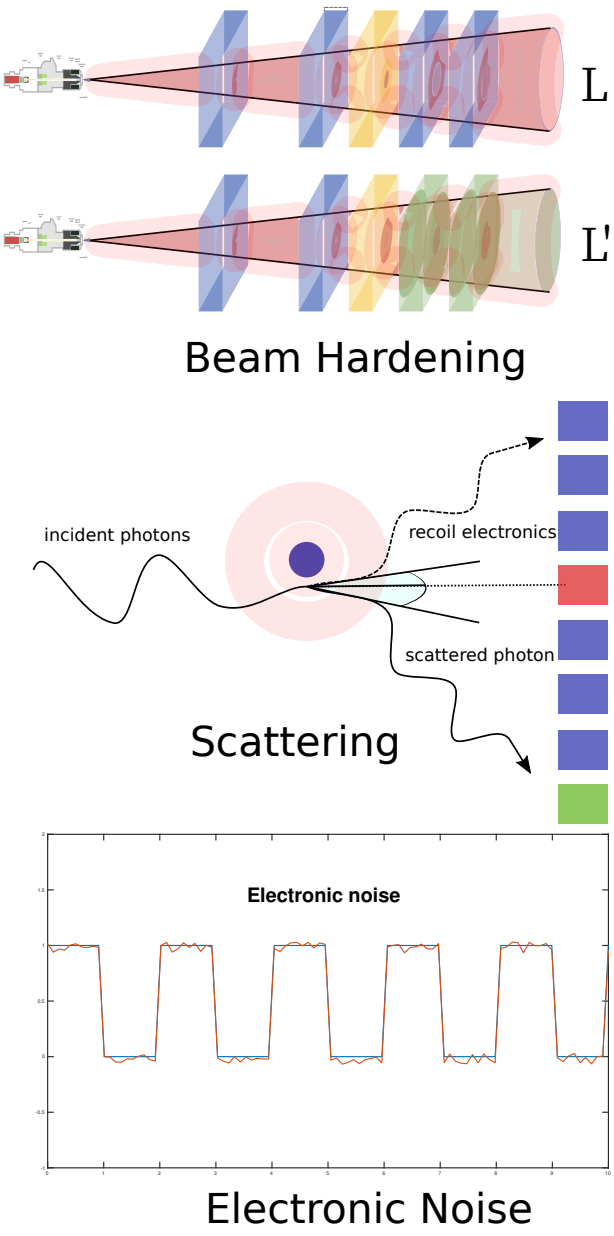

Figure 2.4: Three main physical effects are considered for simulating metal artifacts. (Top) Beam hardening. The metal part, shown in gold has nonlinear X-ray energy absorption, thus violating the Beer-Lambert law. This generates an underestimation of the material attenuation ratio located after the metal part. (Middle) Scattering effect. A scattered photon is abnormally detected by the green detector, but would have been detected by the red detector in the absence of scatter. (Bottom) The electronic noise (red) and the corresponding ideal signal (blue).

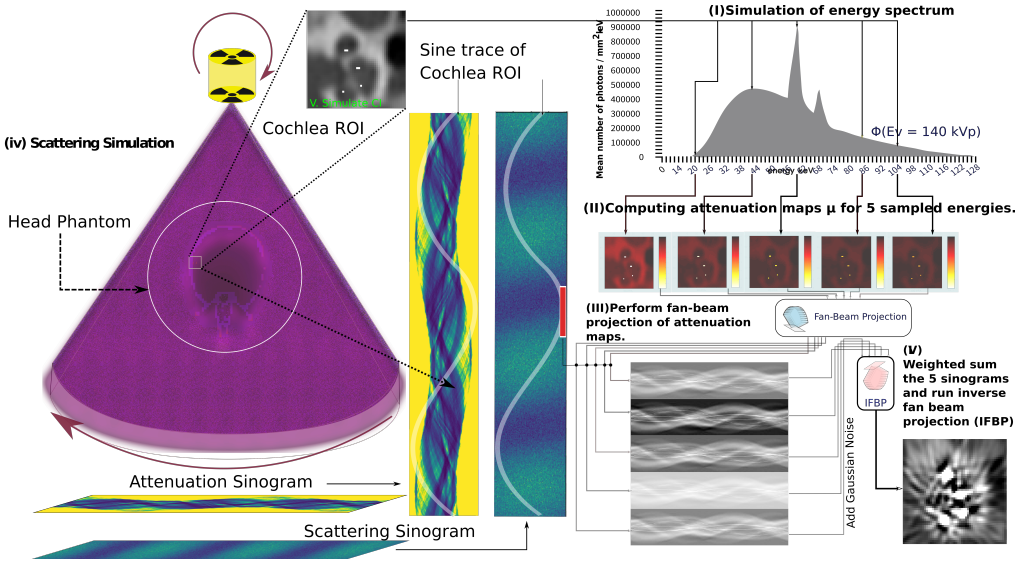

Figure 2.5: Pipeline of metal artifact simulation. Given a preoperative image with simulated implants, the simulation starts from the computation of attenuation maps (steps (I) -(II)) for the cochlea ROI volume based on the energy spectrum of the X-ray tube. Step (III) performs fan-beam projection to simulate the sinograms of the attenuation map. (IV) Monte Carlo simulation of scattering effects is performed offline on a head phantom for the generation of the scattering sinograms whose traces in the ROI are randomly chosen, then normalized and added to the combined attenuation map sinograms. (V) Gaussian electronic noise is added and then inverse fan-beam projection is performed to get the final simulated artifact images.

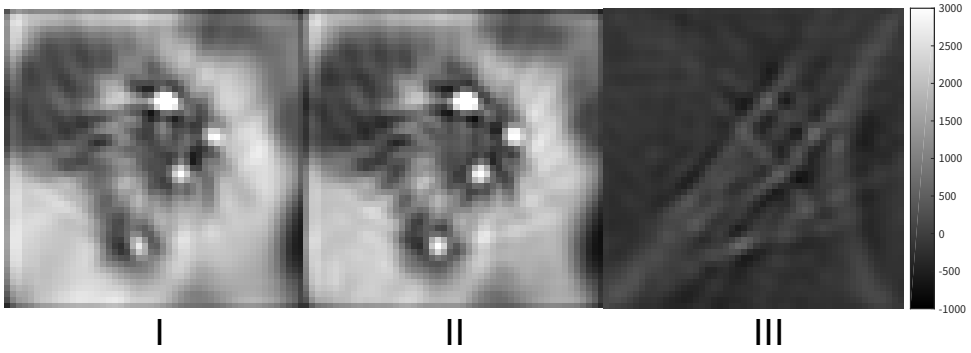

Figure 2.6: Noise image and beam hardening image. (I) Simulation with scattering effect and electrical noise. (II) Simulation with only beam hardening. (III) The subtraction map between the two simulations.

2.3.1 Network Overview

In MARGAN, two neural networks are used: the generator network produces the MAR images and the discriminator network indicates whether the input image contains metal artifacts or not. The generator network, G_{w_g}, with weights, w_g, aims at modeling the mapping between the image with artifacts, I^{m}, and the simulated artifact-free image, I^{train}. We denote by I^{MAR} the 3D image created by the generator network, which should be as close as possible to I^{train}. The discriminator neural network, D_{w_d}, tries to detect the presence of artifacts in the generated MAR images, I^{MAR}. To train G_{w_g} and D_{w_d} networks, the sum of discriminator and generator losses is optimized as detailed below.

2.3.2 Network Architecture

The generator network architecture is similar to U-Net with convolution and deconvolution layers, skip connections and batch normalization layers to improve the training efficiency (see Fig 2.2). Moreover, unlike [Sanchez and Vilaplana, 2018] which is patch based, the input to the network consists of full 3D images as it is compatible with GPU memory. The number of filters increases gradually from 1 to 512, the maximum number of feature maps that will fit on an 11 Gb video-memory GPU card. The discriminator network follows that of [Sanchez and Vilaplana, 2018] with eight groups of convolution layers and batch normalization layers combined sequentially.

2.3.3 Loss Functions

Discriminator Loss The discriminator network, D_{w_d}, is trained using output images from the generator network, $I^{MAR} = G_{w_g}(I^{\mathrm{m}})$, and images without any metal artifacts, I^{nm}. Following [Sanchez and Vilaplana, 2018], the discriminator loss enforces the ability of the discriminator network to distinguish the artifact-free images, I^{nm}, from the generated ones, I^{MAR} :

$$\arg\max_{w_d} \; L_D = \mathbb{E}_{x\sim I^{\mathrm{nm}}} \log\left(D_{w_d}(x)\right) + \mathbb{E}_{y\sim I^{\mathrm{m}}} \log\left(1 - D_{w_d}(G_{w_g}(y))\right)$$

Generator Loss The objective of the generator network is to produce an image, $I^{MAR} = G_{w_g}(I^{\mathrm{m}})$, as close as possible to the target image, I^{train}. This is why the first loss term is the mean square error (MSE), $L_{mse} = \mathbb{E}_{y\sim I^{\mathrm{m}}}(|I^{\mathrm{train}} - G_{w_g}(y)|^2)$, to encourage a similarity between generated and target voxels. But using only the MSE loss leads to blurred MAR images with a lack image detail at high frequencies. To avoid this excessive smoothing, we propose a new loss term based on Retinex theory [Land and McCann, 1971]. This theory is mostly used to improve images seriously affected by environmental illumination. The Retinex theory assumes that a given image can be considered as the product of environmental brightness (or illumination), $L(x, y)$, and the object reflectance, $R(x, y)$. This reflectance map contains high frequency details and is unaffected by the illumination condition, a property referred to as the color constancy phenomenon. The objective of Retinex-based algorithms is to recover the reflectance image from the original one. In single-scale Retinex approaches [Zhang et al., 2011], the environmental brightness is simply a Gaussian blur version of the input image and therefore $\log(R(x, y)) = \log(I(x, y)) - \log(I(x, y) * \mathcal{N}(0, \sigma))$ where $\mathcal{N}(0, \sigma)$ is a Gaussian function of standard deviation σ, and $*$ is the convolution operator. This leads us to introduce the following Retinex loss to make its illumination part as close to 1 as possible :

$$L_{retinex} = \mathbb{E}_{Y\sim I^{\mathrm{m}}} \frac{|G_{w_g}(Y) - e^{\log G_{w_g}(Y) - \log G_{w_g}(Y) * \mathcal{N}(0,\sigma)}|}{|Y|} \tag{2.4}$$

where the expectation is taken over the image domain. This loss definition ensures numerically stable evaluations and enforces salient features in the image that would otherwise be attenuated. Combining it with the adversarial term $L_{adv} = \frac{1}{2}|D_{w_d}(G_{w_g}) - 1|^2$ as in [Sanchez and Vilaplana, 2018], the full optimization target of the generator is:

$$\arg\min_{w_g} \; L_{generator} = \alpha \cdot L_{retinex} + L_{mse} + L_{adv} \tag{2.5}$$

where α is a parameter controlling the influence of the Retinex loss.

Table 2.2: Material Mapping Table for Voxel Conversion to MCGPU File

	air	water	bones	muscle	titanium	soft tissue	fat
MC-GPU MATERIAL	1	15	4	2	16	3	6
DENSITY [g/cm^3]	0.001205	1.000	1.990	1.041	4.506	1.038	0.916

2.4 Results

2.4.1 Dataset

Training data The cochlea dataset was collected from the Radiology Department of the Nice University Hospital with a GE LightSpeed CT scanner without any metal artifact reduction filters. The preoperative dataset includes 1000 temporal bone images (493 left and 597 right) from 597 patients. The original CT volumes are registered to a sample image by a pyramidal block-matching algorithm in order to spatially normalize all images, then they are resampled with $0.2 \times 0.2 \times 0.2mm^3$ voxel size. They were then cropped to volumes of $60 \times 50 \times 50$ voxels around the cochlea region. We then simulated on all volumes, the insertion of CI electrodes and the generation of metal artifacts as described in section 2.2. This created a set of 1000 pairs of images, with and without metal artifacts.

Evaluation Data The evaluation dataset #1 includes 33 cadaver temporal bone CT images collected from the same site from different bodies. The imaging protocol was the same as for the training dataset but was performed before and after the implantation of CI, thus leading to 33 pre- and postoperative image pairs. The temporal bones were ground by an ENT (ear, nose and throat) surgeon, approximately along a plane perpendicular to the cochlear modiolar axis at the bottom of the scala tympani as shown in Fig. 2.12. Pictures of the ground bones were acquired in order to visualize the electrode array.

Finally, the second evaluation dataset includes 8 postoperative images that were acquired on a Carestream 9600 cone beam CT (CBCT) following the CI surgery. These images were resampled, registered and cropped following the same processing pipeline as the training set.

Table 2.3: Dataset Summary: Preoperative and postoperative refer to images collected before and after Cochlear Implant, respectively.

Dataset	Pre-Operative	Post-Operative	Photography
Training	800	0	0
Validation	200	0	0
Evaluation CT	33	33	33
Evaluation CBCT	0	8	0

2.4.2 Implementation details

Artifact simulation A polychromatic X-ray source was simulated with MC-GPU v1.3, a GPU-based Monte Carlo simulator of photon transport in voxelized geometry [Badal and Badano, 2009]. To simulate the scatter effects, we simplified the contents of the human head by assuming it consists of air, water, soft tissue, bone, muscle and unalloyed titanium. Cochlear CT voxel values were converted to MC-GPU v1.3 units based on the material mapping in Table 2.2. The simulation of scatter was performed offline on a GPU parallel computing cluster. The beam hardening maps and the final simulation volumes were computed with Matlab 2017a on a Dell Mobile Workstation with Intel(R) Core(TM) i7-7820HQ @ 2.90GHz CPU.

Neural Networks The networks were trained with a RMSprop optimizer [Arjovsky et al., 2017] with learning rate $l_rg = 1e{-}4$ for the generator and $l_rd = 1e{-}3$ for the discriminator. The MARGAN was implemented with Tensorflow and the weight of Retinex loss was set to $\alpha = 5e{-}5$.

2.4.3 Clinical Evaluation

Qualitative Study Fig. 2.9 shows the output of the MARGAN network for four patients on two selected slices together with pre- and postoperative CT images. The streak artifact patterns were largely suppressed by the MARGAN algorithm. As shown inside the yellow boxes, the artifact patterns were significantly reduced compared to postoperative images. The cochlear structures that were slightly

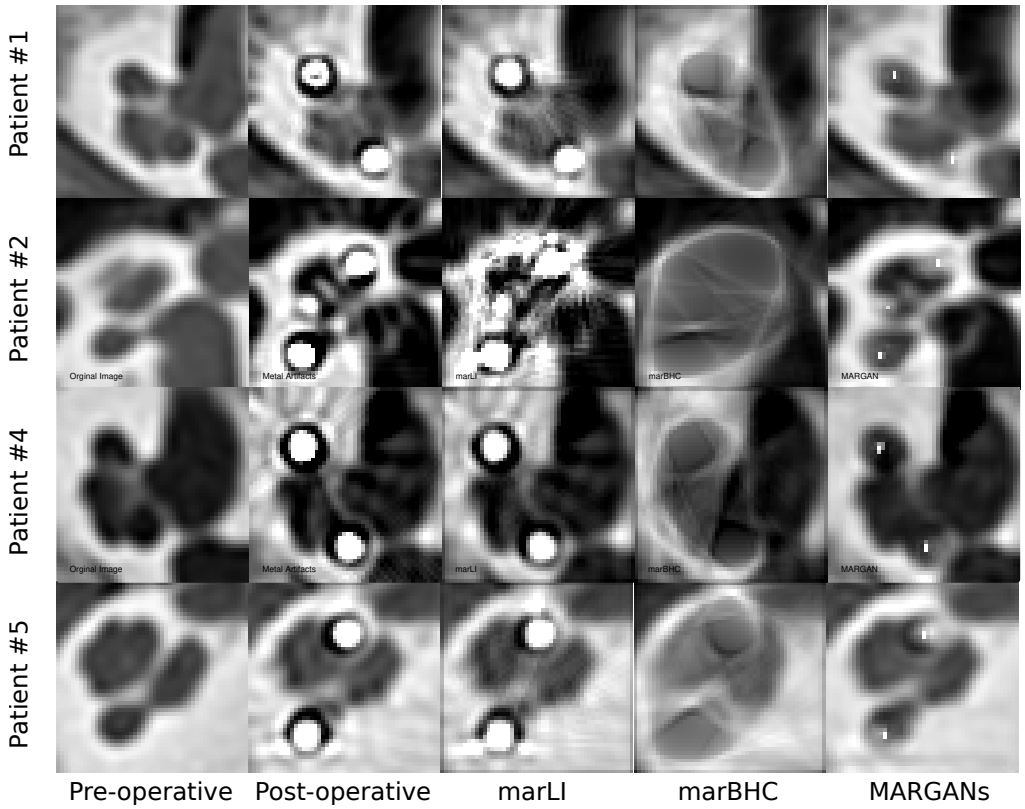

Figure 2.7: Metal artifact reduction visualization of MARGAN in comparison with other approaches for patients #1 - 5.

Table 2.4: Quantitative evaluation of the MARGAN approach compared to marBHC, marLI and Nmar. Mean value shows the advance of MARGAN, STD metric shows the slice consistency of MARGAN.

Metric	Preoperative	marBHC	marLI	Nmar	MARGAN
PSNR	16.33	11.59	16.53	13.58	**18.31**
RMSE	0.15	0.28	0.15	0.52	**0.12**
SSIM	0.58	0.56	0.55	0.52	**0.64**

distorted by the artifacts (indicated by yellow arrows) were mostly recovered in comparison to the preoperative image slices. Finally, the MARGAN-generated images include by design, high intensity pixels at the potential locations of electrode centers. The yellow circles are clearly positioned in the centers of the electrodes and can help otologists visualize the relative positions of electrodes with respect to the scala tympani.

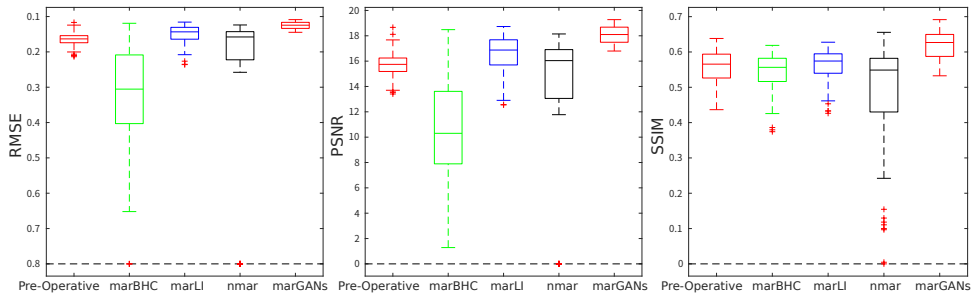

Figure 2.8: The 3D consistency between slices from patient #1 for three different metrics. We see the MARGAN algorithm achieves the best slice consistency in comparison to other approaches.

Quantitative Comparison with other MAR algorithms Similar to Zhang *et al.* [Zhang and Yu, 2018], we compared our approach with three open source MAR algorithms: MAR with projection linear interpolated replacement (marLI) [Kalender et al., 1987b], beam hardening correction (marBHC)[Verburg and Seco, 2012] and NMAR [Meyer et al., 2010]. The visual assessment of the different MAR algorithms is shown in Fig. 2.7. The MARGAN approach clearly outperforms the other three MAR methods in its ability to decrease the texture changes of artifacts

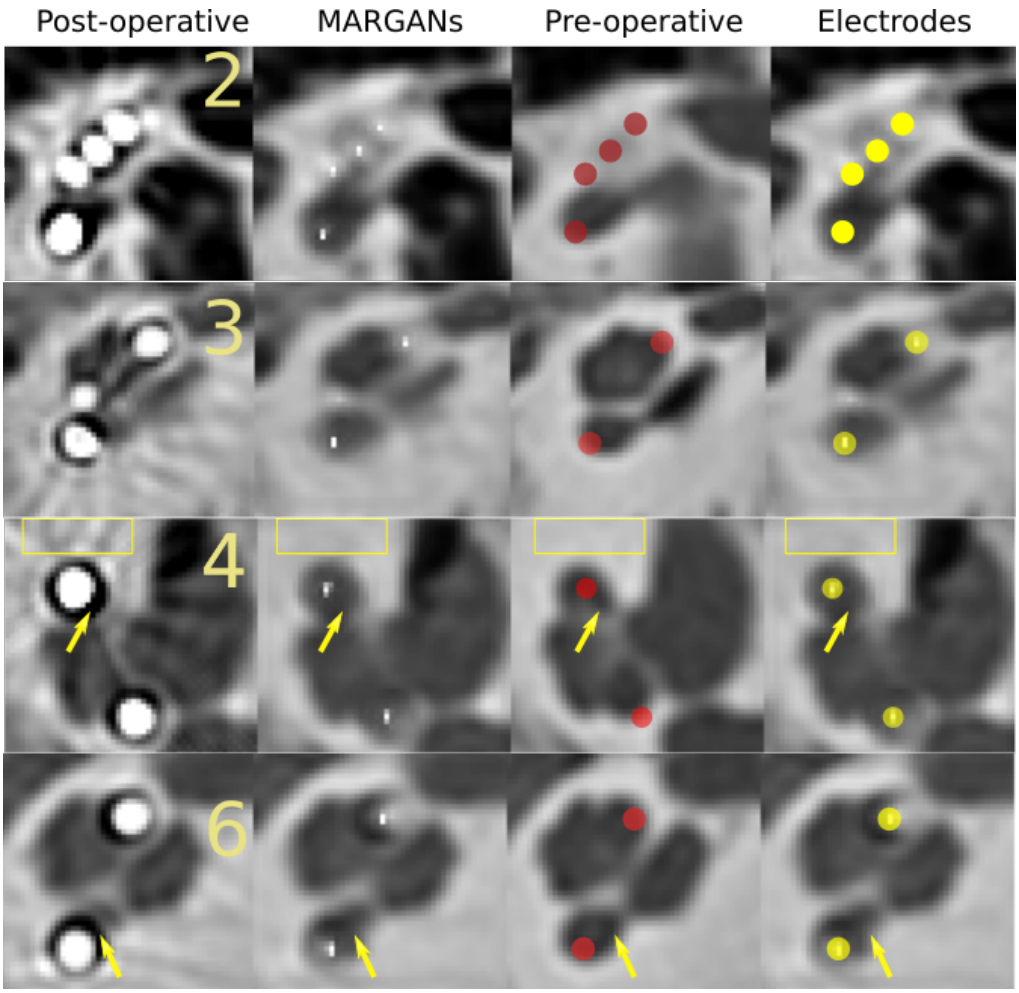

Figure 2.9: Results from patients #2 (top left), #3 (top right), #4 (bottom left) and #6 (bottom right) for two middle slices (first and second rows). The four columns correspond to: original postoperative images, output of MARGANs, registered preoperative images with manually positioned electrodes in red and postoperative images with electrodes appearing in yellow.

and to generate an image similar to the preoperative image. All 33 postoperative images were processed by marLI, marBHC, and NMAR. Three global similarity indices, the root mean square error (RMSE), structural similarity index (SSIM) and peak signal to noise ratio (PSNR), were computed between the preoperative images and the MAR images generated by the three comparison methods and our proposed approach. These three indices are reported in Table 2.4 and capture the preservation of visible structures, the errors and the quality of the reconstructed images. Our method outperforms the other MAR methods for all three metrics (lowest RMSE and largest SSIM and PSNR). In Fig. 2.8, the same indices were computed for all patient #1 image slices to evaluate the spatial consistency of the reconstruction. Clearly the MARGAN approach exhibits the best performance, with a lower mean value and much lower variance. This can be explained by the fact that it is the only MAR algorithm working directly on 3D images.

2.4.4 Impacts of methodological contributions

We assess the importance of our methodological contributions by evaluating their impact on the generated MARGAN images when they are removed from the computational pipeline. More precisely, we consider the following two contributions:

- *Retinex Loss* When zeroing the Retinex scale factor $\alpha = 0$ (instead of setting $\alpha = 5\text{e}{-5}$) during the MARGAN training, only the L_{mse} loss term is used, which is equivalent to minimizing the $L2$ norm between the generated and ground truth images. We also include in the ablation study the replacement of L_{mse} with the $L1$ norm involving $|I^{train} - G_{w_g}(I^m)|$ terms.
- *Simulation of scatter and electronic noise in artifact simulation* We simulated the image training set with only the beam hardening effect (as in [Wang et al., 2019d]) or with the full pipeline as described in section 2.2.2.

In Table 2.5, we used the three similarity measures PSNR, RMSE and SSIM with respect to the preoperative images as a way to quantify the impact of those contributions.

Table 2.5 shows that both the addition of scatter and electronic noise in the simulation and the addition of the Retinex loss can improve the performance of MAR for all three different metrics. We also see that using a single $L1$ loss function performs worse than the proposed loss combination approach. A

Table 2.5: Ablation Study of Retinex and Physical Simulation

Dataset	PSNR	RMSE	SSIM
MARGAN L1 Scatter	16.67	0.1490	0.56
MARGAN L2 Scatter	18.17	0.1257	**0.64**
MARGAN L2+Retinex No-Scatter Sim	18.02	0.1277	0.63
MARGAN L2+Retinex Scatter	**18.31**	**0.1242**	**0.64**

visualization of the image difference output obtained using different training loss functions is shown in Fig. 2.10 with subtraction maps between different output images and the ground truth image. We see from the yellow and red marks in those subtraction maps the effectiveness of the proposed Retinex loss function in comparison with using pure L1 and L2 losses.

2.4.5 Out-of-sample Test

To assess the generalization ability of this MARGAN approach, we explore its performance on 8 postoperative CBCT images, noting that the network was trained on CT images.

In Fig.2.11, we see that metal artifacts in CBCT are more extensive and complex than in CT images. Yet, the MARGAN can cope well with those CBCT images and is able to recover most of the cochlear structures.

2.4.6 CI Electrode Position Prediction

The positioning of CI electrodes in postoperative imaging provides important information for establishing a hearing prognosis [Kós et al., 2005, Todt, 2009] and can be used to improve the cochlear implant programming strategy [Noble et al., 2014]. The proposed MARGAN algorithm output images where the electrode centers are outlined by voxels in hypersignal as shown in Fig. 2.9 and 2.7. To qualitatively evaluate the positional accuracy of those electrode centers in generated MARGAN images, we use pictures of the cochlea acquired after the dissection and grinding of post-mortem temporal bones following CI surgery (see Fig. 2.12(b)). On each generated MARGAN image, a slice having roughly the same position and

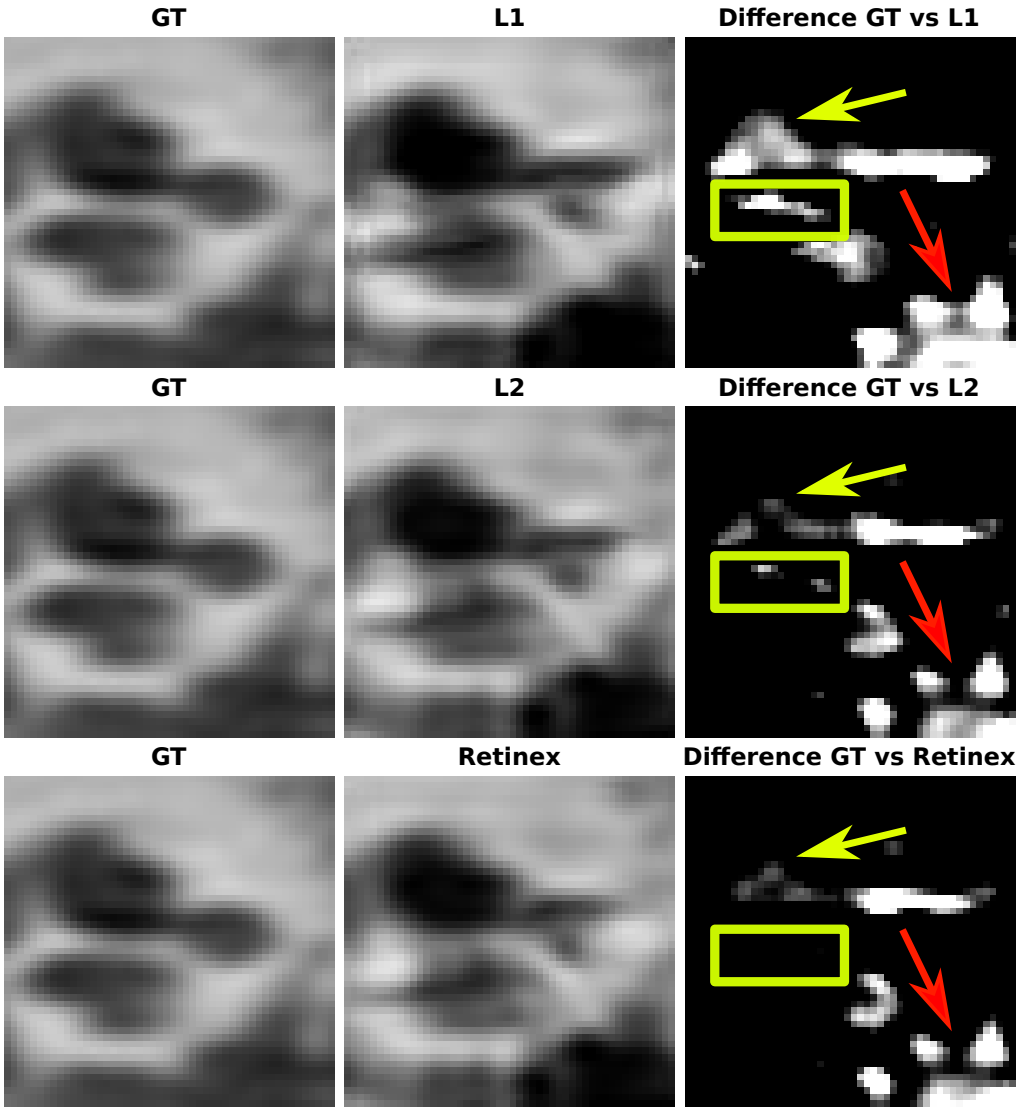

Figure 2.10: Qualitative ablation study of Retinex loss effectiveness. The first column is a middle slice of patient #1, the second column is the corresponding outputs of MARGAN with different loss functions and the last column shows subtraction maps between the first two columns.

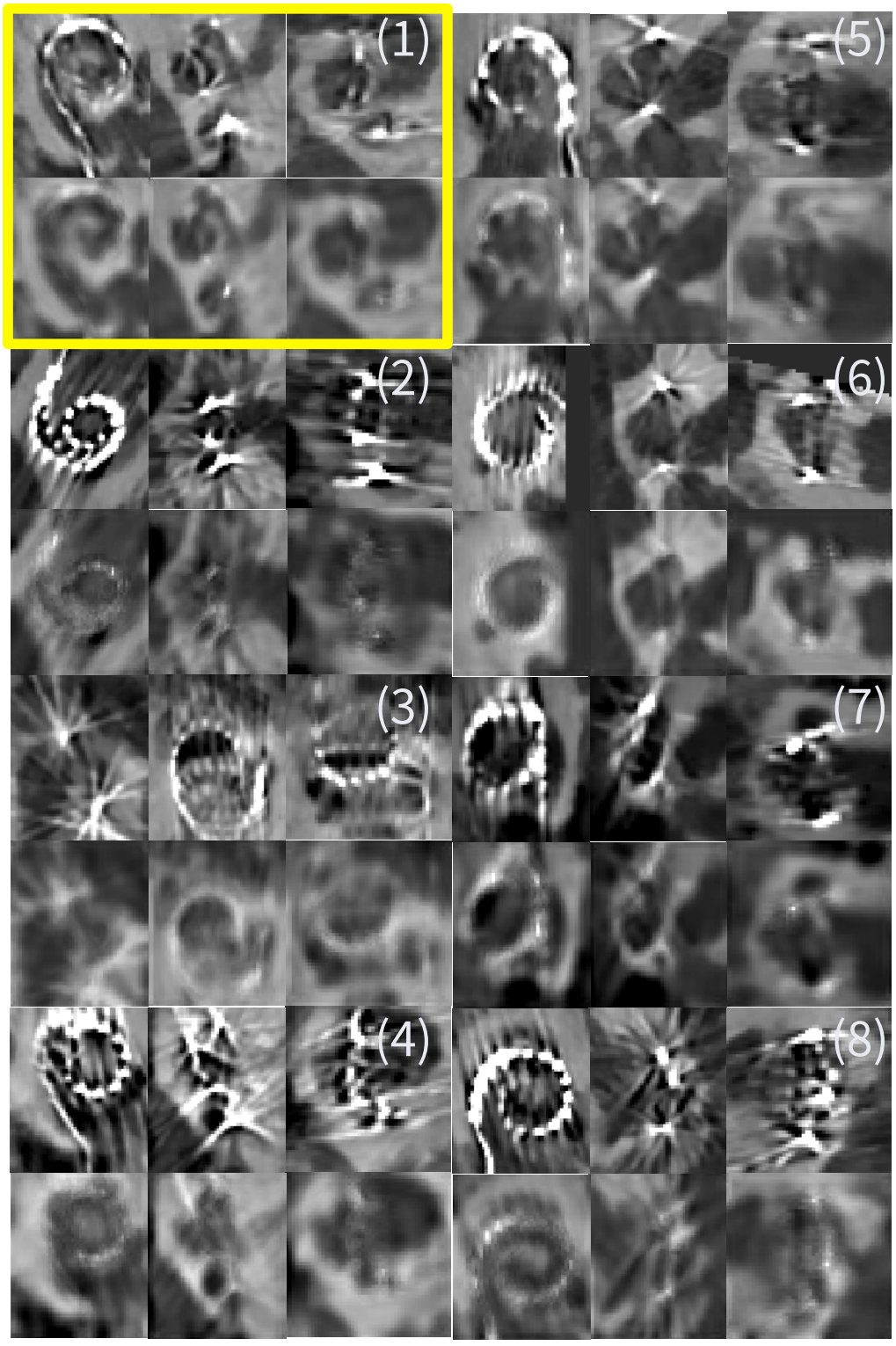

Figure 2.11: Performance of MARGAN on 8 CBCT postoperative images. The yellow box shows three views of postoperative images and MARGAN-processed images for patient #1.

orientation as the dissection picture has been manually extracted (Fig. 2.12(a)). Semi-transparent red circles have been manually positioned on the MARGAN slice at high intensity voxels while green dots have been positioned by an ENT surgeon on the electrodes visible in dissection pictures. Furthermore, those two images have been registered with an affine transform estimated after selecting two corresponding electrodes. The two registered images are fused in Fig. 2.12(c) thus showing the good overlap between green and red circles. This experiment shows that information about the position of the electrodes causing the artifacts was kept after the application of the MARGAN algorithm.

2.5 Discussion

Our MARGAN approach combines an artifact simulation pipeline with a 3D GAN network that generates *augmented* preoperative images from postoperative images. The artifact generation algorithm relies on three physical phenomena: beam hardening, scatter and electronic noise. The scatter and noise effects clearly have less impact on the output image compared to beam hardening. Yet, these effects were shown in Table 2.5 to improve the realism of the output of MARGAN when compared to preoperative images. The simulation pipeline could easily be refined in many ways, for example, using a more hardware-specific energy spectrum, increasing the number of sample energies in the approximation, or including more application-dependent scatter to primary ratios. This approach could also be extended to other imaging systems, such as cone beam CT, dual energy CT or trimodal low-dose X-ray tomography[Zanette, 2012]. The use of 3D GANs allowed us to generate MAR images with spatial coherence across neighboring slices, which is not guaranteed when using 2D slice-by-slice MAR methods. Furthermore, Retinex loss was introduced to improve the sharpness of the MAR images. We show in Tab. 2.5 that the Retinex loss can improve the performance of the MARGAN with a scale coefficient $\alpha = 5e{-}5$. However, an inappropriate α value can introduce distortions in the MARGAN output. Furthermore, the influence of other hyperparameters in the simulation pipeline on the artifact reduction needs to be further investigated.

The MARGAN approach is both data driven (for the generation of MAR images) and model driven (for the generation of training image pairs). This is in contrast to purely data-driven MAR methods that either rely on pairs of pre- and postop-

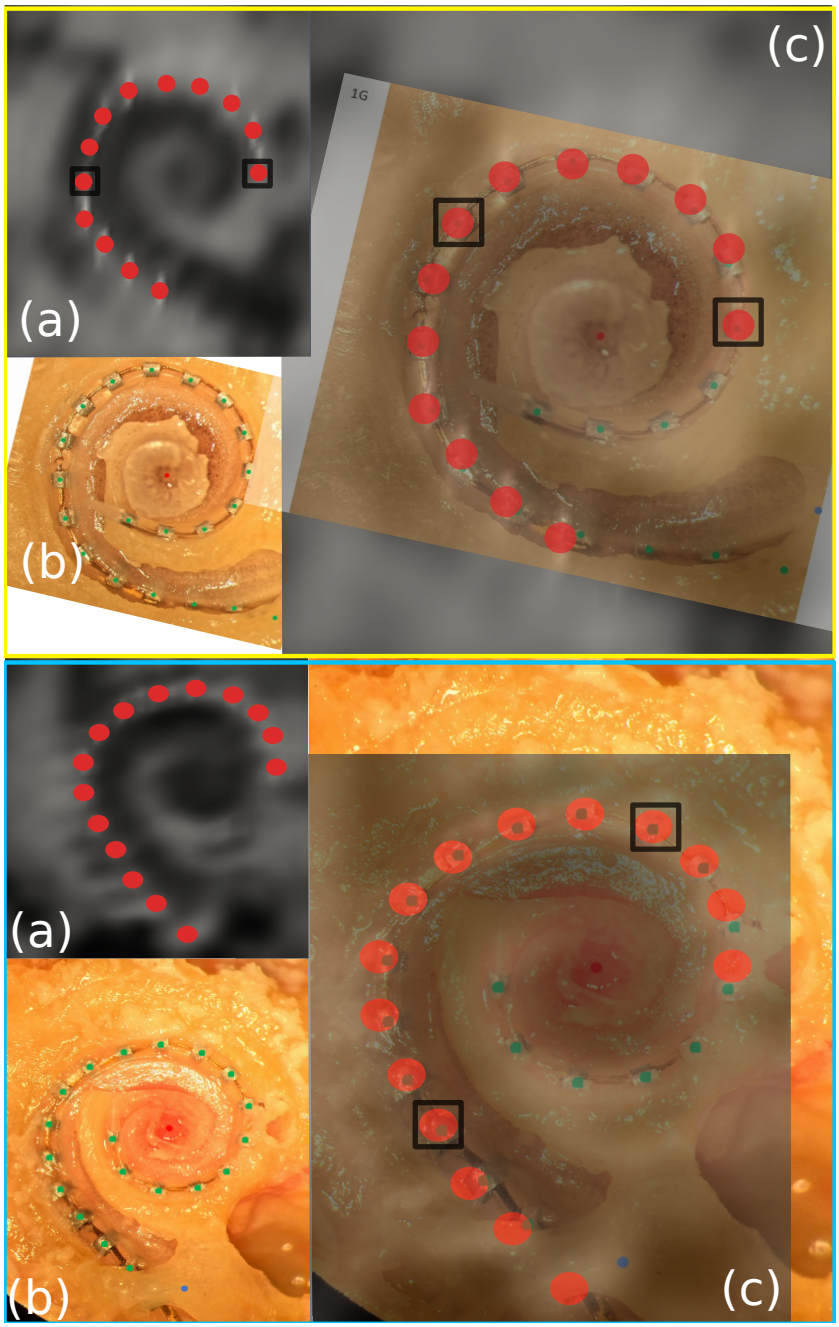

Figure 2.12: Evaluation of the electrode position after the application of the MARGAN algorithm on 2 subjects (top and bottom); (a) Reformat of a 3D MARGAN image along a plane orthogonal to the modiolar axis. Red circles were manually added at the location of high intensity voxels; (b) Image of the cochlea with electrodes inserted after dissection and grinding of the temporal bone; (c) fusion of images (a) and (b) after an affine transform based on the manual correspondence of the centers of the two circles outlined by black squares. A good overlap of green and red squares is observed.

erative images [Wang et al., 2019b] or on non-paired data [Nakao et al., 2020]. The collection of image pairs, with and without artifacts, is mostly restricted to images acquired before and after an intervention such as CI insertion. The use of such pairs makes the 3D GAN fairly effective at removing artifacts in postoperative images. However, the collection of those images may be difficult and an intra-patient image registration is required. The MARGAN pipeline aims at reaching the same efficiency but by replacing the postoperative image with a simulated one. This makes the MARGAN algorithm applicable to a larger set of clinical cases where such image pairs cannot be gathered, for instance in the case of hip, shoulder or knee prostheses. The use of CycleGAN on non-paired images as in [Nakao et al., 2020] is very appealing, because it avoids both artifact simulation and collection of paired images. However, it has only been tested to remove large artifacts, such as those caused by dental fillings, and with limited quantitative assessment.

Another advantage of the artifact simulation approach in MARGAN is its ability to augment the generated MAR image with voxels indicating the location of the metal part. In the case of CI postoperative images, it enables visualization in the same image of both the cochlea and the implant electrode centers. Note that the augmentation of the MARGAN image is only optional in this framework, because the metal-free image can replace the augmented image as I^{train} in the loss function of the 3D GAN.

Specifically, in the cochlear metal artifact reduction problem, we see from Fig. 2.7 and Tab. 2.4 that almost all the traditional MAR approaches have degradation problems in terms of reconstruction image quality. It was reported in [Meyer et al., 2009] and [Diehn et al., 2017] that sinogram inpainting-based methods can introduce new artifacts. These artifacts can have a severe impact on image quality if the metallic parts and artifacts occupy a large area of space in the image, which is typically the case for the CI electrodes discussed here. However, the risk of quality degradation is not applicable for MARGAN as the image domain-based methods do not need to access the sinogram and the Radon transform.

A limitation of MARGAN lies in the relative complexity of implementing the simulation of metal artifacts in CT images. This is especially true for the scatter effect, which only adds a marginal gain in realism to the generated images. A more thorough study should be performed to evaluate the level of realism

required in the simulation pipeline to improve the MARGAN output. Simulating the insertion of metal parts can also be complex as it requires a segmentation algorithm to locate the region of insertion. But this complexity is rewarded by the ability to generate a vast training set accounting for variations in patient anatomy or implant design.

Learning-based MAR methods were shown to outperform traditional sinogram-based MAR approaches in several previous works [Wang et al., 2019b,d, Zhang and Yu, 2018]. But by design, the performance of those supervised methods depends on the chosen training set and they are application-specific algorithms. Their integration into a clinical workflow remains to be demonstrated, in particular due to their potential lack of robustness. The successful application of MARGAN on CBCT images unseen during training is an encouraging sign of the generalization ability of MARGAN, though further studies are required.

Finally a limitation common to all MAR methods is the difficulty of evaluating performances quantitatively, due to the lack of ground truth data. The use of paired pre- and postoperative image data enables quantitative comparison through global similarity indices (such as PNSR, RMSE) but is also dependent on the registration quality of the two images. Images with synthetic artifacts created by image processing were also considered in [Nakao et al., 2020], for instance, but they are computationally intensive to reach sufficient realism. Physical anthropomorphic phantoms are a useful alternative for MAR assessment [Bolstad et al., 2018] but are limited by the number of phantoms considered.

2.6 Conclusion

In this chapter, we have introduced a simulation-based 3D GAN to attenuate metal artifacts in CT images. The network is trained on a thousand regular CT images without any artifacts and their corresponding images where metal artifacts have been simulated. We have demonstrated the introduction of scatter and electronic noise effects in addition to beam hardening in an efficient computational pipeline. The complexity of scatter simulation has been alleviated by precomputing the impact of scatter on a generic head phantom where metal parts have been introduced. A Retinex loss was introduced to enhance visible edges in the generated images. The MARGAN approach was evaluated on CT and CBCT

images of the inner ear with cochlear implants inserted. The proposed approach provided images close to preoperative images and outperformed open source MAR methods. Furthermore, images generated by MARGAN included the location of the electrode centers, which is useful for assessing the quality of implant surgery.

The trade-off between the complexity of artifact simulation and MARGAN output requires additional study, and we will also investigate the impact of MARGAN images on the automatic registration of pre- and postoperative images.

Chapter 3

Bayesian Logistic Shape Model Inference : application to cochlea image segmentation

Incorporating shape information is essential for the delineation of many organs and anatomical structures in medical images. While previous work has mainly focused on parametric spatial transformations applied on reference template shapes, in this paper, we address the Bayesian inference of parametric shape models for segmenting medical images with the objective to provide interpretable results. The proposed framework defines a likelihood appearance probability and a prior label probability based on a generic shape function through a logistic function. A reference length parameter defined in the sigmoid controls the trade-off between shape and appearance information. The inference of shape parameters is performed within an Expectation-Maximisation approach where a Gauss-Newton optimization stage allows to provide an approximation of the posterior probability of shape parameters.

This framework is applied to the segmentation of cochlea structures from clinical CT images constrained by a 10 parameter shape model. It is evaluated on three different datasets, one of which includes more than 200 patient images. The results show performances comparable to supervised methods and better than previously proposed unsupervised ones. It also enables an analysis of parameter distributions and the quantification of segmentation uncertainty including the effect of the shape model. This chapter is based on our preprint journal article [Wang et al., 2021a] which is under peer review.

3.1 Introduction

Several anatomical structures have a typical shape, such that a medical expert can easily recognize them from their three-dimensional representation. This is for instance the case of basal ganglia within the brain [Ashburner and Friston, 2005a], but also of abdominal structures, such as the liver or kidneys. Another emblematic example is the cochlea which is a small organ within the inner

ear having a remarkable spiraling configuration where mechanical waves are transformed into electrical stimulation of the auditory nerve. The cochlea shape is complex as it completes around two and a half turns with its centerline closely resembling a logarithmic spiral helix [Baker, 2008, Cohen et al., 1996]. Its segmentation from CT images of the temporal bone is challenging since those images have low resolution with respect to the anatomy of the cochlea: the cochlea dimension is about 8.5x7x4.5 mm^3 while the typical CT voxel size is larger than 0.2 mm which is weakly visible for the fine structures of the chambers.In addition, the cochlea is filled with fluids that can be found in the vestibular system and other neighbouring structures, with similar appearance in CT images.

Supervised learning (e.g. Deep Learning) is an effective way to perform image segmentation or processing in many cases. Specifically, in inner ear CT imaging analysis, many works achieved impressive results [Alshazly et al., 2019, Heutink et al., 2020, Li et al., 2021, Lv et al., 2021, Raabid et al., 2021, Wang et al., 2019c, 2020d, Zhang and Yu, 2018]. However, supervised learning methods have also many limitations. First, creating dataset annotations is time consuming, possibly preventing the creation of massive training datasets. In the cochlea case, a well trained ENT surgeon would need at least ten minutes to segment each 3D cochlea volume. Second, due to the potential overfitting related to the limited training set, the output of such supervised algorithm is likely to fall outside the shape space of the structure of interest.

Shape-based image segmentation can overcome the above limitations since the optimization of the model can be done in an unsupervised or weakly supervised way. Besides, the recovered shape parameters make a natural compact representation that is useful for shape analysis and even clinical applications. In this chapter, we consider shapes that are either defined as an explicit $\mathcal{S}(\boldsymbol{\theta}_S) \in \mathbb{R}^d$ or implicit $\mathcal{S}(\boldsymbol{\theta}_S, \mathbf{x}) = 0$ parametric shape models where $\boldsymbol{\theta}_S$ is a set of shape parameters and $\mathbf{x} \in \mathbb{R}^d$, is any point in space ($d = 2, 3$).

Those parametric shape models serve to guide the delineation of such anatomical structures by constraining the shape space of the segmented object. We can roughly split the shape-based image segmentation methods into two sets of methods. A first set optimizes the shape parameters $\boldsymbol{\theta}_S$ by minimizing the sum of a regularizing term $E_R(\boldsymbol{\theta}_S)$ and an image term $E_I(\mathcal{S}(\boldsymbol{\theta}_S), I, \theta_I)$: $\hat{\theta}_S = \arg\min_{\boldsymbol{\theta}_S} E_I(\mathcal{S}(\boldsymbol{\theta}_S), I, \theta_I) + E_R(\boldsymbol{\theta}_S)$ where θ_I is a set of image parameters that

may also be optimized. This iconic shape fitting principle is typically used in the classical active shape model [Cootes et al., 1995, Heimann et al., 2007] and their extensions [Cremers et al., 2003]. Various generic image terms may be considered for instance as those explored in [Tsai et al., 2003]. A second set of methods uses the shape model $\mathcal{S}(\boldsymbol{\theta}_S)$ as a shape prior instead of a shape space. Several shape constraints have been introduced within several image segmentation frameworks including level-sets [Chan and Zhu, 2005, Cremers, 2003], free-form deformation space [Rueckert et al., 2003a] or implicit template deformation [Prevost et al., 2013]. While those methods have greater shape flexibility for delineating structures, it is often difficult to set the coefficients weighting the shape constraint with other image terms. Those two sets of shape based segmentation methods are expressed as energy minimization problems, thus only allowing to have point estimates of shape parameters and not their posterior probabilities.

Another common shape representation consists in specifying a parametric spatial transformation $\mathcal{T}(\theta_D) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ acting on a template shape $\mathcal{S}(\theta_0) \in \mathbb{R}^d$ leading to an indirect shape parameterization : $\mathcal{S}(\theta_D) = \mathcal{T}(\theta_D) \circ \mathcal{S}(\theta_0)$. This formulation of shape modeling based on a deformable template leads to solving a joint segmentation and registration problem. More precisely, several authors [Ashburner and Friston, 2005b, Pohl et al., 2006a] defined generative image and shape models and performed statistical variational inference to optimize their parameters and hyperparameters. Priors on the deformation space based for instance on minimal elastic energy [Van Leemput, 2009], were applied on triangular or tetrahedral mesh templates. Other shape priors were defined as restricted Boltzmann machines [Agn et al., 2019] or as shape-odds [Elhabian and Whitaker, 2017]. In most cases, optimal shape parameters (e.g. mesh vertex positions) are obtained as maximum *a posteriori* but not their posterior probability. Uncertainty quantification of image registration algorithms has been tackled in some research papers [Le Folgoc et al., 2017, Simpson et al., 2012, Wang et al., 2018] based on a low dimensional representation of deformation space and Laplace approximation.

In this chapter, we propose a novel Bayesian framework for shape constrained image segmentation based on parametric shape models (instead of parametric spatial transformations) where the output segmentation is driven by a shape model but without restricting it to a low dimensional space. The proposed approach is

generic as it is suitable for any explicit $\mathcal{S}(\boldsymbol{\theta}_S)$ and implicit $\mathcal{S}(\boldsymbol{\theta}_S, \mathbf{x}) = 0$ parametric shape models associated with any appearance models representing the intensity distributions inside background or foreground regions. It is based on a logistic shape prior defined as the sigmoid of a shape function (e.g. signed distance map) defined over the image domain. Inferences of shape and intensity parameters are performed by maximizing the joint image and shape parameters probability $p(\boldsymbol{\theta}_S, \theta_I, I)$ with an Expectation-Maximization algorithm. We show that this optimization boils down to having the posterior label distribution as close as possible (in terms of Kullback Leibler divergence) from both the likelihood and shape prior distributions. A Gauss-Newton optimization method is introduced to optimize the shape parameters leading to closed form updates similarly to iterative reweighted least squares schemes. It outputs the most probable shape and imaging parameters but also an approximation of the posterior shape parameter probability which is essential for estimating the segmentation uncertainty.

This framework is applied to the problem of cochlea segmentation on CT images based on a parametric shape model with 10 parameters, and an imaging model defined as a mixture of Student's t-distributions. It results in the reconstruction of cochlea structures in 2 small datasets consisting of paired CT and μCT post-mortem images and one large dataset of nearly 200 patients CT images. We showed that the proposed framework leads to state of the art reconstruction performances as well as the recovery of consistent shape parameter distributions and the estimation of segmentation uncertainty.

The main contributions of this chapter are:

- A novel framework for image segmentation that combines probabilistic appearance and shape models. It is generically defined for parametric shape functions rather than parametric space transformations. The trade-off between the appearance and shape models is governed by an interpretable parameter : the reference length.
- A Gauss-Newton optimization method of the shape parameters which also produces a posterior approximation of those shape parameters.
- A method for uncertainty quantification of image segmentation which takes into account the shape uncertainty.
- A segmentation method of the cochlea in clinical CT images which provides

state-of-the-art results and interpretable shape parameters.

We present below the framework of the logistic shape model (section 3.2), the shape and intensity models used specifically for cochlea segmentation (section 3.3), and the segmentation results on 3 clinical and pre-clinical datasets (section 3.4).

3.2 Method

3.2.1 Shape-based Generative Probabilistic Model

We consider an observed image I consisting of N voxels $I_n \in \mathbb{R}$, $n = 1, \ldots, N$, for which we seek to solve a binary segmentation problem guided by a shape model. That model is defined either as in a parametric form as $\mathcal{S}(\boldsymbol{\theta}_S) \in \mathbb{R}^d$, $d = 2, 3$ or in an implicit form as $\mathcal{S}(\boldsymbol{\theta}_S, \mathbf{x}) = 0$. In the case of parametric shape models, one can define an associated implicit function $\text{SDM}(\mathcal{S}(\boldsymbol{\theta}_S), \mathbf{x}) = 0$ as the signed distance map defined at point $\mathbf{x}$. Therefore, we propose to unify notations for both parametric and implicit cases by stating the existence of a *shape function* $\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}) \in \mathbb{R}$ whose zero level defines a shape and whose sign indicates if a point is inside (positive) or outside (negative). Note that with this hypothesis, a shape corresponds to a (smooth) manifold of co-dimension 1 without borders, thus defining a partition of the image into inside and outside regions.

A binary label variable $Z_n \in \{0, 1\}$ is defined at each voxel specifying if voxel n belongs to the background or foreground regions. A probabilistic intensity distribution model is defined for each region $p(I_n | Z_n = k, \theta_I^k)$, $k = 0, 1$ controlled by the intensity parameter array θ_I^k. The arrays for background ($k = 0$) and foreground ($k = 1$) are concatenated into the intensity parameter array θ_I. This appearance model can be either supervised , e.g. a trained convolutional neural network, or unsupervised, e.g. a Gaussian mixture model. In the remainder, we assume the latter case and therefore we define mechanisms to optimize the appearance parameters θ_I. In the supervised case, the steps involving the update of θ_I should be ignored.

We enforce a spatial correlation between the label of each voxel by specifying their *a priori* dependence on the shape model $\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x})$. More precisely, we define

a prior probability for voxel n to belong to the foreground region as follows:

$$p(Z_n = 1|\boldsymbol{\theta}_S) = \sigma\left(\frac{\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}_n)}{l_{\text{ref}}}\right) \tag{3.1}$$
$$p(Z_n = 0|\boldsymbol{\theta}_S) = 1 - p(Z_n = 1|\boldsymbol{\theta}_S) = \sigma\left(-\frac{\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}_n)}{l_{\text{ref}}}\right)$$

where $\sigma(x)$ is the sigmoid (or logistic) function, $\mathbf{x}_n$ is the position of voxel n and l_{ref} is a reference length. With that definition, the prior probability will be close to 1 inside the object, close to 0 outside and equal to 0.5 on the shape boundary. We call this formulation of the label prior, the *logistic shape model* as it combines shape information into a probability distribution through a logistic function. This definition of the shape prior is related to several prior work in the literature such as probabilistic atlases and LogOdds maps [Pohl et al., 2006b], continuous STAPLE [Commowick and Warfield, 2009], a nd label fusion [Sabuncu et al., 2010].

The quantity l_{ref} is a characteristic length which controls the slope of the prior probability next to the object boundary. This parameter also influences the trade-off between intensity and shape information in the segmentation process as discussed in section 3.2.5. The shape parameters $\boldsymbol{\theta}_S$ are themselves regarded as random variables with a multivariate Gaussian prior controlled by hyper-parameters $\boldsymbol{\alpha}$: $p(\boldsymbol{\theta}_S|\boldsymbol{\alpha})$. The intensity parameters may also optionally be considered as random variables with hyper-parameter $\boldsymbol{\beta}$ as $p(\theta_I|\boldsymbol{\beta})$. The shape based generative model is summarized in Fig. 3.2:(a).

3.2.2 Logistic Shape Model Framework

With the proposed generative model, given an image, the objective is to infer the most probable values of the intensity $\hat{\theta}_I$ and shape parameters $\hat{\theta}_S$ which will lead to the estimation of the posterior label probabilities given by :

$$p(Z_n = 1|I_n, \theta_I, \boldsymbol{\theta}_S) = \frac{p(I_n|Z_n = 1, \theta_I^1)p(Z_n = 1|\boldsymbol{\theta}_S)}{\sum_{k=0}^{1} p(I_n|Z_n = k, \theta_I^1)p(Z_n = k|\boldsymbol{\theta}_S)} \tag{3.2}$$

That posterior probability is clearly a compromise between shape information stored in the prior $p(Z_n = 1|\boldsymbol{\theta}_S)$ and appearance information stored into the likelihood $p(I_n|Z_n = 1, \theta_I^1)$. The *segmented region of interest* (SROI) then

corresponds to voxels for which $p(Z_n = 1|I_n, \theta_I, \boldsymbol{\theta}_S) \geq \frac{1}{2}$. In addition, the logistic shape model framework recovers the most likely shape parameters $\hat{\theta}_S$ that corresponds to the *segmented shape instance* (SSI) which is the best fit of the shape model in that image. Finally, we will show that we can approximate the posterior shape parameter $p(\boldsymbol{\theta}_S|I)$ in order to capture the uncertainty in the shape parameter estimation.

The optimization of the intensity and shape parameters is done by maximizing the the log-joint intensity and parameters probability :

$$\begin{aligned}(\hat{\theta}_S, \hat{\theta}_I) &= \arg\max_{\boldsymbol{\theta}_S, \theta_I} \log p(I, \boldsymbol{\theta}_S, \theta_I) = \arg\max_{\boldsymbol{\theta}_S, \theta_I} \mathcal{L}(\boldsymbol{\theta}_S, \theta_I) \\ \mathcal{L}(\boldsymbol{\theta}_S, \theta_I) &= \log p(I|\boldsymbol{\theta}_S, \theta_I) + \log p(\boldsymbol{\theta}_S) + \log p(\theta_I) \\ &= \sum_{n=1}^{N} \log \left(\sum_{k=0}^{1} p(I_n|Z_n = k, \theta_I)\ p(Z_n = k|\boldsymbol{\theta}_S) \right) + \\ &\quad \log p(\boldsymbol{\theta}_S) + \log p(\theta_I)\end{aligned} \tag{3.3}$$

In the log-joint probability $\mathcal{L}(\boldsymbol{\theta}_S, \theta_I)$ we have marginalized out the hidden label variables Z_n and used the conditional independence of variables I_n given $\boldsymbol{\theta}_S$.

3.2.3 Expectation-Maximization Inference

The direct optimization of $\mathcal{L}(\boldsymbol{\theta}_S, \theta_I)$ can be done by any optimization toolbox but it is difficult due to the possible encountered overflows/underflows caused by the log-sum-exp expressions.

This is why we propose to follow the Expectation-Maximization (EM) algorithm which relaxes that optimization problem into several optimizations over simpler problems. We proceed by introducing N variables u_n that are surrogates for the posterior label probability $p(Z_n = 1|I_n, \boldsymbol{\theta}_S, \theta_I)$ such that $u_n \in [0, 1]$. Writing $U = \{u_n\}$, we introduce a new augmented criterion $\mathcal{L}^*(\boldsymbol{\theta}_S, \theta_I, U) = \log p(I, \boldsymbol{\theta}_S, \theta_I) - D_{\mathrm{KL}}(U||p(Z|I, \boldsymbol{\theta}_S, \theta_I))$ by adding the negative Kullback-Leibler divergence between u_n and the posterior label $p(Z_n|I_n, \boldsymbol{\theta}_S, \theta_I)$.

Maximizing $(\boldsymbol{\theta}_S, \theta_I, U)$ over the augmented criterion $\mathcal{L}^*(\boldsymbol{\theta}_S, \theta_I, U)$ leads to the same optima in $(\boldsymbol{\theta}_S, \theta_I)$ than the maximization of $\mathcal{L}(\boldsymbol{\theta}_S, \theta_I)$ but with simpler

expressions:

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\theta}_S, \theta_I, U) &= \sum_{n=1}^{N}\sum_{k=0}^{1} u_n^k \log(p(I_n|\boldsymbol{\theta}_S, \theta_I)p(Z_n^k = k|I_n, \boldsymbol{\theta}_S, \theta_I)) \\ &\quad - \sum_{n=1}^{N}\sum_{k=0}^{1} u_n^k \log u_n^k + \log p(\boldsymbol{\theta}_S) + \log p(\theta_I) \\ &= Q(U, \boldsymbol{\theta}_S, \theta_I) + \sum_{n=1}^{N} H(u_n) + \log p(\boldsymbol{\theta}_S) + \log p(\theta_I)\end{aligned}$$

where $Q(U, \boldsymbol{\theta}_S, \theta_I) = \mathbb{E}_U(\log p(I, Z|\boldsymbol{\theta}_S, \theta_I))$ is the conditional expectation of the complete marginal log-likelihood (a.k.a. evidence) and $H(u_n)$ is the entropy of variable u_n. The quantity $Q(U, \boldsymbol{\theta}_S, \theta_I)$ is a lower bound of the log-likelihood since $H(u_n) > 0$.

The maximization of the augmented criterion $\mathcal{L}^*(\boldsymbol{\theta}_S, \theta_I, U)$ is performed by the successive maximization over the U, θ_I and $\boldsymbol{\theta}_S$ variables. The **E-step** corresponds to the maximization of $\mathcal{L}^*(\boldsymbol{\theta}_S, \theta_I, U)$ with respect to U which sets the surrogate variable U to the posterior label probability $u_n = p(Z_n = 1|I_n, \boldsymbol{\theta}_S, \theta_I)$.

The **MI-step** optimizes the log-joint probability with respect to the appearance variables θ_I, which is equivalent to the maximization of $\mathcal{L}_I = -D_{\mathrm{KL}}(U||p(I|Z, \theta_I)) + \log p(\theta_I|\boldsymbol{\beta})$. When the appearance parameters are independent between classes, then $\log p(\theta_I|\boldsymbol{\beta}) = \sum_{k=0}^{K} \log p(\theta_I^k|\boldsymbol{\beta}_k)$ and the MI-step splits into 2 independent maximization over θ_I^k, $k = 0, 1$ of $\mathcal{L}_I^k = -\sum_{n=1}^{N} D_{\mathrm{KL}}(u_n^k||p(I_n|Z_n = e_k, \theta_I^k)) + \log p(\theta_I^k|\boldsymbol{\beta}_k)$. For certain well chosen intensity models such as Gaussian mixture models, this optimization leads to closed-form updates of θ_I.

Finally, we perform the **MS-step** corresponding to the maximization over shape variables $\boldsymbol{\theta}_S$ which is equivalent to the maximization of $\mathcal{L}_S$:

$$\mathcal{L}_S = -D_{\mathrm{KL}}(U||p(Z|\boldsymbol{\theta}_S)) + \log p(\boldsymbol{\theta}_S|\boldsymbol{\alpha})$$

We can see that the EM algorithm preserves an interesting symmetry between shape and appearance information. Indeed, the iterative application of the E, MS and MI steps makes the posterior labels distribution U as close (in terms of KL divergence) as possible from the likelihood $p(I|Z, \theta_I)$ and shape prior $p(Z|\boldsymbol{\theta}_S)$ that the minimization of $D_{\mathrm{KL}}(U||p(Z|\boldsymbol{\theta}_S)) + D_{\mathrm{KL}}(U||p(I|Z, \theta_I))$. At convergence,

the posterior distribution is therefore clearly a compromise between shape and appearance information.

3.2.4 Optimization of shape parameters $p(\boldsymbol{\theta}_S|I)$

The functional $\mathcal{L}_S$ is a non trivial function of the parameters $\boldsymbol{\theta}_S$ as it combines 2 non-linear functions : the sigmoid $\sigma()$ and the shape function $\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}_n)$:

$$\mathcal{L}_S = -\sum_{n=1}^{N} \left(u_n \log \sigma \left(\frac{\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}_n)}{l_{\text{ref}}} \right) + (1 - u_n) \log \sigma \left(-\frac{\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}_n)}{l_{\text{ref}}} \right) \right) + \log p(\boldsymbol{\theta}_S|\boldsymbol{\alpha}) + \text{cst} \tag{3.4}$$

The functional gradient $\nabla_{\boldsymbol{\theta}_S} \mathcal{L}_S$ cannot be written in closed form since it requires the computation of the gradient of the scaled shape function at each voxel : $\mathbf{d}_n = \frac{\nabla_{\boldsymbol{\theta}_S} \tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}_n)}{l_{\text{ref}}} \in \mathbb{R}^{|\boldsymbol{\theta}_S|}$. Those gradient vectors may be computationally costly to compute, for instance when the shape function is based on a signed distance map of parametric shape models $\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}) = \text{SDM}(\mathcal{S}(\boldsymbol{\theta}_S), \mathbf{x})$. In that case, the $\mathbf{d}_n$ values are computed by a costly finite difference approximation except for translation and rotation parameters for which they can be computed efficiently (see 3.10.1). After combining all $\mathbf{d}_n$ terms in a gradient matrix $\mathbf{d} \in \mathbb{R}^{|\boldsymbol{\theta}_S| \times N}$, the functional gradient can be simplified as $\nabla_{\boldsymbol{\theta}_S} \mathcal{L}_S = -\mathbf{d}(\mathbf{u} - \boldsymbol{\mu}) + \nabla_{\boldsymbol{\theta}_S} \log p(\boldsymbol{\theta}_S|\alpha)$ where $\mathbf{u} = (u_1 \dots u_N)^T \in \mathbb{R}^N$ and $\boldsymbol{\mu} = \left(\sigma \left(\frac{\tilde{\mathcal{S}}(\boldsymbol{\theta}_S^i, \mathbf{x}_1)}{l_{\text{ref}}} \right) \dots \sigma \left(\frac{\tilde{\mathcal{S}}(\boldsymbol{\theta}_S^i, \mathbf{x}_n)}{l_{\text{ref}}} \right) \right)^T \in \mathbb{R}^N$.

Thus, a first approach for optimizing the shape parameters is to use any quasi-Newton optimization method such as the BFGS algorithm (similarly to [Demarcy, 2017b]), since it only requires the computation of the functional gradient and iteratively estimates the Hessian matrix. Yet, this generic optimization was found to be fairly time consuming and sometimes unstable.

Instead, we propose to adopt a Gauss-Newton optimization approach where we approximate the Hessian matrix by ignoring the term involving second order derivatives. More precisely, the Hessian of the functional is computed as $\mathbf{H} = \nabla^2_{\boldsymbol{\theta}_S} \mathcal{L}_S = -\nabla_{\boldsymbol{\theta}_S} \mathbf{d} \otimes (\mathbf{u} - \boldsymbol{\mu}) - \mathbf{d} \otimes \nabla_{\boldsymbol{\theta}_S} \boldsymbol{\mu} + \nabla^2_{\boldsymbol{\theta}_S} \log p(\boldsymbol{\theta}_S|\alpha)$. After dropping the first term, we get the following approximate Hessian $\mathbf{H} \approx \tilde{\boldsymbol{H}} = -\mathbf{d} \otimes \nabla_{\boldsymbol{\theta}_S} \boldsymbol{\mu} + \nabla^2_{\boldsymbol{\theta}_S} \log p(\boldsymbol{\theta}_S|\alpha)$. When inserting the expression of the gradient of the prior, we get : $\tilde{\boldsymbol{H}} = \mathbf{d}\, \text{Diag}(\boldsymbol{\mu} \circ (1 - \boldsymbol{\mu}))\, \mathbf{d}^T + \nabla^2_{\boldsymbol{\theta}_S} \log p(\boldsymbol{\theta}_S|\alpha)$ where $\circ$ is the element-wise product between two vectors. This approximate Hessian matrix is positive definite

by construction and is then used to perform several Newtons steps.

The sketch of the MS step is shown as algorithm 1 where the shape parameter prior $p(\delta\boldsymbol{\theta}_S^i|\boldsymbol{\alpha})$ is arbitrarily chosen as a zero mean Gaussian distribution with covariance $\Sigma_{\boldsymbol{\theta}_S}^0$. It consists of two intertwined loops, the innermost performing iteratively the Newton updates and updating the mean, gradient and Hessian values. The outer loop updates the shape function gradient which is potentially a costly step. In line 15 of the algorithm, the U variable is updated in an E-step in order to speed-up the convergence of the overall EM algorithm. Since the parameter range is bounded, we perform in practice a truncated Newton step as proposed in [Nash, 1984].

This Gauss-Newton approach was inspired by the iterative re-weighted least squares algorithm [Bishop, 2006] developed for solving logistic regression (LR) problems. Indeed the first term of $\mathcal{L}_S$ is similar to the log likelihood of LR after replacing u_n with a binary variable and linearizing the shape function. The proposed approach is also related to the Fisher scoring algorithm (see [Sourati et al., 2019] as an example in medical image analysis) when the point-wise Hessian matrix of the log likelihood is replaced by its expectation thus leading to more stable evaluation. In this particular case, the approximate Hessian is not the expectation of the Hessian since the first term of $\mathcal{L}_S$ is the expectation of the log-prior with respect to binary variable U instead of Z.

Finally, the proposed algorithm also outputs a Laplace approximation of the shape parameter posterior $p(\boldsymbol{\theta}_S|I)$ as a Gaussian distribution where the mean is the optimized shape parameter $\boldsymbol{\theta}_S^\star$ and the covariance is the inverse approximate Hessian matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_S}^\star = (\tilde{\boldsymbol{H}})^{-1}$.

The overall optimization finally consists in iterating a series of outer loop, each loop consisting in optimizing the shape parameters as in Alg. 1 then followed by a series of MI-steps until the relative change of intensity parameters is less than a threshold. The stopping criterion for the outer loop is the relative change of foreground intensity parameters as it is the most impactful parameter.

3.2.5 Influence of the characteristic length l_{ref}^k

Based on Eq.3.2.5 and Eq.3.1, it is easy to see that for infinitely small value of the characteristic length $l_{\text{ref}} \to 0$, then the label prior becomes more and more sharp

Algorithm 1 MS step to compute $p(\boldsymbol{\theta}_S|I)$

$i \leftarrow 0$ $u_n = p(Z_n = 1|I, \boldsymbol{\theta}_S)$ // E-Step, Update U

repeat

  $\boldsymbol{V} \leftarrow \frac{\tilde{\mathcal{S}}(\boldsymbol{\theta}_S^i, \mathbf{x}_n)}{l_{\text{ref}}} \in \mathbb{R}^N$ // Shape function

  $\mathbf{d} \leftarrow \frac{\nabla_{\boldsymbol{\theta}_S} \tilde{\mathcal{S}}(\boldsymbol{\theta}_S^i, \mathbf{x}_n)}{l_{\text{ref}}} \in \mathbb{R}^{|\boldsymbol{\theta}_S| \times N}$ // Shape function gradient

  $\delta\boldsymbol{\theta}_S^0 \leftarrow \mathbf{0}$, $t \leftarrow 0$ **repeat**

    $\boldsymbol{\mu} \leftarrow \sigma(\boldsymbol{V} + \mathbf{d}^T \boldsymbol{\delta\theta}_S^t)$ // Current Prior probability

    $\mathbf{g} \leftarrow -\mathbf{d}(\mathbf{u} - \boldsymbol{\mu}) - (\boldsymbol{\Sigma}_{\boldsymbol{\theta}_S}^0)^{-1} \delta\boldsymbol{\theta}_S^t$ // Functional Gradient

    $\tilde{\boldsymbol{H}} \leftarrow \mathbf{d}\,\text{Diag}(\boldsymbol{\mu} \circ (1 - \boldsymbol{\mu}))\ \mathbf{d}^T - (\boldsymbol{\Sigma}_{\boldsymbol{\theta}_S}^0)^{-1}$ // Approximate Functional Hessian

    $\boldsymbol{\Sigma}^\star \leftarrow (\tilde{\boldsymbol{H}})^{-1}$ // Covariance

    $\delta\boldsymbol{\theta} \leftarrow -\boldsymbol{\Sigma}^\star \mathbf{g}$ // Truncated Gauss Newton Update

    $\delta\boldsymbol{\theta}_S^{t+1} \leftarrow \delta\boldsymbol{\theta}_S^t + \delta\boldsymbol{\theta}$, $t \leftarrow t + 1$ // Update shape parameters

  until $\|\delta\boldsymbol{\theta}\|/\|\boldsymbol{\theta}\| < \epsilon$

  $u_n = p(Z_n = 1|I, \boldsymbol{\theta}_S)$ // E-Step, Update U

  $\boldsymbol{\theta}_S^{i+1} \leftarrow \boldsymbol{\theta}_S^i + \delta\boldsymbol{\theta}_S^{t+1}$, $i \leftarrow i + 1$ // end inner loop

until $\|\delta\boldsymbol{\theta}_S^{t+1}\|/\|\boldsymbol{\theta}_S^{t+1}\| < \epsilon$

$\boldsymbol{\theta}_S^\star \leftarrow \boldsymbol{\theta}_S^i$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_S}^\star = \boldsymbol{\Sigma}^\star$ // Gaussian posterior

$p(Z_n = 1|\boldsymbol{\theta}_S) \rightarrow \delta_{\text{SDM}(\mathcal{S}(\boldsymbol{\theta}_S), \mathbf{x}) > 0}$ and the label posterior becomes equal to the label posterior : $p(Z_n = 1|\boldsymbol{\theta}_S, \theta_I, I_n) \longrightarrow p(Z_n = 1|\boldsymbol{\theta}_S)$. Conversely, for infinitely large value of the characteristic length $l_{\text{ref}} \longrightarrow \infty$, the label prior becomes uninformative $p(Z_n = 1|\boldsymbol{\theta}_S) \longrightarrow \frac{1}{2}$ and the label posterior converges towards the appearance driven label posterior : $p(Z_n = 1|\boldsymbol{\theta}_S, \theta_I, I_n) \longrightarrow p(I_n|Z_n = 1, \theta_I^1)/(p(I_n|Z_n = 0, \theta_I^0) + p(I_n|Z_n = 1, \theta_I^1))$. Therefore the characteristic length controls the relative influence of the shape and appearance information in the probability of assigning a label.

Since it is scaling the signed distance function, l_{ref} can be interpreted as controlling how far the resulting shape given by $p(Z_n = e_1|\boldsymbol{\theta}_S, \theta_I, I_n) = 0.5$ is allowed to deviate from the reference shape given by $\mathcal{S}(\boldsymbol{\theta}_S)$. More precisely, assuming a uniform distribution of the appearance label probability between 0 and 1, one can compute the expectation of the posterior probability for a voxel located as a

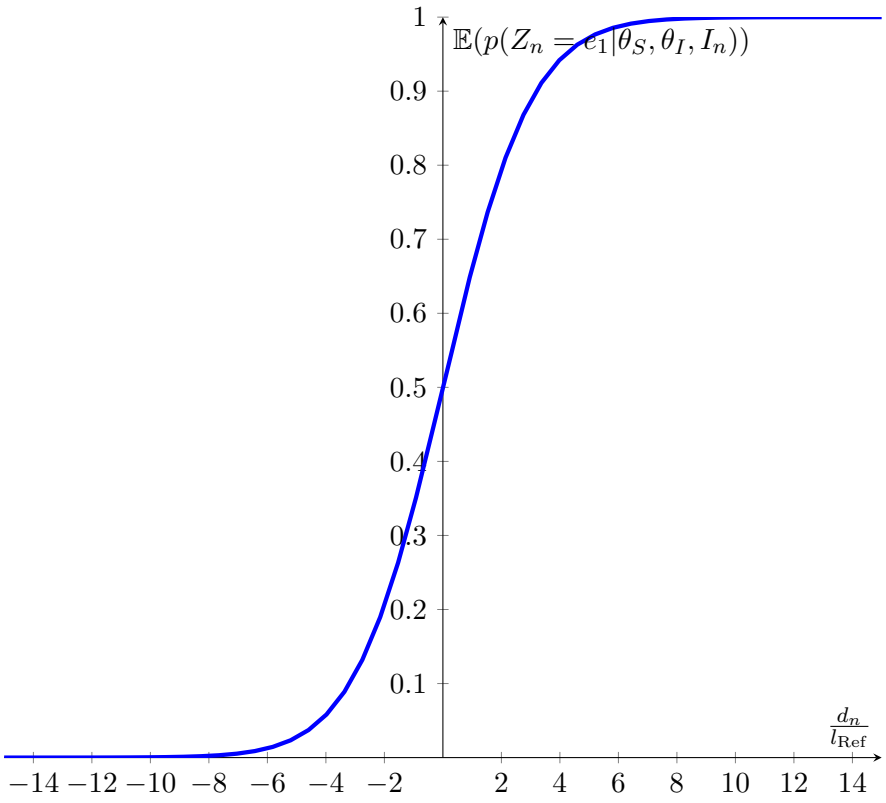

Figure 3.1: Expected label posterior probability as function of the normalized signed distance from the reference shape.

distance d_n from the reference shape :

$$\mathbb{E}(p(Z_n = 1|\boldsymbol{\theta}_S, \theta_I, I_n)) = \int_0^1 \frac{tS(\Delta_n)}{tS(\Delta_n) + (1-t)(1-S(\Delta_n))}\, dt$$
$$= \frac{1 - \Delta_n e^{-\Delta_n} - e^{-\Delta_n}}{(e^{-\Delta_n} - 1)^2} \quad , \quad \Delta_n = \frac{d_n}{l_{\text{ref}}}$$

Based on the graph of Fig.3.1, a voxel located at least at $4l_{\text{ref}}$ inside the boundary of the reference shape $\mathcal{S}(\boldsymbol{\theta}_S)$ ($p < -4$) will have in average at least 95% probability to be classified as belonging in the object.

3.3 Application to Cochlea Shape Recovery

3.3.1 Cochlea shape model

We use a parametric cochlea shape model which is controlled by a set of 4 deformable shape parameters $\boldsymbol{\theta}_{SD} : \{a, b, \alpha, \varphi\}$ as shown in Fig:(3.3). Those 4 parameters control the deformation of the centerline of the cochlea represented as a generalized cylinder and is detailed in 3.10.2. In addition to those 4 *deformable* parameters, we consider the 6 pose parameters $\boldsymbol{\theta}_{SR}$ consisting of rotation $\{rx, ry, rz\}$ (parameterizing a rotation vector) and translation $\{tx, ty, tz\}$ values. Therefore, the total number of shape parameters is 10, controlling the rigid and non rigid (deformable) motion: $\boldsymbol{\theta}_S = \boldsymbol{\theta}_{SD} \cup \boldsymbol{\theta}_{SR}$. To fit in our framework, signed

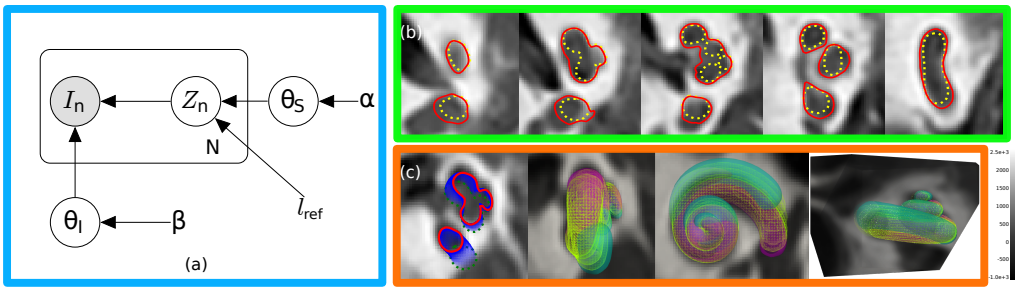

Figure 3.2: (a) graphical model for the shape-based generative model; (b) Cochlea segmentation on CT images is shown in solid red with the associated shape model in dashed yellow lines; (c) Evolution of the cochlea shape model during several MS steps shown as 2D contours (from dotted green to solid red) and 3D models.

distance map $\text{SDM}(\mathcal{S}(\boldsymbol{\theta}_S), \mathbf{x})$ from the cochlea triangular mesh surface must be created. This can be performed for instance by using VTK functions [Maurer

et al., 2003] but that distance map generation may take several seconds on large volumetric images. This is why we have developed a convolution neural network, noted as DLSDM, which outputs an approximation of the signed distance map from the set of deformable shape parameters in few milliseconds on CPU [Wang et al., 2020b].

Figure 3.3: Parametric shape model of the cochlea. (Left) Effect of the radial parameters a (red), and b (yellow) are shown with the reference position in purple; (Right) Effect of the longitudinal parameters α (pink) and φ (blue) parameters.

3.3.2 Cochlea Appearance model

Appearance models describe the intensity patterns inside the foreground and the background classes and can be built in a supervised, semi-supervised or unsupervised manner. Many simple generative models such as Gaussian mixture models (GMM) with spatial corrections [Ashburner and Friston, 2005b, Pohl et al., 2006a] have been proposed in the literature to describe tissue intensity distributions. For the cochlea segmentation in CT images, we propose an unsupervised approach based on mixture of mixtures of Student's t-distributions, i.e. each background and foreground regions are described as mixtures of Student's t-distributions. Those t-distributions are generalized Gaussian distributions with heavy tails and lead to more robust estimations than GMM since they are less sensitive to extreme intensity values [Peel and McLachlan, 2000]. In this context, the probability of observing intensity I_n knowing the label Z_n is parameterized as :

$$p(I_n|Z_n = k, \theta_I) = \sum_{m=1}^{M_k} \pi_m^k \, t(I_n|\mu_m^k, \sigma_m^k, \nu_m^k) \,, \tag{3.5}$$

where M_k corresponds to the number of mixture components for the class k and mixture coefficients π_m^k are positive and sum to one $\sum_{m=1}^{M_k} \pi_m^k = 1$. The mean parameter μ_i^k, standard deviation coefficient σ_i^k and degrees of freedom ν_i^k are

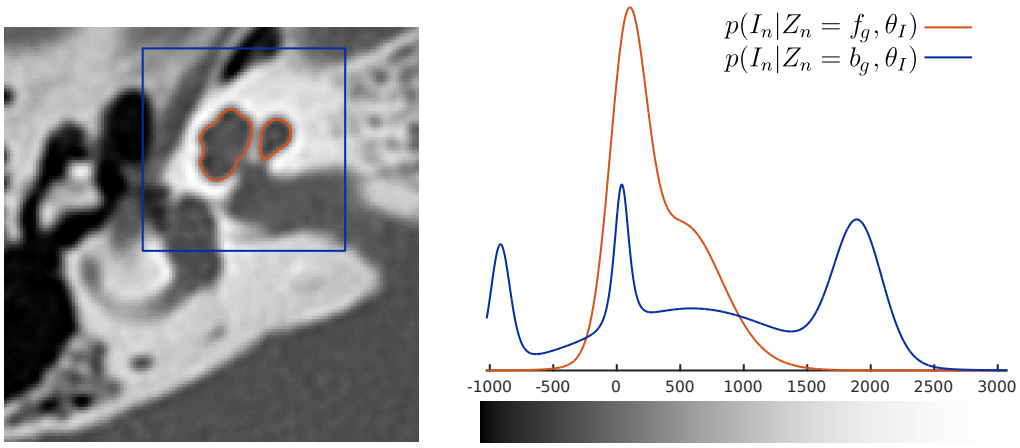

Figure 3.4: Example of intensity probability distributions of the foreground (f_g, in red) and the background (b_g, in blue) as functions of the Hounsfield unit.

parameters of the Student's t-distribution defined as:

$$t(I_n|\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}\sigma} \left(1 + \frac{(I_n-\mu)^2}{\sigma^2\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}, \tag{3.6}$$

where $\Gamma(\cdot)$ is the gamma function. To write the likelihood of this Student's t-distribution mixture of mixtures, we introduce a new categorical variable τ_{nkm} which is a binary 1-of-Mk encoding such that $\tau_{nkm} = 1$ if voxel n belongs to the m-th component of region k, and $\sum_{m=1}^{M_k} \tau_{nkm} = 1$. The likelihood then writes as:

$$p(I_n|Z_n,\tau_n) = \prod_{k=0}^{1} \prod_{m=1}^{M_k} \left[\left(t(I_n|\mu_m^k,\sigma_m^k,\nu_m^k)\right)^{\tau_{nkm}}\right]^{Z_{nk}}$$

The inference is performed with closed-form updates of all parameters [Bishop, 2006, Peel and McLachlan, 2000] after writing the Student's t-distribution as a Gaussian scale mixture. The total number of parameters to estimate is then $|\theta_I| = 4(M_0 + M_1)$. For the cochlea segmentation problem, we assume that the cochlea region mainly consists of two components ($M_1 = 2$) : the fluid (perilymph and endolymph) component centered around 0 HU and the bony walls centered around 500 HU. For the cochlea background, we consider 4 components ($M_0 = 4$) centered around 0 HU (fluid), 2000 HU (bony labyrinth), -1000 HU (air due

to pneumatization) and 600 HU (temporal bone). The corresponding initial distribution of intensity in the background and foreground regions are shown in Fig.3.4 and the exact initialization values are provided in 3.10.3.

3.4 Results

3.4.1 Synthetic Images

We provide a 2D synthetic example to illustrate the influence of the reference length l_{ref} in the proposed segmentation algorithm. We consider the segmentation of an ellipse with Gaussian intensity distribution on both background and foreground (see Fig. 3.5 (Top Left)) by using a circle prior shape $\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x}) = \|\mathbf{x} - \mathbf{C}\|^2 - R^2$. It illustrates the frequent case where the parametric model used as a prior is far simpler than the shape visible in the image. The intensity model consists of two Gaussian distributions initialized with mean and variance offsets and the circle is parameterized by its center coordinates and radius. The trade-off between imaging information (leading to an ellipse) and prior shape (leading to a circle) is controlled by the l_{ref} parameter. The log-likelihood as a function of l_{ref} exhibits a single maximum for $l_{\text{ref}} = l_{\text{opt}} = 0.021$ (Fig. 3.5 (Middle)) corresponding to the white circle in Fig. 3.5 (Left) and to the posterior label distribution in Fig. 3.5 (Right). The resulting segmentation is the isocontour $p(Z_n = 1) = 0.5$, displayed as a yellow curve in Fig. 3.5 (Left), which closely matches the elliptic shape except at its flat part (see arrow). This optimal value of l_{ref} corresponds to a configuration where the area of the circle is roughly equal to the area of the ellipse. A value of $l_{\text{ref}} < l_{\text{opt}}$ leads to isocontours $p(Z_n = e_1) = 0.5$ that fit more closely the ellipse whereas $l_{\text{ref}} > l_{\text{opt}}$ leads to isocontours that look more like a circle.

3.4.2 Inner Ear Datasets

The evaluation of the proposed approach is studied on 3 different datasets.

Dataset #1 includes spiral CT temporal bone images of 210 patients from the radiology department of Nice University Hospital of size $512 \times 512 \times 178$ corresponding to a voxel size of $0.185mm, 0.185mm, 0.25mm$. They have then been registered to a reference image via an automatic pyramidal blocking-matching

(APBM) algorithm [Ourselin et al., 2000b] from the software MedInria [Toussaint et al., 2007] followed by an image reformatting around the cochlea to the dimension $(60, 50, 50)$ with isotropic voxel size of $0.2mm$. The relatively robust registration provides a rough alignment of the cochlea visible in the input image with a cochlea reference frame. From that dataset, 5 CT images were manually segmented by an ENT surgeon (see section 3.4.4).

Dataset #2 includes 9 cadaveric cochlea spiral CT images acquired at the face and neck institute at Nice University Hospital with the same size and voxel spacing as dataset #1. In addition to CT images, high resolution X-ray microtomography (a.k.a. μCT) images with dimension of $(1035, 800, 1095)$ and isotropic voxel spacing of $0.02479mm$ were acquired each subject. The 9 μCT and spiral CT images have been registered together as shown in Fig 3.6 and reformatted around the cochlea to the same physical size as for dataset #1 (i.e. $12mm, 10mm, 10mm$). The cochlea and its two scala have been segmented on both CT and μCT images by an ENT surgeon with a semi-interactive tool [Criminisi et al., 2008]. The high resolution μCT masks serve as ground truth information for the location of the cochlea.

Dataset #3 is a human bony labyrinth dataset [Wimmer et al., 2019] which includes 22 bony labyrinth CT images and their corresponding μCT images having isometric voxel size respectively of $0.1562mm$ and $0.0607mm$. Those images were preprocessed and reformatted as for dataset #2 and also contains manually segmented cochlea masks.

3.4.3 Quantitative evaluation of segmentation on post-mortem μCT/CT datasets #2 and #3

Baseline Approach: We have implemented a 3D atlas based segmentation approach and applied it on dataset #2 and #3 to get a baseline accuracy in terms of Dice score. To this end, we randomly select one image from each dataset as template image and for each input image we perform a multiscale demons deformable registration [Vercauteren et al., 2007b] (as implemented in SimpleITK 1.0.1) to estimate the deformation field. The segmented mask of the template is deformed to match the target image. The average Dice scores are 0.63 for dataset #2 and 0.68 for dataset #3.

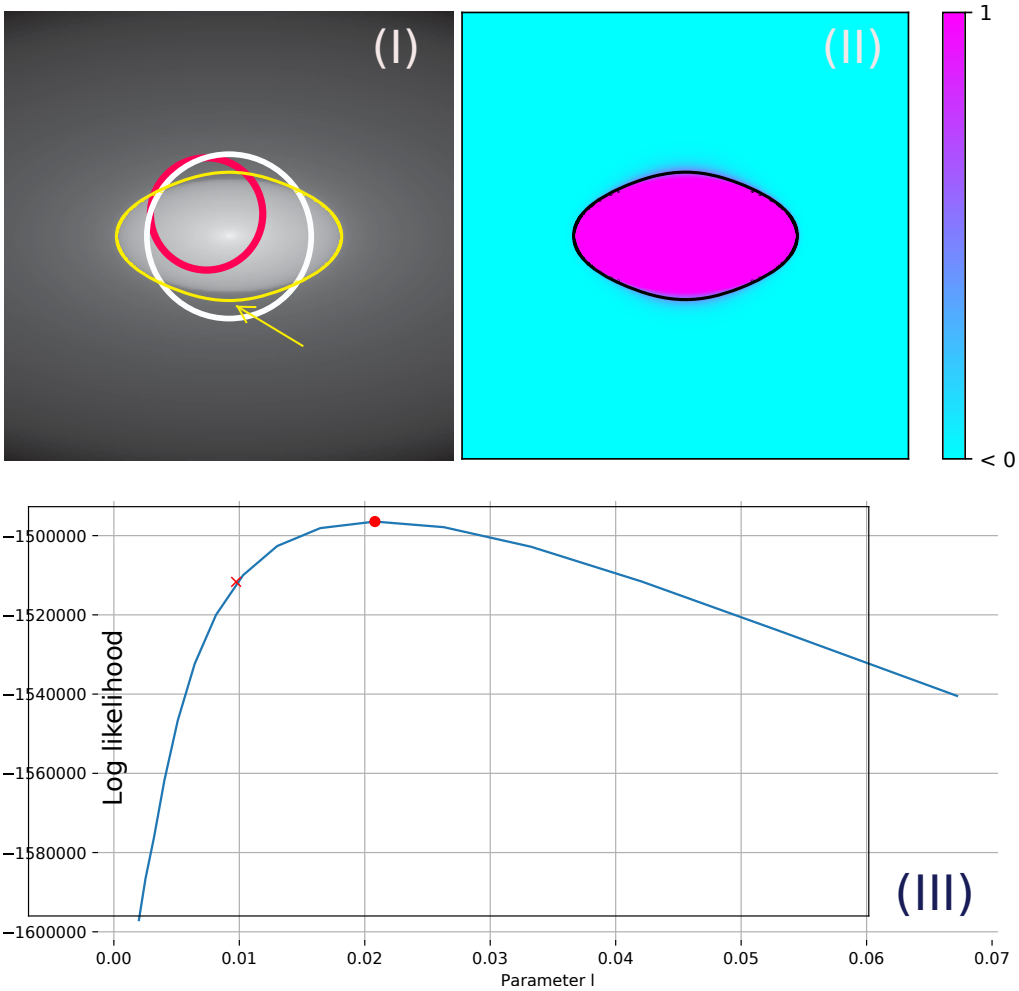

Figure 3.5: (I) Input Ellipse image fitted with a circle shape : initial circle (red), final circle (white) and 0.5 isocontour of posterior label probability for optimal value of l_{ref} (yellow);(II) posterior label probability $p(Z_n = 1|\boldsymbol{\theta}_S, \theta_I)$ for optimal value of l_{ref}; (III) Log likelihood as a function of l_{ref} ;

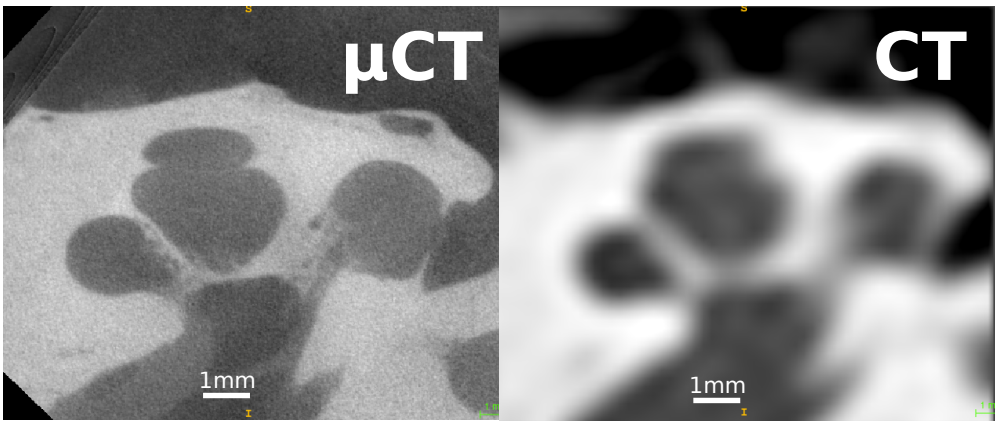

Figure 3.6: A visual comparison of imaging resolution between the μCT and conventional CT for cochlea imaging.

Logistic Shape Model Inference: In all cases, the deformable shape parameters $\boldsymbol{\theta}_{SD}$ were initialized as $(a = 4, b = 0.15, \alpha = 0.6, \varphi = 0.2)$ and the pose parameters were set to zero. The l_{ref} value was set between 0.1 and 0.3 (see section 3.4.3) and the stopping condition is $\frac{\Delta\theta}{\theta} < 0.1$, thus stopping when parameter updates are less than 10% of the parameter values.

Computational efficiency: We analyze the computational cost of several alternative formulations of our algorithm. More precisely, in Table 3.1 we compare the computational time of three different implementations of our approach that differ by the choice of the quasi-Newton optimization method in the MS-step (BFGS vs Gauss-Newton) and by the algorithm used for generating signed distance maps (VTK based vs deep learning based). The various algorithms was applied on the 9 images of dataset #2 and ran on a Dell Precision 7520 computer. It is clear that the Gauss-Newton method described in Algorithm 1 is far more efficient since it uses a much better approximation of the Hessian matrix than in the generic quasi Newton approach. Furthermore, as expected, the trained deep learning method leads to a speedup factor greater than 3.

Influence of the reference length To assess the influence of the hyper parameter reference length: l_{ref}, we analyse the variation of the final Dice score for various reference lengths based on one image of dataset #2. The results are shown in Table 3.3. We see that the reference length within the range

Table 3.1: Computational efficiency proposed methods

	BFGS	VTK SDM	DLSDM
Mean Comput. Time	12h15min	43min	16min

Table 3.2: Performance metrics obtained on dataset #2 and #3.

Compared Labels		**Dice Score**		**Symmetric Hausdorff Distance (voxel size 0.2 mm)**			
				Dataset #2		**Dataset #3**	
		Dataset #2	**Dataset #3**	**95%**	**100%**	**95%**	**100%**
CT	SSI	0.74 ± 0.02	0.77 ±0.023	0.53	1.04	0.70	1.91
	SROI	0.85 ± 0.011	0.91 ±0.015	0.34	0.82	0.36	1.68
μCT	SSI	0.67 ± 0.024	0.76 ±0.068	0.68	1.48	0.67	1.96
	SROI	0.81 ± 0.04	0.91± 0.019	0.50	1.31	0.36	1.68
	CT Manual	0.70 ± 0.084	0.93± 0.021	0.50	1.34	0.19	0.74

$[0.05mm, 0.25mm]$ has a relatively small influence on the Dice score. To minimize the time of computation, we do not optimize the reference length through a greedy search but simply set its value to 0.1 for dataset #1 and #2 and 0.3 for dataset #3 for shape fitting. To compute the final hard segmentation we use a fixed reference length of 0.25.

Robustness analysis To study the robustness of the method, we randomly initialize the cochlea shape parameters by performing a random uniform sampling within their defined value range. Based on 10 initial random samples, we computed the average Dice score for one image of dataset #2 and obtained a mean Dice

Table 3.3: Influence of the hyper parameter: l_{ref} for the segmentation accuracy.

Ref. Length	0.05	0.1	0.15	0.2	0.25
Dice Score	0.84	0.85	0.85	0.82	0.84

score of 0.81 ±0.1 (respectively of 0.68 ±0.23) for the Gauss-Newton method (resp. the BFGS method). This clearly shows the increased robustness with respect to initial shape values obtained by the Gauss-Newton optimization of the MS-step.

Evaluation on CT and μCT images: Datasets #2 and #3 include both CT and μCT images of the same subject that have been registered to each other. Furthermore the cochlea was manually or semi-automatically segmented by an expert on both modalities such that we can use those two binary maps to evaluate the accuracy of the algorithm applied on the CT image. The cochlea binary map from high resolution μCT images have been downsampled and represent a more reliable ground truth than the manual segmentation performed on the CT images.

The proposed algorithm using Gauss-Newton optimization and deep-learning generation of signed distance maps was applied on the $9 + 22$ CT images of the two datasets. Fig. 3.2 (Right) shows the segmented cochlea in red, the associated shape model, and its evolution during the MS step. Clearly, we see that the resulting segmentation is strongly constrained by the shape model.

In Table 3.2, we provide two metrics between pairs of binary masks : the Dice score and the 95% and 100% symmetric Hausdorff distance (HD) (computed as the average of two distances). Furthermore, we compare the segmentations produced by the posterior label probability (SROI for $p(Z_n|I_n, \boldsymbol{\theta}_S, \theta_I) = 0.5$) and the ones produced by the shape model only (SSI for $p(Z_n|\boldsymbol{\theta}_S) = 0.5$) with both manual segmentations obtained on CT and μCT images. To measure the uncertainty in the manual CT segmentation, we also evaluate the metrics between both CT and μCT manual mask images.

The logistic shape model framework produces good segmentation results on both datasets (Dice scores of 0.81 and 0.91) and even slightly outperforms the manual CT segmentation on dataset#2 (0.81 vs 0.7) which is far more challenging dataset #3. The segmented shape instances produced by the shape model are not as accurate as the SROI for the cochlea segmentation (lower Dice score and larger HD). This confirms that the parametric geometric cochlea model is a simplified representation of the cochlea anatomy. Finally, the metrics between the 2 manual segmentations on dataset #2 (DSC of 0.7 with a 95% HD of 0.5mm) shows the

difficulty of performing a manual segmentation of the cochlea due to its limited size and contrast.

3.4.4 Semi-quantitative analysis of segmentation on clinical dataset #1

We ran the segmentation framework (with DLSDM) on the 210 CT of dataset #1 on a Dell 6145 and 6420 CPU clusters.

Unsupervised quality control and semi-quantitative evaluation As manual segmentations of the 210 images are not available, we propose instead an original approach to estimate our algorithm's performance while minimizing the manual annotation effort. First, we apply the unsupervised quality control algorithm of Audelan *et al.* [Audelan and Delingette, 2019] on the whole dataset in order to sort the 210 segmentations according to their hypothesized performance. More precisely, this quality control algorithm computes for each image segmentation, an average distance between the segmentation provided by our algorithm and a segmentation produced by a simple generic probabilistic method. We can then generate an histogram of such average surface error (ASE) in Fig. 3.7. Segmentations having a low ASE correspond to those having good intensity contrast across their boundaries while those on the right tail of the distributions are considered as more challenging and suspect of including segmentation errors.

The histogram exhibits a bell shape with few outliers on its right and left tails. Furthermore, we have manually checked that this unsupervised quality control algorithms worked well on this dataset with visually better segmentations localized on the left side of the histogram. To estimate the relation between the ASE and the Dice score, we picked 5 images in order to sample the histogram at different levels of ASE corresponding to images #213, #210, #53, #264 and #143 (see Fig.3.7) in ascending order of ASE. Those 5 CT images were manually segmented by an ENT surgeon and the Dice scores of the segmentation produced by our algorithm was reported in Table 3.4. We see that the Dice score decreases as the ASE increases which indicates that the ASE may be a proper surrogate for the segmentation performance. The cochlea in image #143 was indeed found to be an outlier in terms of shape probably due to a patient malformation. Inspired by [Audelan and Delingette, 2019], we can make the hypothesis that ASE a good proxy for the Dice score as there is a monotonic relation between ASE and Dice.

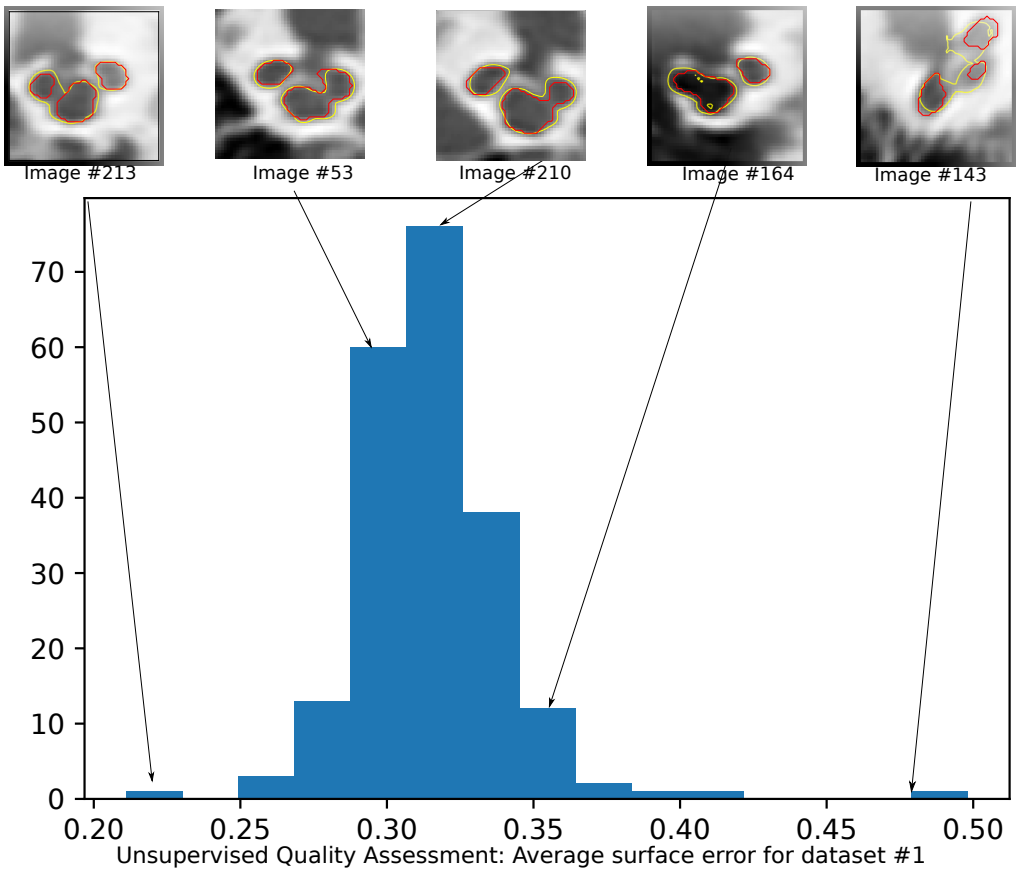

Figure 3.7: Average surface error of segmentations generated from dataset #1 resulting from the unsupervised quality control. Red contours correspond to the manual ground truth while yellow ones are segmentation outputs.

On this basis, we can extrapolate that the median Dice score over the whole dataset is probably above 0.82. Yet, a more thorough study with far more manual segmentations is necessary to be less speculative about the actual performance on clinical CT data.

Parameter analysis The application of the algorithm on dataset #1 resulted in the estimation of 10*210 shape parameters with 210 covariance matrices $\mathbf{\Sigma}^{\star}_{\boldsymbol{\theta}_S}$. In Fig. 3.8(a) the histograms of the 4 deformable shape parameters are displayed in green. Interestingly, the a and α parameters exhibit a bimodal distribution for which a simple explanation may be provided. Indeed, the left highest mode is

Table 3.4: Dice score for selected segmentation samples from dataset #1 based on the histogram of Fig.3.7. The ASE are got from automatic quality control algorithm and the DICE score are computed based on manual segmentation.

Patient ID	213	210	53	164	143
DICE Score	0.84	0.84	0.84	0.76	0.45
ASE	0.21	0.30	0.32	0.36	0.50

probably corresponding to straight centerline profiles whereas the rightmost mode may be associated with the "rollercoaster" longitudinal profiles [Avci et al., 2014]. In Fig. 3.8(b) the average 10×10 covariance computed as the log-Euclidean mean [Arsigny et al., 2007] is displayed showing potential correlations between pose and deformable parameters. Shape parameter b corresponding the exponent of the logarithmic center line curve has a particularly low variance such that it can be well estimated from the data. Conversely the phase parameter φ has much greater variance and is harder to estimate in average. The extraction of the eigenvectors of that covariance matrix confirms that most parameters are independent with other except for parameter a which is a bit correlated with the φ parameter and the translation term ty. The relative independence of the parameters for shape fitting indicates that we do not have an overparametrization of the cochla shape.

Uncertainty segmentation analysis The estimated covariance matrix $\boldsymbol{\Sigma}^{\star}_{\boldsymbol{\theta}_S}$ can be used for studying the uncertainty of the output segmentation. We sampled 100 times the multivariate Gaussian approximate posterior distribution of the parameters $p(\boldsymbol{\theta}_S|I) \approx \mathcal{N}(\boldsymbol{\theta}^{\star}_S; \boldsymbol{\Sigma}^{\star})$ and generated accordingly 100 random posterior labels $p(Z|I, \boldsymbol{\theta}_S, \theta_I)$ that are then averaged to estimate with Monte-Carlo sampling the marginal posterior $p(Z|I, \theta_I) = \int_{\mathbb{R}^{|\boldsymbol{\theta}_S|}} p(Z|I, \boldsymbol{\theta}_S, \theta_I) p(\boldsymbol{\theta}_S|I)\ d\boldsymbol{\theta}_S$. In Fig 3.9 we show several slices of the resulting probability maps with the 0.5 level curve together with the posterior probability $p(Z|I, \boldsymbol{\theta}^{\star}_S, \theta_I)$ obtained with the most likely shape parameter $\boldsymbol{\theta}^{\star}_S$. We see that we have a much larger uncertainty in the resulting segmentation when accounting for the uncertainty in the shape parameters than without them. This is a far better approximation of the true uncertainty $p(Z|I)$ than the posterior label probability $p(Z|I, \boldsymbol{\theta}_S, \theta_I)$.

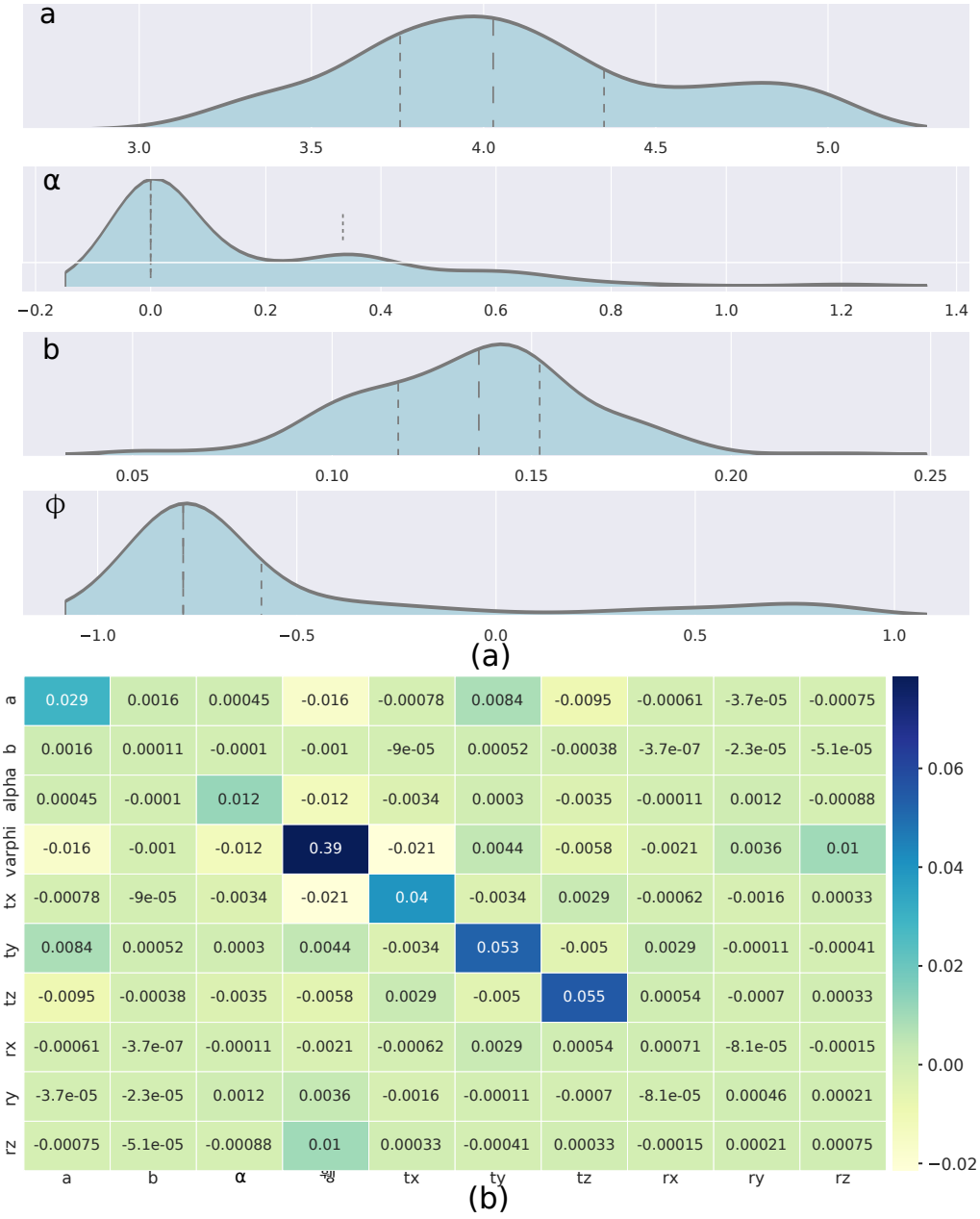

Figure 3.8: (a) Distribution plots for shape parameters variance. (b) Average covariance matrix of the 10 shape parameters.

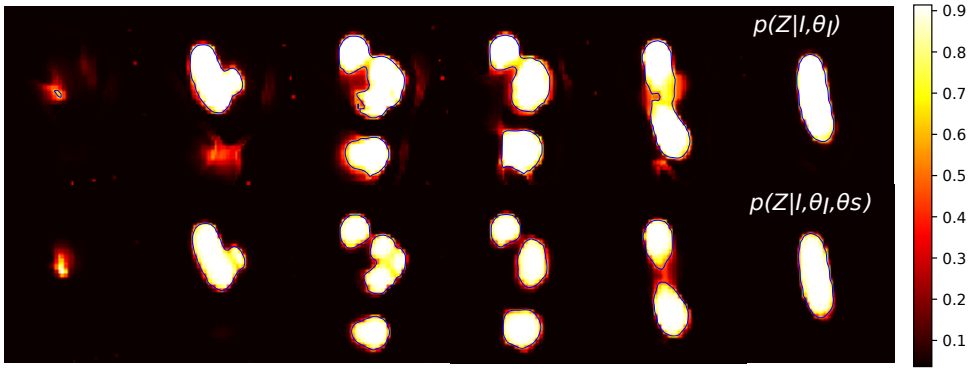

Figure 3.9: Marginal posterior probability $p(Z|I, \theta_I)$ (Top) versus posterior probability $p(Z|I, \theta_I, \boldsymbol{\theta}_S)$ (Bottom) computed on patient #1 of dataset #1.

3.4.5 Comparison with the state-of-the-art

We consider below the prior work on cochlea segmentation evaluated on clinical CT images while discarding the literature on the segmentation of μCT images [Kjer et al., 2014, Ruiz Pujadas et al., 2016a,b] or of the scala tympani and vestibuli located inside the cochlea [Noble et al., 2012, 2013]. Table 3.5 summarises the relevant publications on cochlea segmentation that are split into unsupervised and supervised methods. The former approaches are mostly based on cochlear shape fitting based on template image registration [Baker and Barnes, 2005], parametric shape model [Baker, 2008]. The supervised methods are based on statistical deformation models [Ruiz Pujadas et al., 2018] and deep learning [Heutink et al., 2020, Lv et al., 2021, Raabid et al., 2021].

Quantitative comparison of performances is not straightforward due to differences in image modality (CT, μCT or ultra high resolution CT), in metrics (Dice, precision, mean surface error), in subject population (cadaveric vs patient) but also in the target anatomical structures (cochlea vs cochlea labyrinth). In most cases, cochlea segmentation from μCT images are used as ground truth information and a direct comparison between our work with [Raabid et al., 2021] is possible since they used a subset of dataset #3 which is a public database [Wimmer et al., 2019]. We see that our unsupervised approach performs as well as the supervised methods with Dice scores in the range $[0.85, 0.91]$ and outperforms previous unsupervised methods.

Table 3.5: Performances of prior work on cochlea segmentation. *NL (resp.* NT) indicates the number of training (resp. testing) images. *Unsup (resp.* Sup) refers to unsupervised (resp. supervised) learning methods.

Method Group	Study	Comparison	Metrics		Proposed method (Dataset#2 N=9)	Proposed method (Dataset#3 N=22)
UnSup	Baker [2008] (NT= 4)	CT	Precision	0.72 ± 0.09	**0.75** ± **0.03**	**0.83** ±**0.03**
UnSup	Kjer and Paulsen [2015] (NT = 2)	post-mortem μCT	Mean (±1 std) surface error	0.22 ±0.17	**0.11** ±**0.06**	**0.063** ±**0.03**
Sup	Kjer et al. [2017] (NL = 18 /NT = 14)	post-mortem CT	Dice	0.88		
Sup	Heutink et al. [2020] (NL=48/NT = 75)	Ultra-high Resolution CT	Dice	0.90 ± 0.03		
Sup	Lv et al. [2021] (NL=24/NT=6)	Cochlea Labyrinth	Dice	0.90		
Sup	Raabid et al. [2021] (NL + NT = 17)	post-mortem CT	Dice	**0.90** ±**0.07**	0.85 ±0.011	**0.91** ±**0.03**

3.5 Discussion

The proposed approach relies on the definition of a generic shape function $\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x})$ which can be for instance a statistical shape model, a deformation image template, or an implicit shape equation. In the case of the cochlea, it was defined as a signed distance function of a parametric shape model $\text{SDM}(\mathcal{S}(\boldsymbol{\theta}_S), \mathbf{x})$. This specific choice makes the computation of the shape function and its gradient fairly costly, despite the use of a fully supervised dedicated neural network (DLSDM). There are several ways to optimize its computation time. One could for instance use a supervised appearance model such as a trained neural network which would remove all MI steps in the EM algorithm and would decrease by at least a factor 2 the time of computation. Another way is to use an implicit shape model $\mathcal{S}(\boldsymbol{\theta}_S, \mathbf{x}) = 0$ for instance based on statistical level sets [Tsai et al., 2003]. The cochlea segmentation example provided in this chapter relies on a fully interpretable intensity and shape parameters at the expense of its computational efficiency. Yet, one could train a deep neural regressor for predicting cochlea shape parameters and segmentation by using the segmentations generated by the proposed framework as training set.

For the cochlea segmentation, excellent results were obtained on cadaveric CT images similarly to the supervised methods. Furthermore, to the best of our knowledge, we introduced a first semi-quantitative assessment of cochlea segmentations on clinical CT images acquired on more than 200 patients. However, for a complete study, one would need to assess thoroughly the inter-rater variability of those manual segmentations and ideally combine them with other high resolution image modalities. Finally, an interesting extension of this work would be to segment the scala vestibuli and tympani in addition to the cochlea.

3.6 Accelerating parametric shape representing through Deep Learning

Signed distance map (SDM) is a common representation of surfaces in medical image analysis and machine learning. The computational complexity of SDM for 3D parametric shapes is often a bottleneck in many applications, thus limiting their interest. In this chapter, we propose a learning based SDM generation neural network which is demonstrated on a tridimensional cochlea shape model

parameterized by 4 shape parameters. The proposed SDM Neural Network generates a cochlea signed distance map depending on four input parameters and we show that the deep learning approach leads to a 60 fold improvement in the time of computation compared to more classical SDM generation methods. Therefore, the proposed approach achieves a good trade-off between accuracy and efficiency.

3.6.1 Signed Distance Map

A Signed Distance Map (SDM) Tsai and Osher [2003] is a scalar field $f(\mathbf{x})$ giving the signed distance of each point $\mathbf{x}$ to a given (closed) surface, which mathematically translates into the relation $\|\nabla f\| = 1$. In practise, SDMs are 2D or 3D images storing the distance of each voxel center and are widely used to tackle various problems in computer vision or computer graphics fields. In machine learning, SDMs are useful to encode the probability to belong to a shape through log-odds maps Pohl et al. [2006b]. For instance, given a surface $\mathcal{S}(\boldsymbol{\theta}_S)$ and a scalar l_{ref}, the probability for a voxel n having position $\mathbf{x}_n$ to belong to the surface can be provided through the SDM $\text{SDM}(\mathcal{S}(\boldsymbol{\theta}_S), \mathbf{x}_n)$ at that voxel as $p(Z_n = 1) = \sigma\left(\frac{\text{SDM}(\mathcal{S}(\boldsymbol{\theta}_S), \mathbf{x}_n)}{l_{\text{ref}}}\right)$ where $\sigma(x)$ is the sigmoid function.

While there exist fast (linear complexity) sweeping methods Maurer et al. [2003] for computing SDM from binary shapes, the naive computation of an SDM from triangular meshes has complexity $O(Nn_T)$ where N is the number of image voxels and n_T is the number of triangles describing the shape. An example of a generic computation of SDM from meshes is available in VTK Baerentzen and Aanaes [2005], Quammen et al. [2011] through the *vtkImplicitPolyDataDistance* class. Since many algorithms are relying on the SDM generation, it is critical to optimize its computation time in various ways Jia et al. [2018]. In medical image analysis, the naive approach leads to poor performances due to the fact that volumetric images and complex shapes are considered. To improve the performance of the SDM calculation, several authors Roosing et al. [2019], Wu et al. [2014] proposed 2D and 3D SDM computation methods that take advantage of graphics processing units (GPU) in order to accelerate the computation. Yet, there does not exist any generic library for fast computation of SDM on GPU, and the availability of specific GPU at test time is a significant limitation for machine learning applications.

Algorithmic optimizations were proposed by various authors Jones et al. [2006] by adopting hierarchical data structures to reach an $O(N \log n_T)$ complexity. For instance, Complete Distance Field Representation (CDFR) Jian Huang et al. [2001] were introduced with triangles structured into 3D grids cells.

Fast approximations of SDM was proposed in Wu and Kobbelt [2003] based on structured piece-wise linear distance approximation. Those approaches often require a significant pre-computation stage that can override their computational benefits at later stage.

Recent works of Chen and Zhang [2019] and Park et al. [2019] developed neural networks for the generation of SDM for various of shapes. They rely on an decoder network that takes as input shape parameters and position, and outputs the SDM at that point. The training of those deep SDFs is based on a continuous regression from random samples involving a clamp loss Park et al. [2019]. Those networks are used for shape inference and are point-based signed distance evaluators (without any convolution operation) rather than being generators of SDM. As discussed later in this chapter, this is a major issue for fast generation of large images of signed distance maps.

Despite those prior works, there does not exist any generic and efficient way to compute SDM from a triangular mesh on a grid on CPU resources. In this chapter, we propose an alternative method for fast computation of SDM based on Convolutional Neural Network (CNN) which does not rely on the rasterization of mesh triangles and does not require any hardware acceleration at test time. Results showed that our approach reduces the SDM computational time complexity significantly without any significant impact on the accuracy of shape recovery.

3.7 Methods and Evaluation

The cochlea is an organ that transforms sound signals into electrical nerve stimuli to the cortex. Cochlea lesions can lead to hearing loss that can be improved by inserting Cochlear Implant(CI) on patients at a middle stage of the disease. Cochlea shape recovery from images is a pivotal step for CI, and the work of Demarcy [2017a] is a state-of-art method for cochlea shape analysis which makes a computationally intensive use of SDM computations inside Expectation-

Maximization loops.

3.7.1 Cochlea Shape Model and Dataset

We rely on a parametric cochlea shape model that represents the shape variability of the human cochlea. It is represented as a generalized cylinder around a centerline having four shape parameters a, α, b, ϕ, two of them for the longitudinal (resp. radial) extent of the centerline. To compute the SDM of the shape model, the parametric surface was discretized as triangular meshes whose edge lengths are approximately 0.30±0.15 mm Demarcy [2017a]. The SDM was then generated by using VTK library and the *vtkImplicitPolyDataDistance* class which implements a naive SDM algorithm based on point-to-triangle distance computations.

For training the neural network, we generated a static dataset consisting of 625 ($5 \times 5 \times 5 \times 5$) cochlea SDM datasets of size $50 \times 50 \times 60$ by uniformly sampling the 4 deformation parameters within user specified ranges. In addition, we performed random data augmentation, by generating online SDMs during the training stage through a random sampling of the 4 shape parameters.

3.7.2 Signed Distance Map Neural Network

Our SDM Neural Network (SDMNN) is an encoder-decoder network with merged layers, its structure being inspired by the well known U-net Ronneberger et al. [2015a]. The SDMNN has the four shape parameters as input and generates as output a $50 \times 50 \times 60$ signed distance map (see Fig. 3.10).

3.8 Experiments and Evaluation

The SDMNN was trained on one NVIDIA 1080Ti GPU with both static 625 datasets and online random SDMs with a Mean Square Error (MSE) loss for 168 hours. After training, we generated 100 test SDMs with the naive mesh-based VTK code that are associated with random shape parameters. Those were compared to the SDMs generated by the SDMNN for the same shape parameters and the average MSE on the whole images were MSE $= 0.006mm$ which is small given that the range of a SDM is $(-0.2mm, 1.3mm)$.

Qualitative results are shown in Fig. 3.11 (I) where the comparison of the SDMNN and naive mesh-based generated maps is performed by extracting the

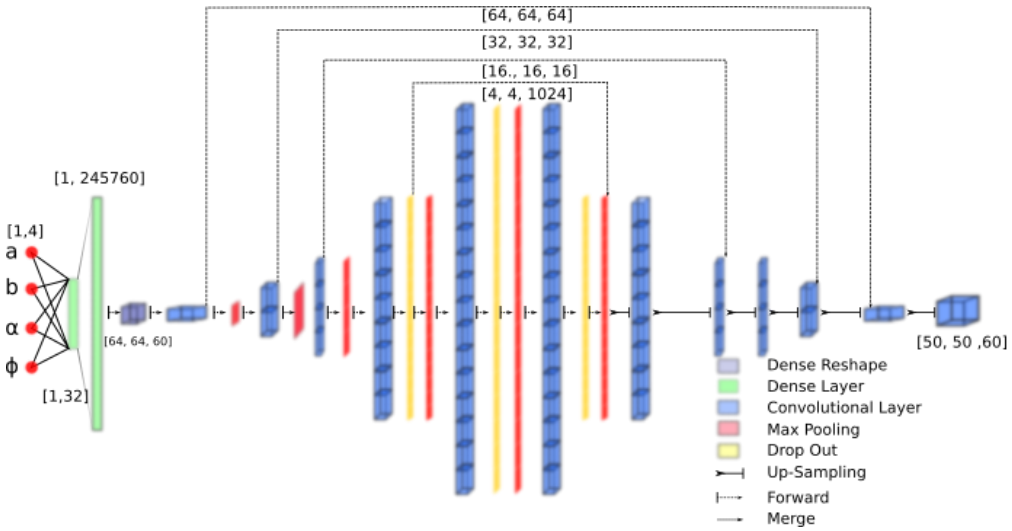

Figure 3.10: The structure of the proposed Signed Distance Map Neural Network (SDMNN).

isocontours associated with the zero (red) and 1mm (yellow) level sets. We see that the isocontours from the SDMNN match closely the ones generated from the mesh. Some small and smooth distorsions appear for the yellow contours. Since in surface reconstruction problems, the main focus of SDM is on the zero level set, the errors of the yellow isocontours are likely not to entail any major reconstruction errors. To verify the accuracy of the zero level isocontour, we have extracted the zero isosurface by the marching cubes algorithm associated with the standard shape values and compared that reconstructed surface with the original triangulated mesh model (the one used to generate the mesh SDM). In Fig. 3.11 (II) the 2 surfaces are overlaid showing that the SDMNN isosurface is as smooth as the original mesh and that the 2 surfaces are very close indeed. The proposed approach is evaluated quantitatively in three ways. First, we compare the computation times between VTK mesh-based SDM generation and the SDMNN-based generation. All evaluations were performed on a Dell Mobile Workstation with Intel(R) Core(TM) i7-7820HQ @ 2.90GHz CPU. We show in Table 3.6 that the SDM neural network is about 66 times more efficient to generate a SDM than the classical method. Second, the performance was also compared for fitting a cochlea shape model on a clinical CT image as in Demarcy

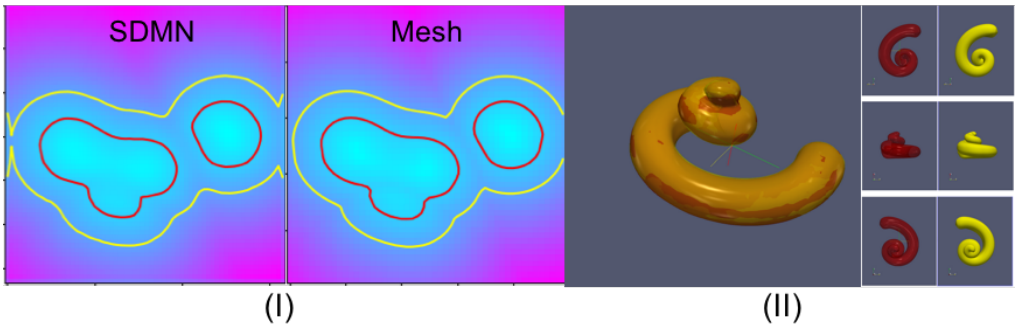

Figure 3.11: (Left-I) comparison between isocontours extracted from an SDM generated by the SDM neural network (left) and classical method (right);(Right-II) comparison of reconstructed 0-isosurface between the two methods.

[2017a] which requires several hundreds of evaluations of signed distance maps. In such case, the speedup was shown to be about 11 times faster than the mesh-based alternative. We also implemented the DeepSDF and IM-NET Chen and Zhang [2019], Park et al. [2019] for the generation of SDM of the cochlea with 4 shape parameters. For a fair comparison, we run DeepSDF (which is very similar to the IM-NET) to test its computational efficiency to fill a (60, 50, 50) SDM grid in one batch. The resulting computing time is 28s as shown in Table 3.6 which is even worse than the default VTK algorithm. This shows that there is high price to pay to have a point-based network rather than a image-based network. Furthermore, we found the accuracy in terms of signed distances of both networks to be significantly worse than our proposed SDMNN.

Table 3.6: Different Methods Computational time for SDM generation (h:m:s)

Generation Time	SDMNN	Mesh based SDM	DeepSDF
Single SDM	0:00:00.2	0:00:10.7	0:00:28.1
Shape Fit	1:05:02.1	12:15:45.4	Failed

Thirdly, we evaluated the difference in terms of estimated shape parameters after fitting 9 clinical CT cochlea volumes using both mesh-based and SDMNN methods. This lead to recover the 4 shape parameters a, α, b, ϕ on each of the 9 cochleas

that are stored in vector P_{mesh} when using mesh-based SDM generation method and vector P_{SDMNN} with SDMNN. The errors in shape parameters $P_{err} = \|P_{mesh} - P_{SDMNN}\|$ are reported in Table 3.8 showing negligible discrepancies given that the parameters magnitude (see head of Table 3.8).

3.9 Conclusion

In this chapter, we have presented a new probabilistic generative approach for combining shape and intensity models for image segmentation. The resulting segmentation is an interpretable compromise between a fidelity to a parametric shape space (captured in the prior distribution) and an appearance model (captured in the likelihood distribution). The proposed method goes well beyond the concept of shape fitting since it also provides an approximation to the posterior distribution of shape parameters. The use of a logistic shape model allows to control the trade-off between appearance and shape with a single parameter: the reference length. When applied to the recovery of cochlea structures from CT images, we were able to provide accurate segmentations with meaningful shape parameter distributions. Furthermore, we have shown how the approximate shape parameter posterior distribution can be exploited to provide realistic uncertainty maps. An interesting application of the proposed approach is to perform model selection with Bayes factors, in order to estimate the optimal complexity of a parametric shape model for a given image segmentation task.

In addition, we have proposed a deep learning-based fast signed distance map generation method. We showed quantitatively and qualitatively that it can generate 3D SDM in less than 300 ms, while having an accuracy suitable for shape-recovery, with no noticeable changes in recovered shape parameters. This CNN based SDM generation model can be used for any parametric shape model for SDM generation and does not require any GPGPU resources after training, which is compatible with a clinical environment. While other point-based approaches such as DeepSDF and IM-NET have been also proposed recently, the time overhead to fill a regular grid appears to be fairly large. The current approach is probably suitable only when the number of shape parameters is small since the number of SDMs in the training set should grow quadratically with the number of shape parameters.

Future work will look at additional strategies to speed-up the training stage and improve the output accuracy. Future work will also explore the application of this framework to other shape representations than explicit parametric shape models in order to find a reasonable trade-off between computational efficiency and interpretability of shape parameters. For instance, in statistical deformation models [Rueckert et al., 2003b], the computation of shape function gradient $\nabla_{\boldsymbol{\theta}_S}\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x})$ is straightforward, but its shape parameters may not be meaningful besides the first modes.

Table 3.7: Appearance model parameters initialization value. The $\#1 f_g$ refers the foreground appearance model for dataset #1. The $\#3 b_g$ refers the background appearance model for dataset #3.

	μ_1	μ_2	μ_3	μ_4	ν_1	ν_2	ν_3	ν_4	σ_1	σ_2	σ_3	σ_4	π_1	π_2	π_3	π_4
#1 b_g	-916.0	40.0	1890.0	593.0	4.0	4.0	4.0	4.0	5.7	7.1	7.8	9.8	0.04	0.0027	0.82	0.135
#1 f_g	51.1	524.9			1.8	4.8			6.0	7.2			0.12	0.88		
#3 b_g	-917.6	40.0	1890.3	593.0	4.0	4.0	3.6	4.0	4.6	10.0	9.0	9.8	0.0	0.0	0.99	0.0073
#3 f_g	-400.0	524.9			5.6	4.5			6.9	8.9			0.6	0.40		

Table 3.8: Shape parameters estimation error for SDMNN compared to mesh based SDM

Parameters Name	a	α	b	φ
Parameters Range	(2.0, 5.0)	(0.0, 1.2)	(0.05, 0.25)	$(-\pi/4, \pi/4)$
Mean shape parameters errors P_{err} on 9 cases.	2.06e-08	2.53e-08	5.4e-08	1.00e-09

3.10 Appendix

3.10.1 Gradient of shape function

In this section, we detail the computation of the shape function gradient $\nabla_{\boldsymbol{\theta}_S}\tilde{\mathcal{S}}(\boldsymbol{\theta}_S, \mathbf{x})$ when rigid and deformable shape parameters are considered. More precisely, writing the parameters controlling the non-rigid deformation as $\boldsymbol{\theta}_{SD}$, the shape function writes as $\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD}, \mathbf{R}\mathbf{x}_n + \mathbf{t})$. The rotation matrix $\mathbf{R}$ is parameterized with rotation vector $\mathbf{r}$, whose norm is the rotation angle and whose direction is the rotation axis. The gradients with respect to the translation and rotation vectors are then given in closed form as :

$$\begin{aligned}\nabla_{\mathbf{t}}\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD}, \mathbf{R}\mathbf{x}_n + \mathbf{t}) &= \nabla_x\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD}, \mathbf{R}\mathbf{x}_n + \mathbf{t}) \\ \nabla_{\mathbf{r}}\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD}, \mathbf{R}\mathbf{x}_n + \mathbf{t}) &= \left(-\mathbf{R}S_{\mathbf{x}_n}\frac{\mathbf{r}\mathbf{r}^T + ((\mathbf{R})^T - \mathbf{I}_3)S_{\mathbf{r}})}{\|\mathbf{r}\|^2}\right)^T \\ &\quad \nabla_x\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD}, \mathbf{R}\mathbf{x}_n + \mathbf{t})\end{aligned}$$

where ∇_x is the spatial gradient, $S_{\mathbf{x}}$ is the $3x3$ anti-symmetric matrix associated with vector $\mathbf{x}$. For a deformable parameter $\boldsymbol{\theta}_{SD}$, if the shape function is not given in an analytical form as it is the case for parametric shapes, the shape function gradient can be computed with finite differences based on a parameter increment $\delta\boldsymbol{\theta}^i_{SD}$:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}^i_{SD}}\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD}, \mathbf{R}\mathbf{x}_n + \mathbf{t}) = \frac{1}{2\delta\boldsymbol{\theta}^i_{SD}}\Big(&\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD} + \delta\boldsymbol{\theta}^i_{SD}), \mathbf{R}\mathbf{x}_n + \mathbf{t}) \\ &-\tilde{\mathcal{S}}(\boldsymbol{\theta}_{SD} - \delta\boldsymbol{\theta}^i_{SD}, \mathbf{R}\mathbf{x}_n + \mathbf{t})\Big)\end{aligned}$$

3.10.2 Cochlea Shape Model

We are interested in the cochlea structure in CT images which is defined as a generalized cylinder, i.e. as cross-sections swept along a centerline.

Centerline The centerline is parameterized in a cylindrical coordinate system by its *radial* $r(\theta_c)$ and *longitudinal* $z(\theta_c)$ components. The range of polar angle θ_c is $[0, \theta_{\max}]$ where $\theta_{\max}$ is the maximum polar angle controlling the total number of cochlear turns.

The radial component is defined piecewise with a polynomial function and a

logarithmic function of the angular coordinate θ_c in the cylindrical coordinate system as:

$$r(\theta_c) = \begin{cases} p_2\theta_c^2 + p_1\theta_c + p_0 & \text{if } \theta_c < \theta_0 \\ ae^{-b\theta_c} & \text{if } \theta_c \geq \theta_0 \end{cases} \tag{3.7}$$

where $\theta_0 = 5\pi/6$ and $p_0 = 5$ mm. Furthermore to obtain a continously differentiable curve, we set :

$$\begin{aligned} p_2 &= \frac{C_1\theta_0 - C_2 + p_0}{\theta_0^2} & p_1 &= \frac{-C_1\theta_0 + 2C_2 + 2p_0}{\theta_0} \\ C_2 &= ae^{-b\theta_0} & C_1 &= -C_2 b\,. \end{aligned} \tag{3.8}$$

The longitudinal component of the centerline is the sum of an exponentially damped sinusoidal and a linear function:

$$z(\theta_c) = \begin{cases} \alpha e^{-\beta\theta_c}\cos(\theta_c + \phi) + q_1\theta_c & \text{if } \theta_c < \theta_1 \\ a_2\theta_c^2 + a_1\theta_c + a_0 & \text{if } \theta_c \geq \theta_1 \end{cases}, \tag{3.9}$$

where $\beta = 0.2$ rad^{-1}, $q_1 = 0.225$ mm.rad^{-1} and $\theta_1 = \theta_{\max} - \pi$. The polynomial function is used to flatten out the last half turn so that $\mathrm{d}z(\theta)/\mathrm{d}\theta|_{\theta=\theta_{\max}} = 0$ and similarly a_2, a_1, a_0 are set to obtain a continuously differentiable curve.

Cross-Sections The cross-sections are modeled by a closed *planar shape* on which a varying *affine transformation* is applied along the centerline. The scala tympani and the scala vestibuli are modeled with two half pseudo-cardioids while the cochlear cross-section corresponds to the minimal circumscribed ellipse of the union of the tympanic and vestibular cross-sections. The affine transform of cross-sections is parameterized by a *rotation*, and a *width* and *height scalings*. All cross-sectional parameters are fixed because their variability was found to be small compared to the variability of the centerline.

Shape parameter vector We have chosen a compact description of the cochlea shape to limit as much as possible the correlation between the shape parameters and therefore make them uniquely identifiable. Finally, only 10 free parameters are considered in $\boldsymbol{\theta}_S$:

- 6 translation and rotation parameters : $\mathbf{t} = (tx, ty, tz)$, $\mathbf{r} = (rx, ry, rz)$

- 2 radial component parameters of the centerline, a and b
- 2 longitudinal component parameters, α and ϕ

Note that there are no free cross-section parameters which implies that $\boldsymbol{\theta}_S$ can be used to define uniquely the cochlea.

The prior probabilities on the 10 shape parameters were modeled as being an uniform distribution (uninformative prior) such that all regularization terms $\log p(\boldsymbol{\theta}_S|\boldsymbol{\alpha})$ can be ignored.

3.10.3 Initialization of intensity parameters

The $4 * 6 = 24$ initial intensity parameters for the mixture of Student's t distributions in datasets #1 and #3 are presented in Table 3.7.

Chapter 4

Quasi-Symplectic Langevin Variational Inference for unsupervised learning

Variational autoencoder (VAE) is a very popular and well-investigated generative model in neural learning research. To leverage VAE in practical tasks dealing with a massive dataset of large dimensions, it is required to deal with the difficulty of building low variance evidence lower bounds (ELBO). Markov Chain Monte Carlo (MCMC) is an effective approach to tighten the ELBO for approximating the posterior distribution. In particular, Hamiltonian Variational Autoencoder (HVAE) is an effective MCMC inspired approach for constructing a low-variance ELBO that is amenable to the reparameterization trick. In this paper, we

propose a Langevin dynamics flow-based inference approach that incorporates the gradients information in the inference process through the Langevin dynamics which is a kind of MCMC based method similar to HVAE. Specifically, we employ a quasi-symplectic integrator to cope with the prohibit problem of the Hessian computing in naive Langevin flow. We show the theoretical and practical effectiveness of the proposed framework in comparison with other methods, as it reaches the best negative log-likelihood on benchmark dataset. This chapter is adapted from a preprint paper [Wang and Delingette, 2021b].

4.1 Introduction

Variational Autoencoder (VAE) is a popular generative neural model, which is applied in a vast number of practical cases to perform unsupervised analysis or to modeling a dataset. It has the advantages of offering a quantitative assessment of generated model quality and being less cumbersome to train compared to Generative Adversarial Networks (GANs). One of the key factors influencing the performance of VAE models is the quality of the marginal likelihood approximation in the corresponding evidence lower bound (ELBO).

A common method to make the amortized inference efficient is to constraint the posterior distribution of the latent variables to follow a given closed-form distribution, often multivariate Gaussian [Wolf et al., 2016]. However, this severely limits the flexibility of the encoder. To increase the flexibility of the posterior modeling, the Hamiltonian Variational Inference (HVI) is proposed to remove the requirement of an explicit formulation of the posterior distribution by forwarding sampling a Markov chain based on Hamiltonian dynamics [Salimans et al., 2015]. It can be seen as a type of normalizing flows (NFs) [Rezende and Mohamed, 2015] where repeated transformations of probability densities are replaced by time integration of space and momentum variables. To guarantee the convergence of HVI to the true posterior distribution, Wolf et al. proposed to add an acceptance step in HVI algorithm. Further more, Caterini et al. [2018] first combined VAE and HVI in Hamiltonian Variational Autoencoders (HVAE) which include a dynamic phase space where momentum component ρ and position component z are integrated. The using of Hamiltonian flow for the latent distribution inference can introduce the target information (gradient flow) into the inference steps for improving the variational inference efficiency.

In this work, we propose a novel inference framework named quasi-symplectic Langevin variational auto-encoder (Langevin-VAE) that leads to reversible Markov kernels and phase quasi-volume invariance. The major contributions of this paper are:

- The proposed MCMC method is a Langevin-flow-based asymptotic low variance unbiased lower bound estimator.
- Different from prior Langevin normalizing flow, this approach is a generalized Langevin flow-based inference framework, which avoids computing the Hessian.
- The method shows comparable quantitative performance with conventional VAE frameworks on benchmark and real-world datasets.

4.2 Preliminary

4.2.1 Variational Inference and Normalizing Flow

Variational Inference One core problem in the Variational Inference (VI) task is to find a suitable replacement distributions $q_\theta(z)$ of the posterior distribution $p(z|x)$ for optimizing the ELBO: $\text{argmax}_\theta \, \mathbb{E}_q[\log p(x,z) - \log q_\theta(z)]$. To tackle this problem, Ranganath et al. proposed black-box variational inference by estimating the noisy unbiased gradient of ELBO, which performs direct stochastic optimization of ELBO. Kingma and Welling proposed to use some multivariate Gaussian posterior distributions of latent variable z generated by a universal function ω, which makes reparameterization trick is possible. To better approximate potentially complex posterior distributions of latent variables, the use of simple parametric distributions like multivariate Gaussian is a limitation. Yet only a few of distributions are compatible with the reparameterization trick. Normalizing Flows (NFs) is an effective way to deal with this limitation, which constructs a mapping between the complex and simple distributions by gradients transform.

Normalizing Flows Rezende and Mohamed proposed the NFs as a way to deal with more general parametric posterior distributions that can still be efficiently optimized with amortized inference [Papamakarios et al., 2019]. Briefly, NFs are a class of methods that use a series of invertible transformations $\mathcal{T}_I ... \circ ... \mathcal{T}_0$ to map a simple distribution z_0 into a complex one z_i: $z_i = \mathcal{T}_I ... \circ ... \mathcal{T}_0(z_0)$, By applying

a cascade of transformations, the corresponding logarithm prior probability $p(z_i)$ of the transformed distribution becomes:

$$\log(p(z_i)) = \log(p(z_0)) - \Sigma_0^i \log \left| det \frac{\partial \mathcal{T}_i}{\partial z_{i-1}} \right| \tag{4.1}$$

where the non-zero Jacobian $|det \frac{\partial \mathcal{T}_i}{\partial z_{i-1}}|$ of each transformation ensures the global volume invariance of the probability density. The positivity of each Jacobian terms is guaranteed by the invertibility of each transformation $\mathcal{T}$ and consequently by the reversibility of normalizing flows.

Hamiltonian Flows The Hamiltonian dynamics in HVAE can also be seen as a type of NFs, for which Eq: (4.1) also holds. Briefly, HVAE employs an I steps Hamiltonian transformation process: $\mathcal{H}_I$ to build an unbiased estimation of posterior $q(z)$ by extending $\tilde{p}(x, z)$ as $\tilde{p}(x, \mathcal{H}_I(z_0, \rho_0))$ leading to: $\tilde{p}(x) := \frac{\hat{p}(x, \mathcal{H}_I(z_0, \rho_0))}{q(\mathcal{H}_I(z_0, \rho_0))}$, where: $\hat{p}(x, z_I, \rho_I) = \hat{p}(x, \mathcal{H}_I(z_0, \rho_0)) = \hat{p}(x, z_I)\mathcal{N}(\rho_I|0, I)$. In particular, the HVAE enables the use of the reparameterization trick during inference thus leading to efficient ELBO gradients computation. The Hamiltonian dynamics is such that the distribution of phase space (z, ρ) [1] remains constant along each trajectory according to Liouville's theorem (symplectic) [Fassò and Sansonetto, 2007]. When using the leapfrog integrator with step size t for discretizing the Hamiltonian dynamics, the Jacobian remains to 1 (ignoring numerical rounding errors) with $|det \frac{\partial \mathcal{H}_i}{\partial z_i}|_t^{-1} = 1$. This property simplifies the Jacobian calculations at each discretization step [Caterini et al., 2018]. In HVAE, the posterior approximation is constructed by applying I steps of the Hamiltonian flow: $q^I(\mathcal{H}_i(\theta_0, \rho_0)) = q^0(\mathcal{H}_i(\theta_0, \rho_0)) \prod_{i=1}^I |det \nabla \Phi^i(\mathcal{H}_i(\theta_0, \rho_0))|^{-1}$, where Φ^i represents the leapfrog discretization transform of Hamiltonian dynamics. When combined with the reparameterization trick, it allows to compute an unbiased estimator of the lower bound gradients $\nabla_\theta \mathbb{L}$.

4.2.2 Langevin Monte-Carlo and Langevin Flow

A Langevin dynamics describes a stochastic evolution of particles within the particle interaction potential $U(x)$ that can be treated as a log probability density, it has recently attracted a lot of attention in the machine learning community [Giro-

[1] https://en.wikipedia.org/wiki/Phase_space

lami and Calderhead, 2011, Mou et al., 2020, Stuart et al., 2004, Welling and Teh, 2011] for the stochastic sampling of posterior distributions $p_\Phi(z|x)$ in Bayesian inference. Langevin Monte-Carlo methods [Girolami and Calderhead, 2011] rely on the construction of Markov chains with stochastic paths parameterized by Φ based on the discretization of the following *Langevin–Smoluchowski* SDE [Girolami and Calderhead, 2011] related to the overdamped Langevin dynamics :

$$\delta\Phi(t) = \frac{1}{2}\nabla_\Phi \log(p(x,\Phi))\delta t + \delta\sigma(t) \tag{4.2}$$

The stochastic flow in Eq (4.2) can be further exploited to construct Langevin dynamics based normalizing flow and its derived methods for posterior inference [Kobyzev et al., 2020, Wolf et al., 2016]. The concept of Langevin normalizing flow was first briefly sketched by Rezende and Mohamed [2015] in their seminal work. To the best of our knowledge, little work has explored practical implementations of Langevin normalizing flows. In [Gu et al., 2019], the authors proposed a Langevin normalizing flow where invertible mappings are based on overdamped Langevin dynamics discretized with the Euler–Maruyama scheme. The explicit computation of the Jacobians of those mappings involves the Hessian matrix of $\log(p_\Phi(x))$ as follows :

$$\log\left|det\frac{\partial\mathcal{L}_i}{\partial z_{k-1}}\right|^{-1} \sim \nabla_z\nabla_z \log(p(x,z)) + \mathbb{O}(z) \tag{4.3}$$

Yet, the Hessian matrix appearing in Eq (4.3) is expensive to compute both in space and time and adds a significant overhead to the already massive computation of gradients. This makes the method of [Gu et al., 2019] fairly unsuitable for the inference of complex models. In a more generic view, in the Langevin flow, the forward transform is modelled by the Fokker-Plank equation and the backward transform is given by Kolmogorov's backward equation which is discussed in the work of Kobyzev et al. and is not detailed here.

4.2.3 Quasi-symplectic Langevin and Corresponding Flow

Trivial Jacobian by Quasi-symplectic Langevin Transform To avoid the computation of Hessian matrices in Langevin normalizing flows, we propose to revert to generalized Langevin dynamics process as proposed in [Sandev T.,

2019]. It involves second order dynamics with inertial and damping terms:

$$\begin{aligned}\delta\Phi(t) &= K\delta t \\ \delta K(t) &= -\frac{\partial ln(p(x,\Phi))}{\partial\Phi}\delta t - \nu K(t) + \delta\sigma(t)\end{aligned} \tag{4.4}$$

where $\Phi(t)$ and $K(t)$ are the stochastic position and velocity fields, and ν controls the amount of damping. We can see that the Langevin–Smoluchowski type SDE of Eq.:(4.2) is nothing but the special case of high friction motion [Sandev T., 2019] when Eq.: (4.4) has an over-damped frictional force (proof is in 4.5.1).

To get simple Jacobian expressions when constructing Langevin flow, we need to have a symplectic Langevin transformation kernel. To this end, we introduce a quasi-symplectic Langevin method for building the flow [Milstein et al., 2002]. The quasi-symplectic Langevin differs from the Euler–Maruyama integrator method which diverges for the discretization of generalized Langevin SDE. Instead, the quasi-symplectic Langevin method makes the computation of the Jacobian tractable during the diffusion process and keeps approximate symplectic properties for the damping and external potential terms.

More precisely, the quasi-symplectic Langevin integrator is based on the two state variables (K_i, Φ_i) that are evolving according to the mapping $\Psi_\sigma(K_i, \Phi_i) = (K_{i+1}, \Phi_{i+1})$ where σ is the kernel stochastic factor. It is known as the *second order strong quasi-symplectic* method (4.5) and is composed of the following steps for a time step t:

$$\begin{aligned}K_{II}(t,\phi) &= \phi e^{-\nu t} \\ K_{1,i} &= K_{II}(\frac{t}{2}, K_i); \quad \Phi_{1,i} = \Phi_i - \frac{t}{2}K_{1,i} \\ K_{2,i} &= K_{1,i} + t\frac{\partial \log(p(x,\Phi_{1,i}))}{\partial\Phi_{1,i}} + \sqrt{t}\sigma\xi_i; \quad \xi_i \sim N(0,I) \\ K_{i+1} &= K_{II}(\frac{t}{2}, K_{2,i}); \quad \Phi_{i+1} = \Phi_{1,i} + \frac{t}{2}K_{2,i}\end{aligned} \tag{4.5}$$

where initial conditions are $K_0 = \kappa_0; \Phi_0 = \phi_0$.

The above quasi-symplectic integrator satisfies the following two properties:

Property 1. *Quasi-symplectic method degenerates to a symplectic method when* $\nu = 0$.

Property 2. *Quasi-symplectic Langevin transform* $\Psi_0(K_i, \Phi_i)$ *(4.5) has constant Jacobian :*

$$|\partial\Psi_0(K_i, \Phi_i)| = \frac{\partial\Phi_{i+1}}{\partial\Phi_i}\frac{\partial K_{i+1}}{\partial K_i} - \frac{\partial\Phi_{i+1}}{\partial K_i}\frac{\partial K_{i+1}}{\partial\Phi_i} = exp(-\nu t) \tag{4.6}$$

The first property shows that the VAE constructed based on the Quasi-Symplectic Langevin (QSL) dynamics is conceptually equivalent to a HVAE in the absence of damping $\nu = 0$. The second property implies that the Langevin-VAE integrator leads to transformation kernels that are reversible and with trivial Jacobians. The proofs of those two properties can be found in appendix 4.5.2 and more discussion about the quasi-symplectic integrator can be found in Milstein [2003]. The advantage of the QSL flow compared to the regular Langevin flow is that it avoids computing the Hessian of the log probability, which is a major advantage given the complexity of the Hessian computation.

We give below the formal definition of the quasi-symplectic Langevin normalizing flow.

Definition 1. *An* I *steps discrete quasi-symplectic Langevin normalizing flow* $\mathcal{L}^I$ *is defined by a series of diffeomorphism, bijective and invertible mapping* $\Psi_0 : \sigma_{\mathcal{A}} \to \sigma_{\mathcal{B}}$ *between two measurable spaces* $(\mathcal{A}, \sigma_{\mathcal{A}}, \mu_\alpha)$ *and* $(\mathcal{B}, \sigma_{\mathcal{B}}, \mu_\beta)$*:*

$$\begin{aligned}\mathcal{L}^I\mu_\alpha(\mathcal{S}_{\mathcal{A}}) :&\Psi_i \circ \mu_\alpha(\mathcal{S}_{\mathcal{A}}) = \mu_\alpha(\Psi_{i-1}^{-1}(\mathcal{S}_{\mathcal{B}})),\\ &\forall\mathcal{S}_{\mathcal{A}} \in \sigma_{\mathcal{A}}, \mathcal{S}_{\mathcal{B}} \in \sigma_{\mathcal{B}}, i = \{1, ..., I\}.\end{aligned} \tag{4.7}$$

where $\sigma_{(\cdot)}$ *and* $\mu_{(\cdot)}$ *are the* σ*-algebra and probability measure for set* $(\cdot)$ *respectively,* Ψ_i *is the* i_{th} *quasi-symplectic Langevin transform given by Eqs:(4.5).*

Example for single step quasi-symplectic Langevin flow We illustrate below definition 1 of a quasi-symplectic Langevin normalizing flow in case of a single transform applied on a single random variable. We consider a probability measure $p(x)$ of random variable set $x \in X$. Then a single step Langevin flow transforms the original random variable x to a new random variable $y = \Psi_0(x), y \in Y$. According to definition 1, the new probability measure $q(y)$ of random variable y is given by:

$$q(y) = \mathcal{L}^0 p(x) : \Psi_0 \circ p(x) = p(\Psi_0^{-1}(y)) \tag{4.8}$$

By Eq.(4.1), we conclude for a Langevin flow that:

$$q(y) = p(x) \cdot |det \frac{\partial \Psi_0}{\partial x}|^{-1} \tag{4.9}$$

The defined quasi-symplectic Langevin flow is a generalization of the Langevin flow with a quasi-symplectic structure for the parameters phase space. The quasi-symplectic Langevin normalizing flow has a deterministic kernel Ψ_0 when the kernel stochastic factor $\sigma = 0$, and degenerates to a symplectic transition when $\nu = 0$.

4.2.4 Lower Bound Estimation With Langevin-VAE

In the quasi-symplectic Langevin-VAE, we use an augmented latent space consisting of position ϕ_I and velocity κ_I variables of dimension $\zeta : z = (\phi_I, \kappa_I)$. The objective of the autoencoder is to optimize its parameters as to maximize the evidence lower bound $\tilde{\mathbb{L}}$:

$$\log p(x) = \log \int_\Omega p(x, z) dz \geq \int_\Omega \log \tilde{p}(x) q(\tilde{z}|x) d\tilde{z} \equiv \tilde{\mathbb{L}} \tag{4.10}$$

where Ω is the measure space of the latent variables and as $\tilde{p}(x)$ is an unbiased estimator for $p(x)$. The lower bound is equal to the evidence when the posterior approximation is equal to the true posterior. Thus maximizing the lower bound is equivalent to minimize the gap between the true posterior $p(z|x)$ and its approximation $q(z|x)$ [Blei et al., 2017].

Algorithm 2 Quasi-symplectic Variational Inference

Inputs: Data X, Inference steps I, damping ν, time step t, prior $q^0_{\omega_E}(\phi_0)$
Output: Encoding and decoding parameters $\omega = (\omega_E, \omega_D)$

Initialize all parameters, variables Define: $K_{II}(t,p) = pe^{-\nu t}$ **while** *NOT* ω *converged* **do**
  Get minibatch: $X_N \overleftarrow{}^{N} X$ **while** *NOT* $j = N$ **do**
    $x_j \overleftarrow{}^{j} X_N$ `// Get` x_j `in minibatch.`
    $\phi_0 \sim q^0_{\omega_E}(\phi_0|x_j)$ `// Sampling latent variable from variational prior`
    $\kappa_0 \sim \mathcal{N}(0, E_\zeta)$ `// Sampling velocity from unit Gaussian.`
    for $i = 1;\ i < I;\ i++$ **do**
      `// Quasi-symplectic Langevin Transform`
      $\kappa_{1,i} \leftarrow K_{II}(\frac{t}{2}, \kappa_i);\ \phi_{1,i} \leftarrow \phi_i - t\frac{\partial \log(p(x,\phi_))}{2\partial\phi_i};$
      $\kappa_{i+1} \leftarrow K_{II}(\frac{t}{2}, \kappa_{1,i});\ \phi_{i+1} \leftarrow \phi_{1,i} + \frac{t}{2}\kappa_{1,i}$
    end
    $p^*_\omega \leftarrow \hat{p}_{\omega_D}(x, \phi_I) \cdot \mathcal{N}(\kappa_I|0, E_\zeta)\ \ q^*_\omega \leftarrow q^0_{\omega_E}(\phi_0) \cdot \mathcal{N}(\kappa_0|0, E_\zeta) exp(\nu t)\ \ \tilde{\mathbb{L}}^*_j \leftarrow \log(p^*_\omega) - \log(q^*_\omega);$ `// Quasi-symplectic Langevin ELBO`
    $j \leftarrow j + 1$
  end
  $\tilde{\mathbb{L}}^* \leftarrow \sum_{i=1}^N \tilde{\mathbb{L}}^*_i / N$ `// Minibatch average ELBO`
  $\text{argmax}_{\omega \in \mathbb{R}^n} \tilde{\mathbb{L}}^*$ `// Optimize average ELBO over parameters subset`
end

The posterior approximation $q(z)$ is computed through a series of Langevin transformations which is the Langevin flow: $q_{\omega_E}(z|x) = q^I(\mathcal{L}^I(\phi_0, \kappa_0)|x) = q^0_{\omega_E}(\phi_0, \kappa_0|x) \prod_{i=1}^I |det \nabla \Psi_0(\phi_i, \kappa_i))|^{-1} = q^0_{\omega_E}(\phi_0, \kappa_0|x) \exp(I\nu t)$, where $q^0_{\omega_E}(\phi_0, \kappa_0|x)$ is an initial approximation parameterized by ω_E which can also be seen as the prior on random variables (ϕ_0, κ_0).

We then give the lower bound for the quasi-symplectic Langevin-VAE as:

$$\tilde{\mathbb{L}} := \int_\Omega q_{\omega_E}(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\phi_0, k_0)) - log(q^0_{\omega_E}(\phi_0, k_0)) + I\nu t d\tilde{z} \tag{4.11}$$

4.2.5 Quasi-symplectic Langevin-VAE

The quasi-symplectic Langevin lower bound $\tilde{\mathbb{L}}$ lays the ground for the stochastic inference of a variational auto-encoder. Given a set of dataset $X : \{x^i \in X; i \in$

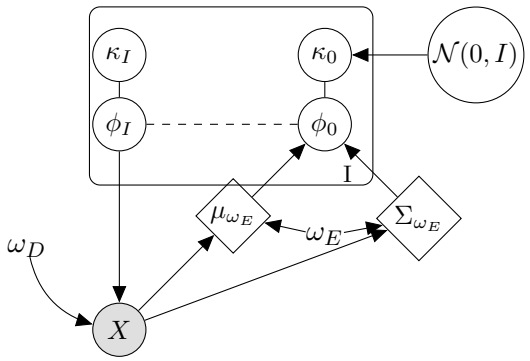

Figure 4.1: Graphical model of the Quasi-symplectic Langevin Variational Autoencoder. The multivariate Gaussian parameters $\mu_{\omega_E}, \Sigma_{\omega_E}$ defining the variational prior of latent variable ϕ_0 are determined from the data X and the parameter ω_E of the encoding network. The initial velocity latent variable κ_0 has a unit Gaussian prior and is paired by initial latent variable ϕ_0. After iterating I times the quasi-symplectic Langevin transform, the latent pair $\{\phi_I, \kappa_I\}$ is obtained from the initial variables $\{\phi_0, \kappa_0\}$. The decoder network with parameters ω_D is then used to predict the data from latent variables ϕ_I through the conditional likelihood $p(x|\phi_I)$. Variables in diamonds are deterministically computed. Network parameters ω_E, ω_D are optimized to maximize the ELBO.

$\mathbb{N}_+\}$, we aim to learn a generative model of that dataset from a latent space with the quasi-symplectic Langevin inference. The generative model $p(x, z)$ consists of a prior on initial variables $z_0 = (\phi_0, \kappa_0)$, $q^0_{\omega_E}(\phi_0, \kappa_0|x) = q^0_{\omega_E}(\phi_0|x) \cdot \mathcal{N}(\kappa_0|0, I_\zeta)$ and conditional likelihood $p_{\omega_D}(x|z)$ parameterized by ω_D. The Gaussian unit prior $\mathcal{N}(\kappa_0|0, I_\zeta)$ is the canonical velocity distribution from which the initial velocity of the Langevin diffusion will be performed. The distribution $q^0_{\omega_E}(\phi_0|x)$ is the variational prior that depends on the data x^i. Thus the generative model $p_{\omega_E,\omega_D}(x, z)$ is parameterized by both encoders and decoders and the quasi-symplectic Langevin lower bound writes as:

$$\begin{aligned}\underset{\omega_E,\omega_D}{\operatorname{argmax}}\, \tilde{\mathbb{L}}^* =& \mathbb{E}_{\phi_0 \sim q^0_{\omega_E}(\cdot), \kappa_0 \sim \mathcal{N}_\zeta(\cdot)}(\log \hat{p}_{\omega_E,\omega_D}(x, \mathcal{L}^I(\phi_0, \kappa_0)) - \\ & \log(q^0_{\omega_E}(\phi_0, \kappa_0)) + K\nu t)\end{aligned} \tag{4.12}$$

The maximization of the lower bound (4.12) can be performed efficiently with the reparameterization trick depending on the choice of the variational prior $q^0_{\omega_E}(\phi_0)$. To have a fair comparison with prior work [Caterini et al., 2018], we

also perform Rao-Blackwellization for reducing the variance of the ELBO in the quasi-symplectic Langevin-VAE:

$$\begin{aligned}\underset{\omega_E,\omega_D}{\operatorname{argmax}}\,\tilde{\mathbb{L}}^* =& \mathbb{E}_{\phi_0\sim q^0_{\omega_E}(\cdot),\kappa_0\sim\mathcal{N}_\zeta(\cdot)}(\log\hat{p}_{\omega_E,\omega_D}(x,\mathcal{L}^I(\phi_0,\kappa_0))\\ &-\log(\hat{q}_{\omega_E}(\phi_0,\kappa_0))+K\nu t-\frac{1}{2}\kappa_I^T\kappa_I)+\frac{\zeta}{2};\quad \forall\phi_0,\kappa_0\in\mathbb{R}^\zeta\end{aligned} \tag{4.13}$$

The resulting algorithm is described in Alg.29.

4.3 Experiment and Result

We examine the performance of quasi-symplectic Langevin-VAE on the MNIST dataset [LeCun et al., 2010] based on various metrics. Caterini et al. [2018] have reported that the Hamiltonian based stochastic variational inference outperforms that of Planar Normalizing Flow, Mean-field based Variational Bayes in terms of model parameters inference error. In this paper, we compare the proposed Langevin-VAE with: VAE, HVAE, Importance Weighted Autoencoder (IWAE) [Huang et al., 2019], Deep Belief Networks (DBNs) [Hinton, 2010], and Deep Autoregressive Networks (DANs) [Gregor et al., 2014] on MNIST dataset.

Given a training dataset $X:\{x^i\in X; i\in\mathbb{N}_+\}$ consisting of binary images of size d, $x^i\in\{0,1\}^d$, we define the conditional likelihood $p(x|z)$ as a product of d Bernoulli distributions. More precisely, we consider a decoder neural network $\text{Dec}_{\omega_D}(\phi)\in[0,1]^d$ that outputs d Bernoulli parameters from the latent variable $\phi\in\mathbb{R}^\zeta$ where $z=(\phi,\kappa)$. Then the conditional likelihood writes as : $p(x^i|z^i)=\prod_{j=1}^d \text{Dec}_{\omega_D}(\phi)[j]^{x^i[j]}\ (1-\text{Dec}_{\omega_D}(\phi)[j])^{1-x^i[j]}$.

4.3.1 Quasi-symplectic Langevin-VAE on binary image benchmark

Implementation details Following the classical VAE approach [Kingma and Welling, 2014], the encoder network parameterized by ω_E outputs multivariate Gaussian parameters : $\mu_{\omega_E}(x)\in\mathbb{R}^\zeta$ and $\Sigma_{\omega_E}(x)\in\mathbb{R}^\zeta$, such that the variational prior is a multivariate Gaussian $q^0_{\omega_E}(\phi_0|x)=\mathcal{N}(\phi_0|\mu_{\omega_E}(x),\Sigma_{\omega_E}(x))$ with diagonal covariance matrix. This choice obviously makes the reparameterization trick feasible to estimate the lower bound. The related graphical model of the quasi-symplectic Langevin-VAE is displayed in Fig. 4.1. The decoder and encoder

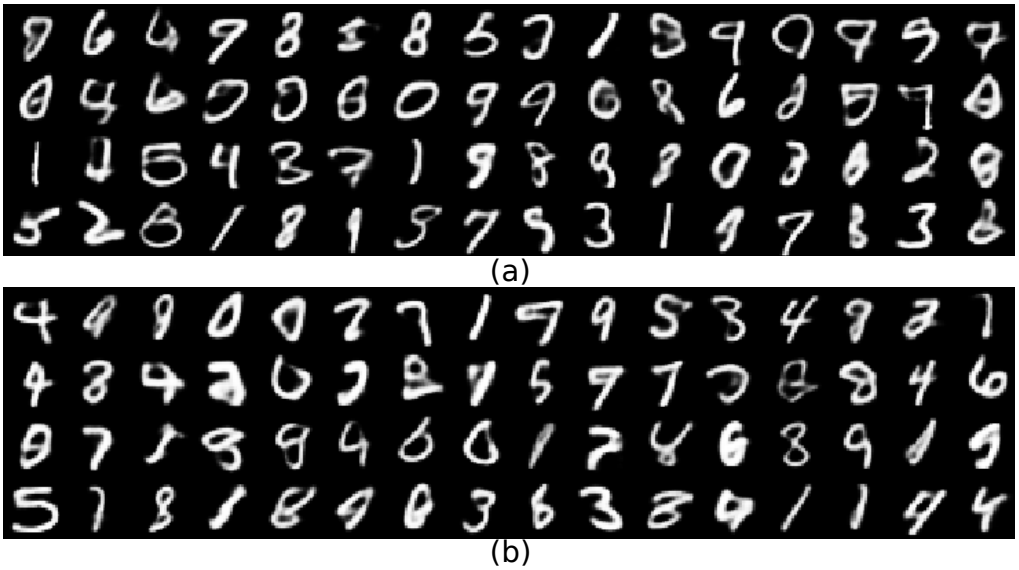

Figure 4.2: Generated samples from Langevin-VAE (b) in comparison with HVAE (a). Upper sub-figures are generated samples of HVAE. The lower sub-figures (b) are samples of Langevin-VAE. In both methods, the number of steps in the flow computation is $K = 5$.

Table 4.1: Quantitative evaluation of the Langevin-VAE in comparison with the HVAE, IWAE, DBN, and DAN methods on MNIST benchmark. It includes the comparison of the negative log likelihoods (NLL), the evidence lower bound (ELBO), and Inception Score (IS) [Borji, 2019]

	Langevin-VAE		**HVAE**		**IWAE**	**DAN**	**DBN**
Flow steps	2	5	2	5	-	-	-
NLL	82.95	*82.40*	83.10	82.75	82.90	84.13	84.55
ELBO	-85.37	*-84.81*	-85.70	-85.29	-	-	-
IS	7.67	*7.76*	7.59	7.38	-	-	-

neural network architectures are similar to the HVAE [Caterini et al., 2018] and MCMCVAE [Salimans et al., 2015], both having three layers of 2D convolutional neural networks for encoder and decoder, respectively. The encoder network accepts a batch of data of size $(N_b \times 28 \times 28)$ with $N_b = 1000$. The dimension

of latent variables is set as $\zeta = 64$ and the damping factor is $\nu = 0$. The discretization step is randomly chosen between ta and tb: $t \in [t_a, t_b]$. The training stage stops when the computed ELBO does not further improve on a validation dataset after 200 steps or when the inference loop achieves 2000 epochs. The scale term σ of Langevin dynamics was set to: $2\sqrt{T}$, where T is the annealing temperature. The initial tempreature is set to: $T_0 = 1.5$

Both tested models Langevin-VAE and HVAE share the same training and testing parameters except for specific Langevin parameters (detailed in 4.5.3). The stochastic ascent of the ELBO is based on the Adamax optimizer with a learning rate $lr = 5e - 5$. All estimation of computation times were performed on an NVIDIA GeForce GTX 1080 Ti GPU. The experiments were implemented with *TensorFlow 2.0* and *TensorFlow Probability* framework to evaluate the different methods in both qualitative and quantitative metrics.

Result on MNIST Both qualitative and quantitative results are studied. The generated samples of Langevin-VAE and HVAE are shown in Fig: (4.2). We qualitatively see that the quality and diversity of the sampled images are guaranteed for both autoencoder models. Quantitatively, Table 4.1 shows the performance in terms of the NLL, ELBO, IS scores for different Langevin-VAE and HVAE. In addition, we compare the negative log-likelihood of the flow-based frameworks(HVAE, LVAE) with 3 non-flow based generative models: IWAE, DBNs, and DANs. The IWAE, which achieved a comparable result (82.90 nats) against (82.40 nats) for the Langevin-VAE. Yet, this was the best performance of IWAE [Huang et al., 2019], which achieved through a k-sample ($k = 5000$) importance weighting MCMC. We notice that the Langivin-VAE reaches the best performance for the 3 evaluation metrics in comparison with the other methods on the MNIST benchmark.

One of the drawbacks of flow-based methods is the time and space overhead of the gradient calculation. The HVAE requires $k + 1$ or $2 \times k$ times computation of the gradient depending whether the gradients are reused or not. If the gradients are reused as in most implementations, they must be stored and retrieved, which requires a compensation between the memory and time cost. Instead, the Langevin-VAE relies on the computation of k gradient vectors without any requirement to store and retrieve them, which is slightly more efficient.

4.3.2 Quasi-symplectic Langevin-VAE on Medical Image dataset

We employ the proposed method for inference of a 3D cochlea CT image dataset. The human cochlea is an auditory nerve organ with a spiral shape. Some severe hearing impairments can be treated with cochlear implants. The shape of the cochlea is of great significance to the definition of preoperative and postoperative plans. The quantitative analysis of the shape of the cochlea is the focus of several research papers [Caversaccio et al., 2019, Manley, 2017, Wang et al., 2020c]. In this experiment, we use the proposed method to create a compact representation of the cochlea shape and appearance.

Dataset and Implementation details The dataset includes 1080 patients 3D images that collected from the radiology department of Nice University Hospital. The original slices sequences are with a spacing size of $0.185mm, 0.185mm, 0.25mm$. We used a reference image to register all the images to the cochlea region (FOI) by using an automatic pyramidal blocking-matching (APBM) framework [Ourselin et al., 2000b, Toussaint et al., 2007]. The FOI volumes are resampled into isometric spacing size of $0.2mm$ with volume of $(60, 50, 50)$. The proposed 3D Langevin-VAE consists with two 3D CNN. The encoder takes tensors with shape of $(N_b = 10 \cdot N_c = 60, N_w = 50, N_h = 50)$ and processes the tensors by three 3D convolutional layers with Softplus non-linear projections. The strides of all the convolutional layers are set to 2. The tensors are then flattened to fully connected layer for estimating the mean and variance as the posterior parameters. Samples of the posterior are then fed to a decoder network which follows inverse operations of the encoder to upscale the feature maps to the original tensor shape. The decoder network uses the deconvolutional operation to compute the marginal likelihood $p(x|z)$.

Result on real dataset Tab. 4.2 shows two inference metrics that represent the inference performance on the medical image dataset. We see that the Langevin-VAE outperforms the classical of naive VAE method on the dataset abstraction ability (as the ELBO and NLL are all better than VAE). The right sub-figure of Fig. 4.3 shows the slices generated by a Langevin-VAE with 2D latent variables. The generated images along the marginal distributions shows changes in shape (position) and appearance (Hounsfield Units), and on the right sub-figure of Fig. 4.3, we can observe the variance of generated samples in both

Table 4.2: Quantitative evaluation of the Langevin-VAE in comparison with the VAE

	VAE	**Langevin-VAE**
Avg. ELBO	-85293.33 +/- 1.538	**-85135.24 +/- 4.82**
Avg. NLL	83204.7+/- 10.92	**83159.44 +/- 5.31**

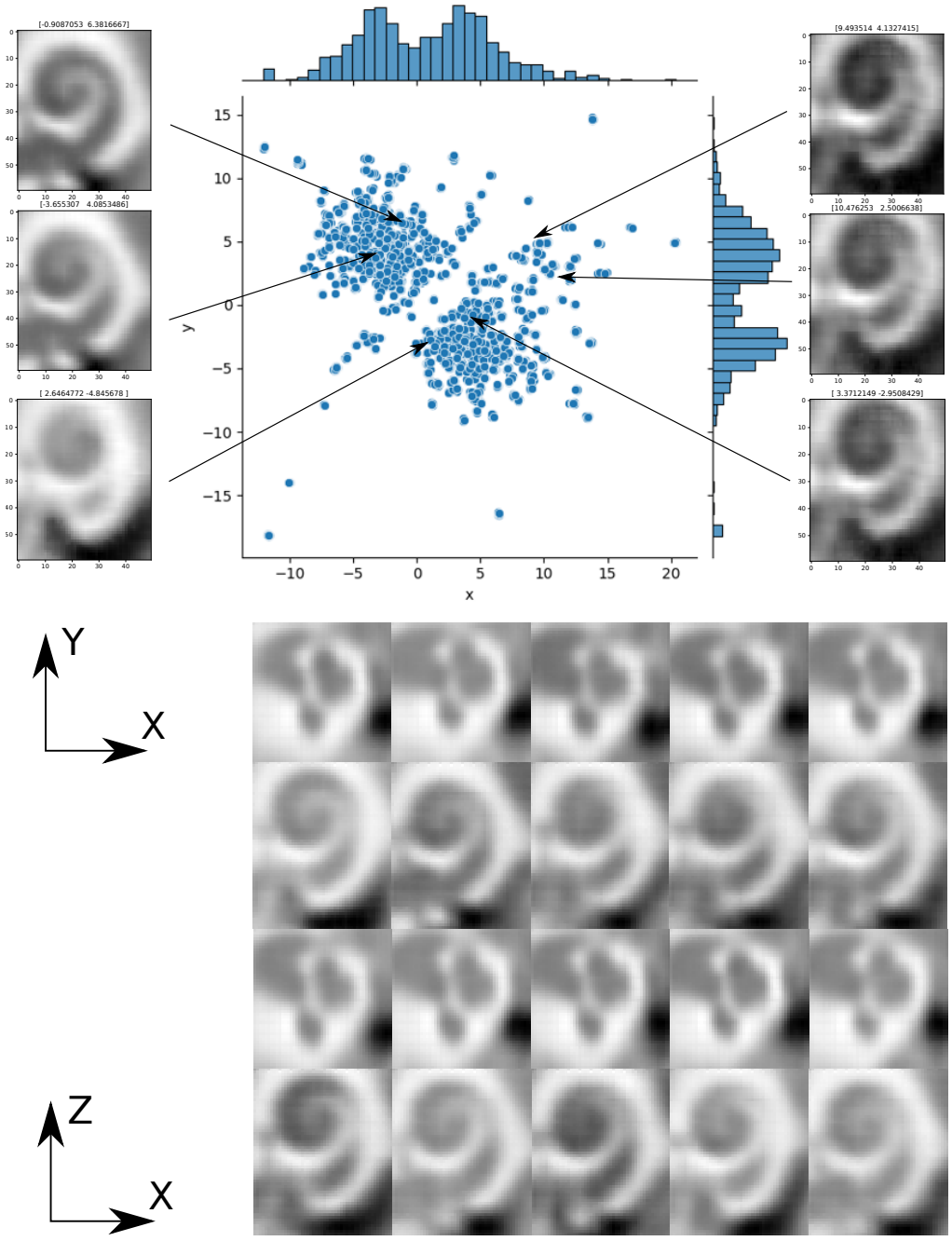

Figure 4.3: Qualitative assessment of the generated samples of Langevin-VAE on inner ear CT images dataset. The upper sub-figure shows a Langevin-VAE with latent parameters in 2D: $\zeta = 2$. The lower sub-figures are middle slices (from different views) of 12 samples which generated by the Langevin-VAE.

shape and appearance. This implies that the Langevin-VAE learns the variance of the cochlea shapes and the diversity of the intensity changes.

4.4 Conclusion

In this paper, we propose a new flow-based Bayesian inference framework by introducing the quasi-symplectic Langevin flow for the stochastic estimation of a tight ELBO. In comparison with conventional VAE and other methods, the proposed method achieves better performance on both toy and real world problems.

More specifically, by introducing the quasi-symplectic Langevin dynamics, we also overcome the limitation of the Langevin normalizing flow [Gu et al., 2019] which requires to provide the Hessian matrix $\nabla\nabla log(p(x,\phi))$ to compute the Jacobian. To the best of our knowledge, the proposed approach is the first Langevin flow based method as a generative model for dataset modeling.

Potential improvements of the quasi-symplectic Langevin inference can arise by investigating the manifold structure of the posterior densities of the latent variables [Barp et al., 2017, Girolami and Calderhead, 2011, Livingstone and Girolami, 2014] to improve the inference efficiency.

4.5 Appendix

4.5.1 Over-damped form of the Generalized Langevin Diffusion

We consider a unit mass $m = 1$ evolving with a Brownian motion. The velocity part of the generalized Langevin type equation is:

$$\partial\Theta(t) = Kdt \quad \partial K(t) = \frac{\partial\Theta(t)^2}{\partial t^2} = \frac{\partial ln(p_\Theta(x))}{\partial\Theta}dt - \nu\Gamma K(t) + \delta\sigma(t) \tag{4.14}$$

In the case of an over-damped frictional force, the frictional force νK overwhelms the inertial force $m \cdot \partial^2\theta/\partial t^2$, and thus $\frac{\frac{\partial\Theta(t)^2}{\partial t^2}}{\nu K(t)} \approx 0$. According to the generalized Langevin diffusion equation, we have :

$$\frac{\frac{\partial\Theta(t)^2}{\partial t^2}}{\nu K(t)} = \frac{\frac{\partial ln(p_\Theta(x))}{\partial\Theta}dt}{\nu K(t)} - \Gamma + \frac{\delta\sigma(t)}{\nu K(t)}$$

Therefore, we get :

$$\nu K(t)\Gamma \approx \frac{\partial ln(p_\Theta(x))}{\partial \Theta}dt + \delta\sigma(t)$$

which is the evolution given in Eq 4.4.

4.5.2 Proof the integrator Eq. 4.5 is quasi-symplectic

Proposition 1. *Eq. 4.5 is asymptotic symplectic:* $\lim_{\nu\to 0} |\partial\Psi_0(K_i, \Phi_i)| = exp(-\nu t)$

Remark 1. *Proposition [1] has propositional equivalences that the exterior power between two integration steps are equivalent as the Jacobian | partial*$\Psi_0(K_i, \Phi_i)|$ *is not dependent on the time step term t. Thus, to prove the proposition 1 is equivalent to proof that:* $dK_{i+1} \wedge d\Phi_{i+1} = dK_i \wedge d\Phi_i$.

Proof:

Let, $\nu \to 0$, the term K_{II} of the composite integrator Eq. 4.5 goes to: $\lim_{\nu\to 0} K_{II}(t, \phi) = \phi$

Then,

$$\begin{aligned} dK_{i+1} &= dK_i + td(\frac{\partial \log(p(x, \Phi_i + \frac{t}{2}K_i))}{\partial \Phi_i}) \\ &= dK_i + d[\frac{t\partial \log(p(x, \Phi_i + \frac{t}{2}K_i))}{\partial \Phi_i}](d\Phi_i + \frac{t}{2}K_i) \\ d\Phi_{i+1} &= d\Phi_i + \frac{t}{2}dK_i + \frac{t}{2}d(K_i + \frac{t\partial \log(p(x, \Phi_i + \frac{t}{2}K_i)}{\partial \Phi_i}) \\ &= d\Phi_i + tdK_i + d\frac{t^2\partial \log(p(x, \Phi_i + \frac{t}{2}K_i)}{2\partial \Phi_i}(d\Phi_i + \frac{t}{2}K_i) \end{aligned} \tag{4.15}$$

Let $U' = \frac{\partial \log(p(x, \Phi_i + \frac{t}{2}K_i)}{2\partial \Phi_i}$, thus,

$$\begin{aligned} dK_{i+1} \wedge d\Phi_{i+1} = dK_i \wedge d\Phi_i + dK_i \wedge tdK_i + dK_i \wedge \frac{t^2}{2}dU'd\Phi_i + \\ dK_i \wedge \frac{t^3}{4}dU'dK_i + tdU'(d\Phi_i + \frac{t}{2}dK_i) \wedge d\Phi_i + tdU'(d\Phi_i + \\ \frac{t}{2}dK_i) \wedge tdK_i + tdU'(d\Phi_i + \frac{t}{2}dK_i) \wedge \frac{t^2}{2}dU'(d\Phi_i + \frac{t}{2}dK_i) \end{aligned} \tag{4.16}$$

Table 4.3: Models hyper parameters and training/testing parameters setting.)

	ν	t_a, t_b	T_0	l_r	$adamax_\epsilon$
Langevin-VAE	1e-2	[1e-2, 5e-1]	1.5	5e-4	1e-7
Hamiltonian VAE	-	[1e-2, 5e-1]	1.5	5e-4	1e-7

According the property of exterior product, therefore:

$$dK_i \wedge tdK_i = tdU' d\Phi_i \wedge d\Phi_i = tdU' \frac{t}{2} dK_i \wedge tdK_i = 0 \tag{4.17}$$

Simplifying Eq. 4.16:

$$\begin{aligned} dK_{i+1} \wedge d\Phi_{i+1} &= dK_i \wedge d\Phi_i + t^2 dU'(dK_i \wedge d\Phi_i) + t^2 dU'(d\Phi_i \wedge dK_i) + \\ &\frac{t^4 dU'^2}{4}(d\Phi_i \wedge dK_i) + \frac{t^4 dU'^2}{4}(dK_i \wedge d\Phi_i) \\ &= dK_i \wedge d\Phi_i + t^2 dU'(dK_i \wedge d\Phi_i) - t^2 dU'(dK_i \wedge d\Phi_i) + \\ &\frac{t^4 dU'^2}{4}(d\Phi_i \wedge dK_i) - \frac{t^4 dU'^2}{4}(d\Phi_i \wedge dK_i) \\ &= dK_i \wedge d\Phi_i \end{aligned} \tag{4.18}$$

Q.E.D.

4.5.3 Parameters of the Experiment Setting

Tab. 4.3 shows the parameters used for the experiment. Except for the parameter ν that is unique for the Langevin-VAE, all the other parameters are the same as the Hamiltonian-VAE.

4.5.4 Evidence lower bound of Langevin Flow

We consider the log-likelihood: $\log p(x)$ with latent variables z, based on Jensen's inequality:

$$\log p(x) \geq \int_\Omega \log \tilde{p}(x) q(\tilde{z}|x) d\tilde{z} \tag{4.19}$$

The data prior is given through the Langevin flow where $\mathcal{L}^I(\theta_0, k_0)$ are the I steps Langevin flows with initialization states (θ_0, k_0):

$$\tilde{p} = \frac{\hat{p}(x, \mathcal{L}^I(\theta_0, k_0))}{q^0(\mathcal{L}^0(\theta_0, k_0))} \tag{4.20}$$

Therefore, we can get the Langevin flow lower bound:

$$\begin{aligned}
\tilde{\mathbb{L}} :\geq & \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log q^0(\mathcal{L}^0(\theta_0, k_0)))d\tilde{z} \\
= & \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log(q^0(\theta_0, k_0) \prod_{k=1}^{I} |det \nabla \Psi_i^{-1}(\theta_0, k_0))|))d\tilde{z} \\
= & \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log(q^0(\theta_0, k_0)) - \sum_{k=1}^{I} \log(|det \nabla \Psi_I^{-1}(\theta_0, k_0))|))d\tilde{z} \\
= & \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log(q^0(\theta_0, k_0)) + \sum_{k=1}^{I} (\nu t))d\tilde{z}
\end{aligned} \tag{4.21}$$

Chapter 5

One-shot Learning based Landmarks Detection

Landmark detection in medical images is important for many clinical applications. Learning-based landmark detection is successful at solving some problems but it usually requires a large number of annotated dataset for the training stage. In this paper, we tackle the issue of automatic landmark annotation in 3D volumetric images from a single example based on a one-shot learning method. It involves the iterative training of a shallow convolutional neural network combined with a 3D registration algorithm in order to perform automatic organ localization and landmark matching. We investigated both qualitatively and quantitatively the performance of the proposed approach on clinical temporal bone CT volumes. The results show that our one-shot learning scheme converges well and leads to a good accuracy of the landmark positions. This chapter is based on a preprint paper [Wang et al., 2020e] which is under peer review.

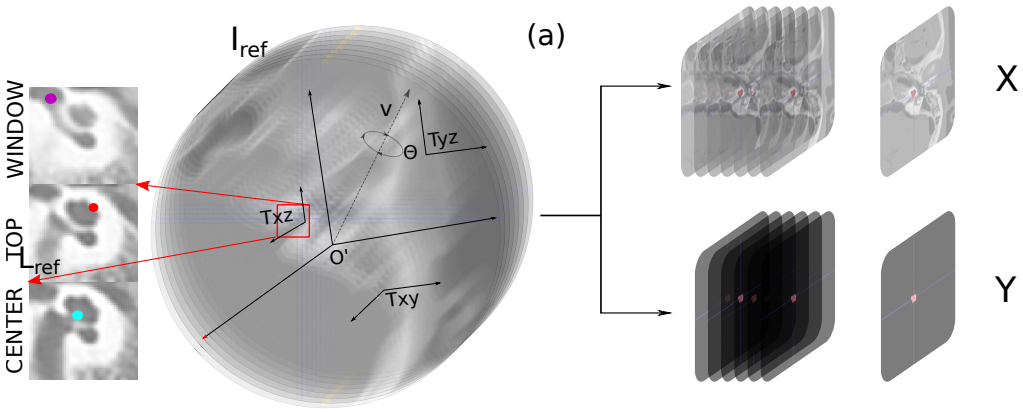

Figure 5.1: Data augmentation for training the CNN.

5.1 Introduction

Landmarks detection for target object localization plays a pivotal role in many imaging tasks. Automatic landmark detection can alleviate the challenges of image annotation by human experts and can also save time for many image processing tasks. The difficulty of landmark detection in clinical images may come from anatomical variability, or changes in body position which can lead to large differences of shape or appearance. The literature on automatic landmarks detection approaches can be roughly split into learning based versus non-learning based algorithms.

Non-Learning based landmarks detection In [Cheung and Hamarneh, 2009] is proposed the augmentation of the scale-invariant feature transform (SIFT) to arbitrary n dimensions (n-SIFT) for 3D-MRI volumes. However, the computation cost for 3D SIFT features is heavy as their complexity is a cubic function of the image size. Wörz *et al.*[Wörz and Rohr, 2006] leverage parametric intensity models for image landmarks detection. Ricardo *et al.* Ferrari et al. [2011] use log-Gabor filters to extract frequency features for 3D Phase Congruency (PC) applied to detect head and neck landmarks.

Learning based landmarks detection Probabilistic graphical models were used for bones landmark labelling in Schmidt et al. [2007] and [Corso et al., 2008].

Potesil et al. [2011] use joint spatial priors and parts based graphical models to improve the landmarks detection accuracy of organs. Shouhei *et al.* Hanaoka et al. [2017] proposed a Bayesian inference of landmarks through a parametric stochastic landmark detector of the candidates. Donner et al. [2013] applied random forest and Markov Random Field (MRF) for vertebral body landmarks detection.

Mothes and Denzler [2017] proposed a one-shot SVM based landmarks tracking method for X-Ray image landmark detection. Suzani *et al.* Suzani A. [2015] proposed to train a convolutional neural network (CNN) with an annotated dataset for automatic vertebrae detection and localization. Liang *et al.* Liang et al. proposed a two-step based residual neural network for landmarks detection. Deep reinforcement learning for landmarks detection was investigated by Ghesu *et al.* Ghesu et al. [2019] where landmarks localization is considered as a navigation problem.

The main drawback of the above deep learning based landmarks detection methods is that the creation of manually annotated dataset with 3D landmarks is time consuming and in practice very difficult to collect. To tackle this problem, Zhang *et al.* Zhang et al. [2017] proposed a deep learning based landmarks detection method that can be used in limited number of annotated medical images. Their framework consists of two CNNs: one for regressing the patches and the second to predict the landmark positions. Yet, this method like the rest of the learning-based methods are not suited when only one annotated image is available. Another source of difficulties is to detect landmarks that are concentrated on a small part of the image. A typical example is the detection of cochlear landmarks from CT images since the human cochlea is a tiny structure.

In this chapter, we tackle the problem of automatic determination of 3D landmarks in a volumetric image from a single example consisting of a reference image with its landmarks. We propose a one-shot learning approach that first localizes a Structure Of Interest (SOI) (e.g. the cochlea in a CT image of the inner ear) which lies next to the landmarks. A 2D CNN is trained offline by generating arbitrary oriented slices of a reference image with the binary mask of the SOI. Given a target image, the location of the SOI is iteratively estimated by applying the 2D CNN on 3 orthogonal sets of slices. After aligning the orientations of the two SOI on the target and reference images, a non-rigid registration algorithm is

applied to propagate the landmarks to the target image. We apply this approach on 200 CT images of the temporal bone to locate 3 cochlear landmarks and show that the positioning error is within the intra-rater variability. To the best of our knowledge, this is the first one-shot learning method for landmark detection which makes it highly applicable for several clinical problems.

5.2 Method

5.2.1 Overview

The proposed algorithm requires as input a reference image I_{ref} where a set of landmarks L_{ref} are positioned. In addition, we require that a binary mask of a visible anatomical or pathological structure $S_{\text{ref}} \subset I_{\text{ref}}$ including the landmarks $L_{\text{ref}} \in S_{\text{ref}}$ be provided. Given a target image I_{target}, landmarks L_{target} are estimated by applying an image registration algorithm between an image patch $P_{\text{ref}} \subset I_{\text{ref}}$ centered on the reference landmarks and an image patch $P_{\text{target}} \subset I_{\text{target}}$ extracted on the target image. The main challenge is to automatically extract the target image patch P_{target} such that it is roughly aligned in position and orientation with the reference image patch in order to ease the non-rigid image registration task.

To extract the centered target image patch, we first train a 2D CNN (shown in Fig. 5.2) to segment the mask S_{ref} on random slices of the reference image. This stage is performed offline and also requires an additional validation image I_{val} where the same visible structure S_{val} has been segmented. Given a target image, the localization stage extracts the target image patch P_{target} by iteratively applying the segmentation network to find the center of mass of the structure and by aligning its axis of inertia. The last stage applies a registration algorithm to estimate the position of landmarks L_{target}.

5.2.2 Offline one-shot CNN training

The objective is to train an algorithm that can roughly segment the structure of interest $S_{\text{ref}} \subset I_{\text{ref}}$. That structure must include the landmarks or must lie in the vicinity of the landmarks L_{ref}. It should also be present in all target images and must be easy to detect in the image with some visible borders. One issue of one-shot learning is the limited amount of training data that can easily lead

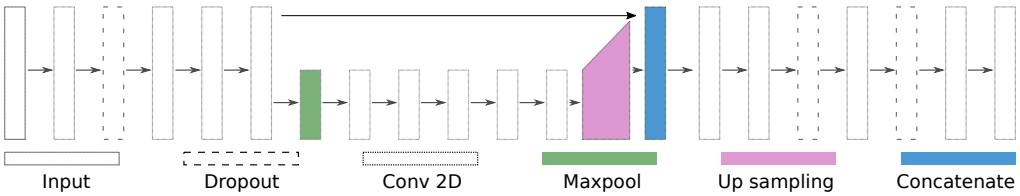

Figure 5.2: The neural network structure.

to overfitting [Wu et al., 2012]. To this end, we chose to train a shallow 2D U-net f_ω segmentation network in order to segment the SOI that surrounds the landmarks. The training set consists of slices of the reference image I_{ref} along arbitrary rotations and translation offsets together with the associated binary masks created by slicing accordingly the reference segmentation S_{ref}. The 2D CNN is trained by minimizing the Binary Cross-Entropy (BCE) loss function. To limit the risk of overfitting, we use a validation set consisting of another volumetric image I_{val} and its segmentation S_{val}. The training is stopped when the segmentation performance of f_ω on the 3 orthogonal slices of I_{val} starts to decrease. The details of the training procedure are provided in algorithm 3. The CNN training can be performed offline and the 2D random image slices are

Algorithm 3 One-shot training of CNN

Inputs: image: $I_{\text{ref}}, I_{\text{val}}$, segmentation: $S_{\text{ref}}, S_{\text{ref}}$
Output: CNN parameters ω

Initialize: f_ω, $\Delta T, \Delta R$ **while** L_{val} *decreases* **do**
  $T \leftarrow (U(-1,1)\Delta_T)^3$; // Uniform Random Translation
  $R \leftarrow (U(-1,1)\Delta_R)^3$ // Uniform Random Rotation
  $I_{trans} \leftarrow \text{Resample}(I_{\text{ref}}, R, T)$ // Transformed Image
  $S_{trans} \leftarrow \text{Resample}(S_{\text{ref}}, R, T)$ // Transformed Segmentation
  for $i = 1;\ i < K;\ i++$ **do**
    $f_\omega \overset{\omega}{\leftarrow} I_{trans}[i]|S_{trans}[i]$ // Train the CNN
  end
  $L_{val} \leftarrow loss(S_{\text{val}}, f_{cnn}(I_{\text{val}}))$ // Validation loss
end

centered on the center of mass $\mathbf{C}_{\text{ref}}$ (for $T = 0$) of the segmented structure of interest S_{ref}. Furthermore, the 2D image size of the CNN input is chosen as to cope with the translation ΔT and rotation ΔR offsets such that random slices do not include any missing pixel values.

5.2.3 Online Structure Detection

Given an input image I_{target}, we seek to locate the structure of interest S_{target} with the proper translation and orientation offsets in order to ease the last image registration stage.

Translation offset estimation To determine the 3D translation offset between I_{target} and I_{ref}, we propose to align the centers of the mass corresponding to the structures of interest S_{target} and S_{ref}. We rely on the trained CNN $f_\omega()$ to determine S_{target} given I_{target}. However, with the limited training set of $f_\omega()$, we need to cope with its possible poor performance. To this end, we propose an iterative method described in algorithm 4 and Fig.5.3, where the estimation of the translation offset is progressively refined. We write as $f_\omega(I^x_{\text{target}}[k])[i,j]$ the output of the CNN applied on the slice k in the X direction of the volumetric image I_{target} which is a 2D probability map. We apply the CNN on the slices of I_{target} extracted along the X,Y,Z directions. To improve the robustness of the center of mass estimation of I_{target}, we combine their output by multiplying the 3 probabilities outputs for each voxel. The joint output of the network at voxel $[i,j,k]$ is then written as :

$$p[i,j,k] = f_\omega(I^Z_{\text{target}}[k])[i,j] \cdot f_\omega(I^Y_{\text{target}}[j])[k,i] \cdot f_\omega(I^X_{\text{target}}[i])[j,k] \tag{5.1}$$

The product of the 3 probability maps favors the pixels where the 3 outputs agree. This helps to filter out the false positive pixels produced by the network that are not correlated on the 3 slice orientations. The center of mass $\mathbf{C}_{\text{target}}$ is then simply estimated as the barycenter of the image voxels weighted by the joint probability $p[i,j,k]$:

$$\mathbf{C}_{\text{target}} = \frac{1}{\sum_{i,j,k} p[i,j,k]} \left(\sum_{i,j,k} x[i,j,k]*p[i,j,k], \sum_{i,j,k} y[i,j,k]*p[i,j,k], \sum_{i,j,k} z[i,j,k]*p[i,j,k] \right)^T \tag{5.2}$$

The target image is then cropped around the detected center $\mathbf{C}_{\text{target}}$ which is written as $\tilde{P}_{\text{target}}$. When the translation offset between the target and reference images is large, the CNN segmentation performances tend to degrade since it has been trained with slices roughly centered on the center of S_{ref}. This is why we propose to iteratively apply the same approach on I_{target} after being centered on $\mathbf{C}_{\text{target}}$. This way, we expect the centered image to be more and more accurately

segmented by the neural network since it sees slices that resemble more and more to its training set. We stop the process when the changes in the detected center $\mathbf{C}_{\text{target}}$ become smaller than a threshold.

Algorithm 4 Iterative center of mass localization

Inputs: image: I_{target}, CNN: $f_\omega(\cdot)$
Output: Center of structure in target image $\mathbf{C}_{\text{target}}$

Initialize: ϵ $\mathbf{C}_{\text{target}} \leftarrow \mathbf{C}_{\text{ref}}$ **while** $|\mathbf{C}_{\text{old}} - \mathbf{C}_{\text{target}}| < \epsilon$ **do**
  $\tilde{P}_{\text{target}} \leftarrow \text{Crop}(I_{\text{target}}, \mathbf{C}_{\text{target}})$ // Patch centered on $\mathbf{C}_{\text{target}}$
  while $o \in \{X, Y, Z\}$ **do**
    for $i = 1; i < K[o]; i++$ **do**
      $out[o][i] \leftarrow f_{omega}(\tilde{P}^o_{\text{target}}[i])$ // apply CNN on slices
    end
  end
  $p \leftarrow out[X] \cdot out[Y] \cdot out[Z]$ // Combine probability maps as Eq.5.1
  $\mathbf{C}_{\text{old}} \leftarrow \mathbf{C}_{\text{target}}$ $\mathbf{C}_{\text{target}} \leftarrow$ Eq. 5.2 // Update center of mass
end
$\tilde{P}_{\text{target}} \leftarrow \text{Crop}(I_{\text{target}}, \mathbf{C}_{\text{target}})$ // Patch centered on $\mathbf{C}_{\text{target}}$

Figure 5.3: Iterative determination of the center of mass of the structure of interest. Steps (1) - (2) show the 2D CNN segmentation of the structure of interest from the 3 set of orthogonal slices; (3) The probability maps of the 3 views are combined; (4) Update of the center of mass from the joint probability maps; (5) The target image is cropped around the center of mass.

Rotation offset estimation After having aligned the center of mass of the two structures of interest, the rotation offset is determined by aligning the moments of inertia of S_{ref} and S_{target}. More precisely, the matrix of inertia captures the ellipsoid appearance of each structure and it determines the structure orientation unambiguously if that structure does not have any axis of symmetry. Therefore the alignment of the matrices of inertia consists in applying a rotation to S_{target} such that the eigenvectors of the 2 matrices coincide Crisco and McGovern [1997], Jaklic and Solina [2003] when they are sorted according to their eigenvalues. The moments of inertia of S_{target} are computed based on the combined probability $p[i, j, k]$ as computed in Eq.5.1. Thus, after performing the eigenvalue decompo-

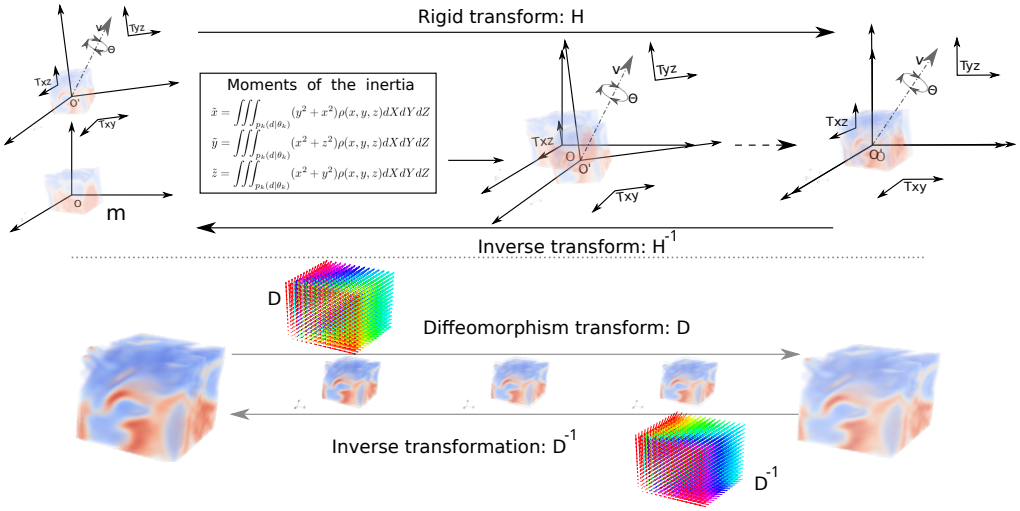

Figure 5.4: Landmarks matching based on inverse rigid and diffeomorphism transformation. Above subfigure shows the rigid transform H: (1) Compute the moments of the inertia of the two volumes. (2) Optimize the alignment position through optimizing the similarity measure metric. Below subfigure shows the diffeomorphism transform D: (1) Compute the diffeomorphism transformation D. (2) Compute the matched landmarks by inverse the displacement field D^{-1}.

sition of the 2 matrices, the rotation matrix centered on $\mathbf{C}_{\text{target}}$ is applied on the image patch $\tilde{P}_{\text{target}}$ to get the final target image patch P_{target}.

5.2.4 Online Image Patch Registration

After the two previous stages, the estimation of the landmarks L_{target} is achieved by performing a non-rigid registration of the reference image patch P_{ref} onto the

target image patch P_{target}. The two image patches have the same size, are both centered on the structure of interest and their orientation roughly coincide. This is a good initialization for applying the standard diffeomorphic demons algorithm Vercauteren et al. [2007a] as implemented in "itk::DiffeomorphicDemonsRegistrationFilter". This algorithm starts with a multi-resolution rigid registration followed by the non-rigid transformation parameterized by a stationary velocity field. It assumes that intensity distribution matches between the two images patches with a sum of square difference as similarity measure. The reference landmarks L_{ref} are then transported to the target image patch P_{target} through the estimated non-rigid transformation. Finally, the landmarks L_{target} on the original target image I_{target} are positioned by inverting the rigid transforms and cropping performed during the first two stages of the method. Fig. 5.4 shows an overview of the entire procedures.

5.3 Experiment and Result

5.3.1 Dataset

The dataset consists of 200 volumetric CT images of the left temporal bones acquired by a GE LightSpeed CT scanner at the Nice University Center Hospital. The image dimensions are $(512, 512, 160)$ in 3D with corresponding spacing of $(0.25mm, 0.25mm, 0.24mm)$. In this case, the structure of interest is the cochlea, a relatively small bone having a spiral shape similar to a snail shell and without any axis of symmetry. The cochlea is easily visible on CT images. Two volumetric images were randomly selected to serve as reference and validation images and their cochlea was then segmented by an expert in a semi-automatic fashion. Three landmarks corresponding the cochlea top, center and round window were manually set on the reference image as shown in Fig. 5.6.

5.3.2 Network architecture and training details

We use a 2D U-net like network Ronneberger et al. [2015b] for segmenting the cochlea in 2D images. The network structure is shown in Fig:5.2 and is relatively shallow in order to minimize its complexity. The network input size is $[nb, 100, 100, 1]$ followed with 4 convolutional layers (shape: $[nb, 100, 100, 64]$) where nb is the number of batches. Feature maps are convoluted with a group of

halved size layers but doubled in channels (shape: $[\cdot, 50, 50, 128]$). Up-sampling layer applied to recover the size of the feature maps to merged with the jump concatenates feature maps (shape:$[\cdot, 100, 100, 64 + 128]$). Finally, 5 convolutional layers (shape:$[\cdot, 100, 100, 64]$, $chn = 64$ for middle layers, $chn = 1$ for the last layer) are used for generating the final feature map. An Adam optimizer is used with a learning rate initialized to $lr = 0.1$ and decreasing with the number of epochs. The neural network was implemented with Tensorflow 2.0 framework and trained on one NVIDIA 1080 Ti GPU. The offline stage of the CNN takes less than 1h for training and the online stages takes around 30s.

5.3.3 Results

The proposed approach was evaluated qualitatively and quantitatively. In Fig: 5.5(a), we show the position of the center of mass of the segmented cochlea $\mathbf{C}_{\text{target}}$ during three iterations of Algorithm 4. We see that the 3 points are getting closer to each other after each iteration thus demonstrating the convergence of the algorithm. In practice, we found that between 2 to 6 iterations are necessary to get a change of mass center position between two iterations less than 1mm.

For a quantitative assessment of performance, an expert positioned twice the 3 landmarks on 20 additional volumes in order to estimate the positioning error and the intra-rater variability. In Fig: 5.5(d) we show the 3 landmarks generated by our algorithm with the same landmarks positioned by the expert. Clearly those points are very close to each other on the 3 views. In Table 5.1(top), we list the average position error of the 3 landmarks on the 20 images with respect to one set of landmarks manually positioned by an expert. In average, the position error of L_{target} is around 0.6mm which corresponds to a difference of position of 2 to 3 voxels. This result is satisfactory when considering the small size of the cochlea (width: 6.53 $\pm 0.35mm$, height: 3.26 $\pm 0.24mm$ Zahara et al. [2019]) within the full CT volume ($128mm \times 128mm \times 55mm$). For a better assessment, we also provide the intra-expert landmark position error in Table 5.1(bottom). It shows that the algorithm error is similar to the intra-expert variability, with a lower error for two (the center and window landmarks) out of the three landmarks. When computing the landmark position error with the second set of landmarks made by the expert, or with the average of the 2 annotations, we also found that the algorithm was performing similarly to the expert. Since the intra-rater

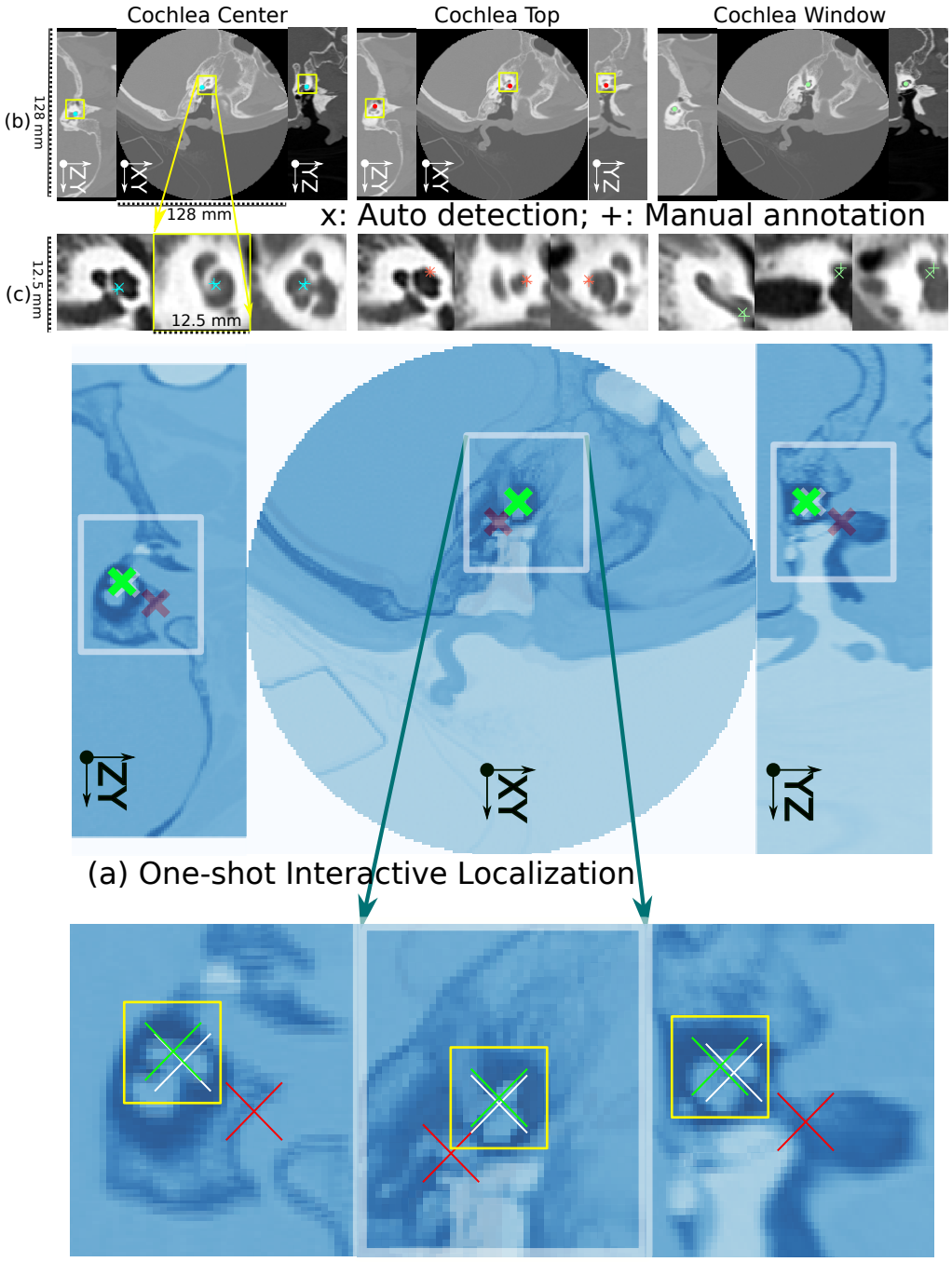

Figure 5.5: (a) Positions of the center of mass of the cochlea during 3 iterations of the translation offset determination. The 3 cross marks in red, white, green correspond to the 1st, 2nd, 3rd iterations; Row (b) shows the result of the landmarks detection in the whole image I_{target}; Row (c) zooms on the detected landmarks before applying the last registration stage; Row (d) zooms on the generated landmarks ('x' marks) after the registration stage and the manually positioned landmarks ('+' marks) by an expert.

variability is in most cases lower than inter-rater variability, we believe that the proposed method is an effective way to automate landmark positioning around the cochlea on CT images. Note that the mean landmark position errors reported by Zhang *et al.* Zhang et al. [2017] also correspond between 2.5 to 3 times the voxel size whereas Grewal et al. [2020] after training on 168 scans reports errors between 2 to 9 times the voxel size ($2 - 9mm$).

Table 5.1: Position errors of the 3 cochlear landmarks (centre, top and window) automatically generated landmarks (AUTO) and a second set of manual (MANU) ones.

Image ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	μ/σ	p < .05
CEN AUTO	0.88	0.28	0.49	0.70	0.72	0.57	0.19	0.49	0.49	0.39	0.65	0.84	0.87	0.67	0.72	0.72	0.73	0.54	0.33	0.39	**0.58**$\pm$0.19mm	0.016
TOP AUTO	0.70	1.33	0.56	0.73	0.72	0.37	0.31	0.78	0.35	0.20	1.63	1.15	1.00	0.26	1.04	0.67	0.80	1.23	0.55	0.39	0.73$\pm$0.39mm	0.003
WIN AUTO	0.86	0.65	0.84	0.55	0.65	1.12	1.35	0.31	0.60	0.49	0.26	1.06	1.54	0.72	0.88	0.81	0.54	0.34	1.43	0.88	**0.79**$\pm$0.36mm	0.007
CEN MANU	0.28	0.56	0.53	1.06	0.65	0.59	0.45	0.57	0.25	1.09	0.94	0.84	1.09	0.53	0.37	0.50	0.25	0.54	0.59	0.30	0.60$\pm$0.27mm	0.006
TOP MANU	0.43	0.38	0.49	0.25	0.31	0.25	0.31	0.19	0.24	1.09	0.00	0.50	0.75	0.25	0.31	0.19	0.60	0.42	0.33	0.66	**0.40**$\pm$0.24mm	0.002
WIN MANU	0.69	0.62	1.11	1.10	0.31	1.07	0.31	0.77	0.43	0.57	0.79	1.22	0.91	0.77	0.97	0.75	0.90	1.01	1.18	1.25	0.84$\pm$0.29mm	0.017

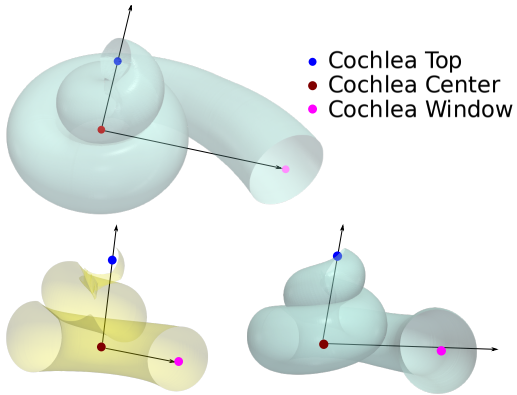

Figure 5.6: Cochlea landmarks shown with three import coordinates (cochlea top, cochlea center and cochlea round window points) of cochlea model.

5.4 Conclusion

To the best of our knowledge, the proposed method is the first one-shot learning approach for 3D landmarks detection in volumetric images. We showed that the proposed approach was effective in localizing 3D landmarks in the cochlea from CT images of the inner ear. It relies on a segmentation stage and the registration of a single user-defined image patch which makes it easy explainable and interpretable. The approach is generic and could be applied to the detection of landmarks in CT imaging and other imaging modalities. In the future, we plan to use more complex image similarity measures in the final registration algorithm and to introduce more annotated data (few-shot learning) to address challenging landmark detection problems. Other network architectures proposed in the literature for one-shot deep learning such as [Chen and Zhang, 2019, Jadon and Srinivasan, 2019, Koch, 2015, Shaban et al., 2017] can be explored.

Chapter 6

Conclusion and Perspectives

The thesis aims to explore the applications of generative models in medical image processing, analysis and modeling. We have presented in previous chapters various practical applications of the generative learning mostly on inner ear CT volumes dataset.

6.1 Main Contributions

6.1.1 Deep Generative Learning for Medical Image Processing

We have shown in chapter 2 how to make use of GANs for cochlear implant CT metal artifacts reduction. We constructed a training dataset through the simulation of the CI insertion and the metal artifacts. The simulation of electrode array insertion of CI is realized through a cochlea shape model fitting framework which is detailed in chapter 3 and artifacts are simulated by accounting for three typical X-ray physical phenomena: the bean hardening, the detector noise, and the scattering effect. The simulation of the metal artifacts allows us to build artifacts and artifacts free pair volumes for training the GAN. The proposed GAN based method is evaluated on the conventional CT images and preliminary experiments showed to be somewhat effective on cone beam CT images. The result implies that the deep generative models are reasonable robust across different modalities.

6.1.2 Shape Based Medical Image Segmentation: a Bayesian perspective

In chapter 3, we have presented a Bayesian logistic shape inference framework for cochlea CT image segmentation. Specifically, we address the issue of the Bayesian inference of parametric shape models in a purely interpretable way. We evaluated the proposed framework on 3 different cochlea CT image datasets which include more than 250 patients' CT images. The segmentation results on clinical CT images show performances comparable to supervised deep learning approaches by quantitative evaluation based on Dice score. The major contributions of this chapter lie in: (1) A novel framework for image segmentation that combines probabilistic appearance and shape models. It is generically defined for parametric shape functions rather than parametric space transformations. The trade-off between the appearance and shape models is governed by an interpretable parameter: the reference length. (2) A Gauss-Newton method of optimizing the shape parameters, which also produces a posterior approximation of those parameters. (3) A method for uncertainty quantification of the image segmentation that takes into account the shape uncertainty. (4) A segmentation method of the cochlea in clinical CT images that provides state-of-the-art results and interpretable shape parameters.

6.1.3 Variational Generative Learning for Medical Data Modeling

An important feature of generative learning is that the learned posterior model can be sampled for generating unseen data. This feature is very useful for medical dataset generation as the data collection is limited. The generative model can better express the target data by dimensionality reduction which can be used for dataset modeling for data sharing. We have shown in the chapter 4 an adapted flow-based variational auto-encoder with practical application in medical dataset modeling. Different from the conventional VAE, the proposed Langevin-VAE introduces the target information by involving the information of the target dataset, which allows us to improve the inference fidelity. A better inference quality results in better modeling of the target dataset and allows us to generate high-quality samples from the latent parameters. The proposed Langevin-VAE is used for cochlea CT images generation for offering high-quality synthetic cochlea CT images.

6.1.4 One-shot Learning for Landmarks Detection

A common flaw of popular used deep learning approaches is the requirement of massive dataset for fitting. However, medical datasets are usually difficult to collect due to many ethical or legal reasons. Massive data collection may require tedious and long procedures to get legal authorizations. Another problem is the annotation of 3D landmarks for a fairly large dataset is very expensive both in terms of time and manpower. Those data barriers limit the potential application of learning approaches in clinical settings. To tackle this problem, in the chapter 5, we proposed a one-shot learning-based landmarks detection approach which allows us to use a single volume to detect landmarks in hundreds of volumes. The method is applied to a large scale CT cochlea volume dataset for detecting the landmarks of the cochlea. This method solves the problem of lack of training data and reduces the burden of human experts for data annotation.

6.1.5 Clinical Impact

This work shows that deep learning and generative learning can bring new solutions to medical imaging analysis in many different aspects. We see that the introduction of shape information through a Bayesian learning framework

can help the human expert to better understand the shape uncertainty of the cochlea which can assist the clinician to reduce the risk of CI electrode array positioning misplacement. The patients' personalized cochlea shape analysis should contribute to overcome the limitations of the traditional used fixed cochlea shape model for CI planning. The personalized cochlea shape analysis is an example of precision medicine.

The use of MARGAN for CI postoperative CT images metal artifacts reduction can reduce the metal artifacts significantly in comparison with traditional artifacts reduction methods. The MARGAN can also bring the positioning information of the electrodes array inside the cochlea. This can help the expert to assess the quality of the CI surgery and to know the correlation between the prognosis and CI electrode array position. Variational auto-encoder for cochlea dataset modeling is another example of generative learning applied in medical dataset modeling. The generation of fake cochlea images can be used for data augmentation, case study, knowledge domain transfer etc. The feature space of the generation model can be used for shape analysis and classification.

One-shot learning has the potential to make a major change in the computerized medical image domain due to the barriers in clinical data collection. The application of one-shot learning in landmarks detection can reduce the workload of data annotation. The one-shot learning-based landmarks detection can also be a prior step for building segmentation algorithms and shape analysis frameworks.

6.2 Perspectives

6.2.1 Trustable Learning based Computed Medical Data Analysis

The fast development of artificial intelligence brings many advantages in various disciplines. A typical example is the computer vision (CV) community which has seen rapid advances in the state of the art algorithms. Medical image analysis which is largely overlapping the CV community, is also fast evolving in many different directions. However, more and more research attention is focused on the performance of algorithms instead of their clinical applicability. The lack of consideration of clinical applicability is harmful to the computerized medical community and may lead to unpredictable results in clinical practice.

Typically, the presence of outliers might be a non-threatening issue in many civil applications, yet may have a significant impact in clinical practice. The use of MARGAN for metal artifacts reduction is an example that a neural network learns from a major group of healthy patients with different degrees of deafness. We concluded that the MARGAN performs well on the cochlea dataset for metal artifacts reduction. Yet, whether the processing of MARGAN will eliminate important features of unseen lesions is unknown. The study of the reliability of current learning based diagnosing approaches is worthwhile and important for clinical propagation.

6.2.2 Parameters coupling between Parametric Shape model and Deep Latent Model

We have shown that the Langevin-VAE can generate cochlea image volumes from a set of latent variables which can be treated as a group of shape and appearance parameters. The latent parameters of Langevin-VAE are independent representations from the cochlea parametric shape model parameters and may also involve intensity representations. Currently, shape and appearance features are represented jointly by the latent variables for Langevin-VAE. However, the representation space of Langevin-VAE and cochlea parametric shape and appearance models may be coupled to get an unified representation of the shapes and other attributions of the volumes. This can help us to get a geometrical meaningful generative model for generating customizable shape volumes.

6.2.3 Attention in medical image analysis

Since the Attention-based learning models were proposed, it has rapidly expanded from the field of natural language processing (NLP) to the entire machine learning community [Carion et al., 2020a, Vyas et al., 2020, Yuan et al., 2021]. As a very successful application of the concept of attention, the Transformer already significantly impacted the NLP field and almost replaced the previously widely-used RNN/LSTM models. Recent works in the computer vision community report that the Transformer based deep learning methods achieved state-of-the-art performance on many datasets. Vision Transformer (ViT) is a representative work that introduces the Transformer in computer vision for image classification [Dosovitskiy et al., 2020]. The ViT takes divided patches p_i of a given image I as

inputs and uses a Transformer as attention features extractor to generate features for classification [Carion et al., 2020b]. Another work that uses a Transformer for object detection is the DETR [Carion et al., 2020a]. The DETR network accepts the image and the detection objects for the Transformer as inputs and outputs the corresponding bounding box information of the target objects on the images.

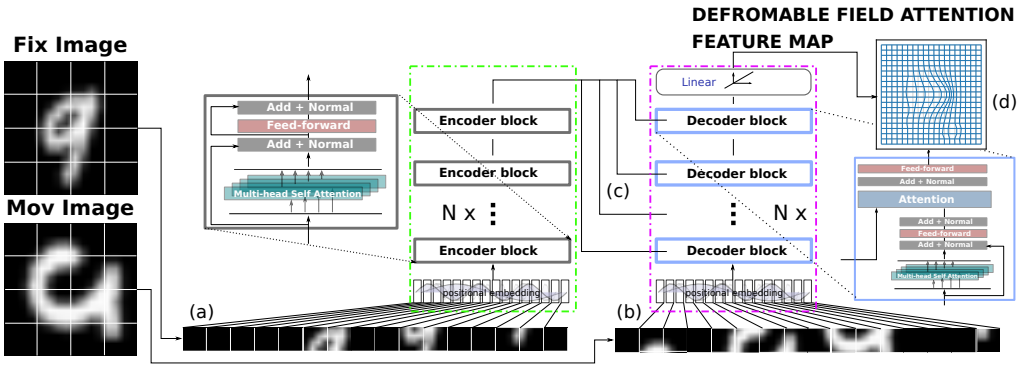

Figure 6.1: The transformer framework used for deformation attention feature map prediction. The proposed transformer consists of encoder and decoder modules. The encoder module (shown in the green dotted line box) takes the fixed images patches as input and learns the representation of the memory attention features with self-attention mechanism. The decoder module (shown in the purple dotted line box) takes the attention features of the fixed image from the encoder (memory) and the self-attentions features of the moving image as input for predicting the deformable features that can transform the moving image into a fixed image.

As a recent model in machine learning, attention-based methods have the potential chance to be the next machine learning model to provide advanced performances. As of now (May 2021), the Transformer based method have only been introduced marginally into medical-related domains. An interesting ongoing project is therefore to use Transformers for image registration.

We detail below the ongoing work of using a Transformer model [Vaswani et al., 2017] for image registration. To feed images to Transformer, they need to be processed as sequence data-points [Dosovitskiy et al., 2020]. To this end, the fixed I_{fix} and moving I_{mov} images pairs are first divided into $i \times i$ (see in Fig .6.1 step (a) where i = 4) patches with added position embedding to get the position information. The fixed image patches are fed to transformer encoder network which encodes the attention features of the fixed images. The transformer

encoder network contains N recursive blocks as shown in the green box of Fig. 6.1. Simultaneously, the moving images patches are fed together with the output attention features to the transformer decoder network (step (c) in Fig. 6.1). The output of the Transformer decoder network is processed with a linear projection layer to generate the deformation attention feature maps.

Our primary experiment shows that the Transformer model can learn a deformable representation between two similar shapes for image registration. Yet, the quality of the learnt transformations is not good in the shape details. This problem may due to the Transformer is not good at pixel-level image processing. Our future work will be focused on the improvement of the current framework for solving current issues of the Transformer based registration model.

6.3 Publications

6.3.1 Journal Articles

AI applications in medical imaging.

- Wang, Vandersteen, Demarcy, Gnansia, Raffaelli, Guevara, and Delingette [2021b]. Inner-ear Augmented Metal Artifact Reduction with Simulation-based 3D Generative Adversarial Networks. CoRR abs/2104.12510 (2021) (Submitted to Journal of Computerized Medical Imaging and Graphics, under major revision.)
- Wang, Demarcy, Vandersteen, Gnansia, Raffaelli, Guevara, and Delingette [2021a]. Bayesian Logistic Shape Model Inference: application to cochlea image segmentation. CoRR abs/2105.02045 (2021) (Submitted to Medical Image Analysis, under review.)

AI applications in signal processing.

- Wang, Xu, He, Hwang, Fan, and Delingette [2020f]. Long Short-Term Memory Neural Equalizer. 2020. ⟨hal-03022865⟩ (Submitted to IEEE Transaction of Signal Processing)
- Xu, Wang, Sun, Hwang, Delingette, and Fan [2021]. Jitter-Aware Economic PDN Optimization With a Genetic Algorithm, in IEEE Transactions on Microwave Theory and Techniques (Early Access), doi: 10.1109/TMTT.2021.3087188.

6.3.2 Conference Papers

- Wang, Vandersteen, Demarcy, Gnansia, Raffaelli, Guevara, and Delingette [2019d]. Deep Learning based Metal Artifacts Reduction in post-operative Cochlear Implant CT Imaging. In MICCAI 2019 - 22nd International Conference on Medical Image Computing and Computer Assisted Intervention, Shenzhen, China, pages 121-129, October 2019.
- Wang, Vandersteen, Demarcy, Gnansia, Raffaelli, Guevara, and Delingette [2020b]. A Deep Learning based Fast Signed Distance Map Generation. In MIDL 2020 - Medical Imaging with Deep Learning, Montréal, Canada, July 2020.
- Jia, Despinasse, Wang, Delingette, Pennec, Jaïs, Cochet, and Sermesant [2018]. Automatically Segmenting the Left Atrium from Cardiac Images Using Successive 3D U-Nets and a Contour Loss. In STACOM: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges, volume 11395 of LNCS, Granada, Spain, pages 221-229, September 2018.
- Wang, Vandersteen, Raffaelli, Guevara, and Delingette [2020e]. One-shot Learning Landmarks Detection. Note: Working paper or preprint, November 2020. (Accepeted by MICCAI DALI 2021 : 1st MICCAI Workshop on Data Augmentation, Labeling, and Imperfections workshop.)
- Wang and Delingette [2021b]. Quasi-Symplectic Langevin Variational Autoencoder, June 2021. (Submitted to AAAI 2022.)

6.3.3 Preprints or working papers.

- Wang and Delingette [2021a]. Attention for Image Registration (AiR): an unsupervised Transformer approach. Note: Working paper or preprint, May 2021. Keyword(s): Transformer, Images Registration, Deep Learning.

References

Graeme Milbourne Clark AC. Graeme Clark- CV Milestones and Achievements 2010. *International Archives of Otorhinolaryngology*, 2010. URL https://graemeclarkfoundation.org/wp-content/uploads/2010/09/Graeme-Clark-CV-Milestones-Achievments-2.pdf.

Mikael Agn, Per Munck af Rosenschöld, Oula Puonti, Michael J. Lundemann, Laura Mancini, Anastasia Papadaki, Steffi Thust, John Ashburner, Ian Law, and Koen Van Leemput. A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Medical Image Analysis*, 54:220 – 237, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.03.005. URL http://www.sciencedirect.com/science/article/pii/S1361841518305103.

Hammam Alshazly, Christoph Linse, Erhardt Barth, and Thomas Martinetz. Ensembles of deep learning models and transfer learning for ear recognition. *Sensors*, 19(19), 2019. ISSN 1424-8220. doi: 10.3390/s19194139. URL https://www.mdpi.com/1424-8220/19/19/4139.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. doi: 10.1137/050637996. URL http://www-sop.inria.fr/asclepios/Publications/Vincent.Arsigny/Arsigny_SIAM_tensors_07.pdf.

John Ashburner and Karl J. Friston. Unified segmentation. *NeuroImage*, 26(3): 839 – 851, 2005a. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.

2005.02.018. URL http://www.sciencedirect.com/science/article/pii/S1053811905001102.

John Ashburner and Karl J. Friston. Unified segmentation. *NeuroImage*, 26(3): 839–851, 2005b. ISSN 10538119. doi: 10.1016/j.neuroimage.2005.02.018.

Benoît Audelan and Hervé Delingette. Unsupervised Quality Control of Image Segmentation based on Bayesian Learning. In *MICCAI 2019 - 22nd International Conference on Medical Image Computing and Computer Assisted Intervention*, Shenzhen, China, October 2019.

Ersin Avci, Tim Nauwelaers, Thomas Lenarz, Volkmar Hamacher, and Andrej Kral. Variations in microanatomy of the human cochlea. *The Journal of Comparative Neurology*, 00:1–17, mar 2014. ISSN 1096-9861. doi: 10.1002/cne.23594. URL http://www.ncbi.nlm.nih.gov/pubmed/24668424.

Andreu Badal and Aldo Badano. Accelerating monte carlo simulations of photon transport in a voxelized geometry using a massively parallel graphics processing unit. *Medical Physics*, 36(11):4878–4880, 2009. doi: 10.1118/1.3231824. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3231824.

J. A. Baerentzen and H. Aanaes. Signed distance computation using the angle weighted pseudonormal. *IEEE Transactions on Visualization and Computer Graphics*, 11(3):243–253, May 2005. ISSN 2160-9306. doi: 10.1109/TVCG.2005.49.

Gavin Baker. *Tracking, modelling and registration of anatomical objects: the human cochlea*. PhD thesis, The University of Melbourne, 2008. URL http://minerva-access.unimelb.edu.au/handle/11343/37464.

Gavin Baker and Nick Barnes. Model-image registration of parametric shape models: fitting a shell to the cochlea. *Insight Journal*, 2005. URL http://users.cecs.anu.edu.au/{~}nmb/papers/ij2005{_}mireg.pdf.

Alessandro Barp, Francois-Xavier Briol, Anthony Kennedy, and Mark Girolami. Geometry and dynamics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 5, 05 2017. doi: 10.1146/annurev-statistics-031017-100141.

Thomas Matthew Benson and Bruno K. B. De Man. Synthetic ct noise emulation

in the raw data domain. *IEEE Nuclear Science Symposium & Medical Imaging Conference*, pages 3169–3171, 2010.

Ricardo Ferreira Bento, Fabiana Danieli, Ana Tereza de Matos Magalh, Dan Gnansia, and Michel Hoen. Residual Hearing Preservation with the Evo A Cochlear Implant Electrode Array: Preliminary Results. *International Archives of Otorhinolaryngology*, 20:353 – 358, 12 2016. ISSN 1809-4864. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1809-48642016000400353&nrm=iso.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Kirsten Bolstad, Silje Flatabø, Daniel Aadnevik, Ingvild Dalehaug, and Nils Vetti. Metal artifact reduction in ct, a phantom study: subjective and objective evaluation of four commercial metal artifact reduction algorithms when used on three different orthopedic metal implants. *Acta Radiologica*, 59:028418511775127, 01 2018. doi: 10.1177/0284185117751278.

Ali Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2018.10.009. URL https://www.sciencedirect.com/science/article/pii/S1077314218304272.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020a. Springer International Publishing. ISBN 978-3-030-58452-8.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm,

editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020b. Springer International Publishing. ISBN 978-3-030-58452-8.

Anthony L. Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder. In *NeurIPS*, 2018.

Marco Caversaccio, Wilhelm Wimmer, Juan Anso, Georgios Mantokoudis, Nicolas Gerber, Christoph Rathgeb, Daniel Schneider, Jan Hermann, Franca Wagner, Olivier Scheidegger, Markus Huth, Lukas Anschuetz, Martin Kompis, Tom Williamson, Brett Bell, Kate Gavaghan, and Stefan Weber. Robotic middle ear access for cochlear implantation: First in man. *PLOS ONE*, 14(8):1–12, 08 2019. doi: 10.1371/journal.pone.0220543. URL https://doi.org/10.1371/journal.pone.0220543.

T. Chan and Wei Zhu. Level set based shape prior segmentation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1164–1170 vol. 2, June 2005. doi: 10.1109/CVPR.2005.212.

Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification, 2021.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Warren Cheung and Ghassan Hamarneh. n-sift: n-dimensional scale invariant feature transform. *Image Processing, IEEE Transactions on*, pages 2012 – 2021, 10 2009. doi: 10.1109/TIP.2009.2024578.

Lars Chittka and Axel Brockmann. Perception space—the final frontier. *PLOS Biology*, 3(4), 04 2005. doi: 10.1371/journal.pbio.0030137. URL https://doi.org/10.1371/journal.pbio.0030137.

Lawrence T Cohen, Jin Xu, Shi Ang Xu, and Graeme M Clark. Improved and simplified methods for specifying positions of the electrode bands of a cochlear implant array. *The American Journal of Otology*, 17(6):859–865, 1996. ISSN 0192-9763. URL http://cat.inist.fr/?aModele=afficheN{&}cpsidt=2488723.

Olivier Commowick and Simon Warfield. A continuous staple for scalar, vector, and tensor images: An application to dti analysis. *IEEE transactions on medical imaging*, 28:838–46, 06 2009. doi: 10.1109/TMI.2008.2010438.

Steve Connor, D.J. Bell, Ruth O'Gorman, and A Fitzgerald-O'Connor. Ct and mr imaging cochlear distance measurements may predict cochlear implant length required for a 360° insertion. *AJNR. American journal of neuroradiology*, 30: 1425–30, 05 2009. doi: 10.3174/ajnr.A1571.

Tim F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, jan 1995. ISSN 10773142. doi: 10.1006/cviu.1995.1004.

Jason J. Corso, Raja' S. Alomari, and Vipin Chaudhary. Lumbar disc localization and labeling with a probabilistic model on both pixel and object features. In Dimitris Metaxas, Leon Axel, Gabor Fichtinger, and Gábor Székely, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, pages 202–210, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-85988-8.

D. Cremers, T. Kohlberger, and C. Schnörr. Shape Statistics in Kernel Space for Variational Image Segmentation. *Pattern Recognition*, 36(9):1929–1943, 2003.

Daniel Cremers. A variational framework for image segmentation combining motion estimation and shape regularization. In *CVPR (1)*, pages 53–58. IEEE Computer Society, 2003.

Antonio Criminisi, Toby Sharp, and Andrew Blake. GeoS: Geodesic Image Segmentation. *ECCV*, pages 99–112, 2008. URL http://link.springer.com/content/pdf/10.1007/978-3-540-88682-2{_}9.pdf.

J.J. Crisco and R.D. McGovern. Efficient calculation of mass moments of inertia for segmented homogenous three-dimensional objects. *Journal of Biomechanics*, 31(1):97 – 101, 1997. ISSN 0021-9290. doi: https://doi.org/10.1016/S0021-9290(97)00108-5. URL http://www.sciencedirect.com/science/article/pii/S0021929097001085.

Thomas Demarcy. *Segmentation and study of anatomical variability of the cochlea from medical images*. Theses, Université Côte d'Azur, July 2017a.

Thomas Demarcy. *Segmentation and study of anatomical variability of the cochlea from medical images*. Theses, Université Côte d'Azur, July 2017b. URL https://tel.archives-ouvertes.fr/tel-01609910.

Thomas Demarcy, Clair Vandersteen, Nicolas Guevara, Charles Raffaelli, Dan Gnansia, Nicholas Ayache, and Hervé Delingette. Automated analysis of human cochlea shape variability from segmented μ CT images. *Computerized Medical Imaging and Graphics*, 59:1–12, 2017. ISSN 08956111. doi: 10.1016/j.compmedimag.2017.04.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0895611117300332.

Anandhan Dhanasingh and Claude Jolly. Review on cochlear implant electrode array tip fold-over and scalar deviation. *Journal of Otology*, 14(3):94–100, 2019. ISSN 1672-2930. doi: https://doi.org/10.1016/j.joto.2019.01.002. URL https://www.sciencedirect.com/science/article/pii/S1672293018300874.

Felix E Diehn, Gregory J Michalak, David R DeLone, Amy L Kotsenas, E Paul Lindell, Norbert G Campeau, Ahmed F Halaweish, Cynthia H McCollough, and Joel G Fletcher. Ct dental artifact: Comparison of an iterative metal artifact reduction technique with weighted filtered back-projection. *Acta radiologica open*, 6(11):2058460117743279, November 2017. ISSN 2058-4601. doi: 10.1177/2058460117743279.

René Donner, Bjoern H. Menze, Horst Bischof, and Georg Langs. Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. *Medical Image Analysis*, 17(8):1304 – 1314, 2013. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2013.02.004.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

S. Elhabian and R. Whitaker. Shapeodds: Variational bayesian learning of generative shape models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 2185–2196, July 2017. doi: 10.1109/CVPR.2017.235. URL doi.ieeecomputersociety.org/10.1109/CVPR.2017.235.

Francesco Fassò and Nicola Sansonetto. Integrable almost-symplectic hamiltonian systems. *Journal of Mathematical Physics*, 48(9):092902, 2007. doi: 10.1063/1.2783937. URL https://doi.org/10.1063/1.2783937.

Ricardo José Ferrari, Stéphane Allaire, Andrew Hope, John J. Kim, David A. Jaffray, and Vladimir Pekar. Detection of point landmarks in 3d medical images via phase congruency model. *Journal of the Brazilian Computer Society*, 17: 117–132, 2011.

Glover GH. Compton scatter effects in ct reconstructions. *Med Phys.*, 9(6): 860-867, 1982. ISSN 1053-8119. doi: doi:10.1118/1.595197.

F. Ghesu, B. Georgescu, Y. Zheng, S. Grbic, A. Maier, J. Hornegger, and D. Comaniciu. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):176–189, Jan 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2017.2782687.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. doi: 10.1111/j.1467-9868.2010.00765.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00765.x.

Lars Gjesteby, Qingsong Yang, Yan Xi, Ye Zhou, Junping Zhang, and Ge Wang. Deep learning methods to guide CT image reconstruction and reduce metal artifacts. In Thomas G. Flohr, Joseph Y. Lo, and Taly Gilat Schmidt, editors, *Medical Imaging 2017: Physics of Medical Imaging*, volume 10132, pages 752 – 758. International Society for Optics and Photonics, SPIE, 2017. doi: 10.1117/12.2254091. URL https://doi.org/10.1117/12.2254091.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Karol Gregor, Ivo Danihelka, Andriy Mnih, Charles Blundell, and Daan Wierstra. Deep autoregressive networks. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1242–1250, Bejing,

China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/gregor14.html.

Monika Grewal, Timo M. Deist, Jan Wiersma, Peter A.N. Bosman, and Tanja Alderliesten. An end-to-end deep learning approach for landmark detection and matching in medical images. *Progress in Biomedical Optics and Imaging*, 11313:1131328–1 – 1131328–10, 2020. ISSN 1605-7422. doi: 10.1117/12.2549302. Medical Imaging 2020: Image Processing ; Conference date: 17-02-2020 Through 20-02-2020.

Minghao Gu, Shiliang Sun, and Yan Liu. Dynamical sampling with langevin normalization flows. *Entropy*, 21:1096, 11 2019. doi: 10.3390/e21111096.

Shouhei Hanaoka, Akinobu Shimizu, Mitsutaka Nemoto, Yukihiro Nomura, Soichiro Miki, Takeharu Yoshikawa, Naoto Hayashi, Kuni Ohtomo, and Yoshitaka Masutani. Automatic detection of over 100 anatomical landmarks in medical ct images: A framework with independent detectors and combinatorial optimization. *Medical Image Analysis*, 35:192 – 214, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.04.001.

Tobias Heimann, Sascha Münzing, Hans-Peter Meinzer, and Ivo Wolf. A shape-guided deformable model with evolutionary algorithm initialization for 3D soft tissue segmentation. *Inf Process Med Imaging*, 20:1–12, 2007. ISSN 1011-2499.

Floris Heutink, Valentin Koch, Berit Verbist, Willem Jan van der Woude, Emmanuel Mylanus, Wendy Huinck, Ioannis Sechopoulos, and Marco Caballo. Multi-scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution ct images. *Computer Methods and Programs in Biomedicine*, 191:105387, 2020. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105387. URL https://www.sciencedirect.com/science/article/pii/S0169260719320231.

Geoffrey Hinton. *Deep Belief Nets*, pages 267–269. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_208. URL https://doi.org/10.1007/978-0-387-30164-8_208.

Chin-Wei Huang, Kris Sankaran, Eeshan Dhekane, Alexandre Lacoste, and Aaron Courville. Hierarchical importance weighted autoencoders. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th*

International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 2869–2878. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/huang19d.html.

Xia Huang, Jian Wang, Fan Tang, Tao Zhong, and Yu Zhang. Metal artifact reduction on cervical ct images by deep residual learning. *BioMedical Engineering OnLine*, 17, 12 2018. doi: 10.1186/s12938-018-0609-y.

Shruti Jadon and Aditya Srinivasan. Improving siamese networks for one shot learning using kernel based activation functions. *ArXiv*, abs/1910.09798, 2019.

A. Jaklic and F. Solina. Moments of superellipsoids and their application to range image registration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(4):648–657, 2003.

Shuman Jia, Antoine Despinasse, Zihao Wang, Hervé Delingette, Xavier Pennec, Pierre Jaïs, Hubert Cochet, and Maxime Sermesant. Automatically Segmenting the Left Atrium from Cardiac Images Using Successive 3D U-Nets and a Contour Loss. In *Statistical Atlases and Computational Modeling of the Heart (STACOM) workshop*, Granada, Spain, September 2018.

Jian Huang, Yan Li, R. Crawfis, Shao-Chiung Lu, and Shuh-Yuan Liou. A complete distance field representation. In *Proceedings Visualization, 2001.*, pages 247–561, Oct 2001.

M. W. Jones, J. A. Baerentzen, and M. Sramek. 3d distance fields: a survey of techniques and applications. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):581–599, July 2006. ISSN 2160-9306. doi: 10.1109/TVCG.2006.56.

W A Kalender, R Hebel, and J Ebersberger. Reduction of ct artifacts caused by metallic implants. *Radiology*, 164(2):576–577, 1987a. doi: 10.1148/radiology.164.2.3602406. URL https://doi.org/10.1148/radiology.164.2.3602406. PMID: 3602406.

W A Kalender, R Hebel, and J Ebersberger. Reduction of ct artifacts caused by metallic implants. *Radiology*, 164(2):576–577, 1987b. doi: 10.1148/radiology.164.2.3602406. URL https://doi.org/10.1148/radiology.164.2.3602406. PMID: 3602406.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

H. Kjer, J. Fagertun, W. Wimmer, N. Gerber, Sergio Vera, Livia Barazzetti, N. Lopez, M. Ceresa, G. Piella, Thomas Stark, M. Stauber, M. Reyes, S. Weber, M. Caversaccio, M. Ballester, and R. Paulsen. Patient-specific estimation of detailed cochlear shape from clinical ct images. *International Journal of Computer Assisted Radiology and Surgery*, 13:389–396, 2017.

Hans Martin Kjer and Rasmus Reinhold Paulsen. *Modelling of the Human Inner Ear Anatomy and Variability for Cochlear Implant Applications*. PhD thesis, Technical University of Denmark (DTU), 2015.

Hans Martin Kjer, Sergio Vera, Frederic Pérez, Miguel Angel González Ballester, and Rasmus Reinhold Paulsen. Semi-automatic anatomical measurements on microCT 3D surface models. In *International Conference on Cochlear Implants and Other Implantable Auditory Technologies, Munich, Germany*, page 711, 2014.

I. Kobyzev, S. Prince, and M. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

Gregory R. Koch. Siamese neural networks for one-shot image recognition. 2015.

Maria-Izabel Kós, Colette Boëx, Alain Sigrist, J.-P Guyot, and Marco Pelizzone. Measurements of electrode position inside the cochlea for different cochlear implant systems. *Acta oto-laryngologica*, 125:474–80, 06 2005. doi: 10.1080/00016480510039995.

Edwin H. Land and John J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, Jan 1971. doi: 10.1364/JOSA.61.000001. URL http://www.osapublishing.org/abstract.cfm?URI=josa-61-1-1.

L. Le Folgoc, H. Delingette, A. Criminisi, and N. Ayache. Quantifying registration uncertainty with sparse bayesian modelling. *IEEE Transactions on Medical Imaging*, 36(2):607–617, 2017. doi: 10.1109/TMI.2016.2623608.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Hongwei Li, Rameshwara G. N. Prasad, Anjany Sekuboyina, Chen Niu, Siwei Bai, Werner Hemmert, and Bjoern Menze. Micro-ct synthesis and inner ear super resolution via generative adversarial networks and bayesian inference, 2021.

Xiaokun Liang, Wei Zhao, Dimitre H. Hristov, Mark K. Buyyounouski, Steven L. Hancock, Hilary Bagshaw, Qin Zhang, Yaoqin Xie, and Lei Xing. A deep learning framework for prostate localization in cone beam ct-guided radiotherapy. *Medical Physics*, n/a(n/a). doi: 10.1002/mp.14355.

Samuel Livingstone and Mark A. Girolami. Information-geometric markov chain monte carlo methods using diffusions. *Entropy*, 16:3074–3102, 2014.

Yi Lv, Jia Ke, Ying Xu, Yu Shen, Junchen Wang, and Jiang Wang. Automatic segmentation of temporal bone structures from clinical conventional ct using a cnn approach. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 17(2):e2229, 2021. doi: https://doi.org/10.1002/rcs.2229. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rcs.2229.

Yuanyuan Lyu, Wei-An Lin, Jingjing Lu, and Shaohua Kevin Zhou. Dudonet++: Encoding mask projection to reduce CT metal artifacts. *CoRR*, abs/2001.00340, 2020. URL http://arxiv.org/abs/2001.00340.

Geoffrey A. Manley. *The Cochlea: What It Is, Where It Came From, and What Is Special About It*, pages 17–32. Springer International Publishing, Cham, 2017. ISBN 978-3-319-52073-5. doi: 10.1007/978-3-319-52073-5_2. URL https://doi.org/10.1007/978-3-319-52073-5_2.

C. R. Maurer, Rensheng Qi, and V. Raghavan. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary

dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):265–270, Feb 2003. ISSN 1939-3539. doi: 10.1109/TPAMI.2003.1177156.

Calvin R. Maurer, Rensheng Qi, Vijay Raghavan, and Senior Member. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 265–270, 2003.

A. Mehranian, M. R. Ay, A. Rahmim, and H. Zaidi. X-ray ct metal artifact reduction using wavelet domain l_0 sparse regularization. *IEEE Transactions on Medical Imaging*, 32(9):1707–1722, 2013.

E. Meyer, F. Bergner, R. Raupach, T. Flohr, and M. Kachelrieb. Normalized metal artifact reduction (nmar) in computed tomography. In *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pages 3251–3255, 2009. doi: 10.1109/NSSMIC.2009.5401721.

Esther Meyer, Rainer Raupach, Michael Lell, Bernhard Schmidt, and Marc Kachelrieb. Normalized metal artifact reduction (nmar) in computed tomography. *Medical Physics*, 37(10):5482–5493, 2010. doi: 10.1118/1.3484090. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3484090.

Esther *et al.* Meyer. Normalized metal artifact reduction (nmar) in computed tomography. *Medical Physics*, 37(10):5482–5493, 2010. doi: 10.1118/1.3484090. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3484090.

G. Milstein. Quasi-symplectic methods for langevin-type equations. *IMA Journal of Numerical Analysis*, 23:593–626, 10 2003. doi: 10.1093/imanum/23.4.593.

G. N. Milstein, Yu. M. Repin, and M. V. Tretyakov. Symplectic integration of hamiltonian systems with additive noise. *SIAM Journal on Numerical Analysis*, 39(6):2066–2088, 2002. doi: 10.1137/S0036142901387440. URL https://doi.org/10.1137/S0036142901387440.

Oliver Mothes and Joachim Denzler. Anatomical landmark tracking by one-shot learned priors for augmented active appearance models. pages 246–254, 01 2017. doi: 10.5220/0006133302460254.

Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I.

Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm. 2020.

Megumi Nakao, Keiho Imanishi, Nobuhiro Ueda, Yuichiro Imai, Tadaaki Kirita, and Tetsuya Matsuda. Regularized three-dimensional generative adversarial nets for unsupervised metal artifact reduction in head and neck ct images. *IEEE Access*, 8:109453–109465, 2020. doi: 10.1109/ACCESS.2020.3002090.

Valery Naranjo, Roberto Lloréns, Mariano Alcañiz, and Fernando López-Mir. Metal artifact reduction in dental ct images using polar mathematical morphology. *Computer methods and programs in biomedicine*, 102(1):64—74, April 2011. ISSN 1872-7565. doi: 10.1016/j.cmpb.2010.11.009. URL https://doi.org/10.1016/j.cmpb.2010.11.009.

Stephen G. Nash. Newton-type minimization via the lanczos method. *SIAM Journal on Numerical Analysis*, 21(4):770–788, 1984. ISSN 00361429. URL http://www.jstor.org/stable/2157008.

J. H. Noble, R. H. Gifford, A. J. Hedley-Williams, B. M. Dawant, and R. F. Labadie. Clinical evaluation of an image-guided cochlear implant programming strategy. *Audiol Neurootol*, 19(6):400–411, 2014.

Jack H Noble, René H Gifford, Robert Frederick Labadie, and Benoit M Dawant. Statistical shape model segmentation and frequency mapping of cochlear implant stimulation targets in CT. *Medical Image Computing and Computer-Assisted Intervention*, 15(Pt 2):421–8, jan 2012. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3559125{&}tool=pmcentrez{&}rendertype=abstract.

Jack H Noble, Robert Frederick Labadie, René H Gifford, and Benoit M Dawant. Image-Guidance enables new methods for customizing cochlear implant stimulation strategies. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(5):820–829, 2013.

S. Ourselin, A. Roche, S. Prima, and N. Ayache. Block matching: A general framework to improve robustness of rigid registration of medical images. In *MICCAI*, 2000a.

Sébastien Ourselin, A Roche, S Prima, and Nicholas Ayache. Block Matching : A General Framework to Improve Robustness of Rigid Registration of Medical

Images. *Medical Image Computing and Computer-Assisted Intervention*, pages 557–566, 2000b.

George Papamakarios, Eric T. Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *ArXiv*, abs/1912.02762, 2019.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, Oct 2000. ISSN 1573-1375. doi: 10.1023/A:1008981510081. URL https://doi.org/10.1023/A:1008981510081.

Kilian M Pohl, John Fisher, W Eric L Grimson, Ron Kikinis, and William M Wells. A Bayesian model for joint segmentation and registration. *NeuroImage*, 31(1):228–239, 2006a. ISSN 10538119. doi: 10.1016/j.neuroimage.2005.11.044.

Kilian M Pohl, John Fisher, Martha E Shenton, Robert W Mccarley, W Eric L Grimson, Ron Kikinis, William M Wells, and W L Eric. Logarithm Odds Maps for Shape Representation. *MICCAI*, 9(Pt 2):955–963, 2006b.

Vaclav Potesil, Timor Kadir, Günther Platsch, and Michael Brady. Personalization of pictorial structures for anatomical landmark localization. In Gábor Székely and Horst K. Hahn, editors, *Information Processing in Medical Imaging*, pages 333–345, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-22092-0.

Raphael Prevost, Remi Cuingnet, Benoit Mory, Laurent D. Cohen, and Roberto Ardon. *Incorporating Shape Variability in Image Segmentation via Implicit Template Deformation*, pages 82–89. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

C. Quammen, C. Weigle, and R. Taylor. Boolean operations on surfaces in vtk without external libraries. *VTK journal*, 05 2011.

Hussain Raabid, Lalande Alain, Girum Kibrom Berihu, and Caroline Guigou. Automatic segmentation of inner ear on ct-scan using auto-context convolutional

neural network. *Scientific Reports*, 1:839 – 851, 2021. ISSN 2045-2322. doi: https://doi.org/10.1038/s41598-021-83955-x.

R. Ranganath, Sean Gerrish, and D. Blei. Black box variational inference. In *AISTATS*, 2014.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/rezende15.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015a. ISBN 978-3-319-24573-7. doi: 10.1007/978-3-319-24574-4_28.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015b. Springer International Publishing. ISBN 978-3-319-24574-4.

A. Roosing, O. Strickson, and Nikolaos Nikiforakis. Fast distance fields for fluid dynamics mesh generation on graphics hardware. 03 2019.

D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8):1014–1025, Aug 2003a. ISSN 0278-0062. doi: 10.1109/TMI.2003.815865.

D. Rueckert, A.F. Frangi, and J.A. Schnabel. Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8):1014–1025, 2003b. doi: 10.1109/TMI.2003.815865.

Esmeralda Ruiz Pujadas, Hans Martin Kjer, Gemma Piella, Mario Ceresa, and Miguel Angel González Ballester. Random walks with shape prior for cochlea segmentation in ex vivo μCT. *International Journal of Computer Assisted Radiology and Surgery*, 11(9):1647–1659, 2016a. ISSN 18616429. doi: 10.1007/s11548-016-1365-8.

Esmeralda Ruiz Pujadas, Hans Martin Kjer, Sergio Vera, Mario Ceresa, and Miguel Angel González Ballester. Cochlea segmentation using iterated random walks with shape prior, 2016b. URL http://dx.doi.org/10.1117/12.2208675.

Esmeralda Ruiz Pujadas, Gemma Piella, Hans Martin Kjer, and Miguel Angel González Ballester. Random walks with statistical shape prior for cochlea and inner ear segmentation in micro-ct images. *Machine Vision and Applications*, 29(3):405–414, 2018. doi: 10.1007/s00138-017-0891-x. URL https://app.dimensions.ai/details/publication/pub.1093061695andhttps://backend.orbit.dtu.dk/ws/files/140598803/10.1007_2Fs00138_017_0891_x.pdf.

Mert Sabuncu, B.T. Thomas Yeo, Koen Van Leemput, Bruce Fischl, and Polina Golland. A generative model for image segmentation based on label fusion. *IEEE transactions on medical imaging*, 29:1714–29, 10 2010. doi: 10.1109/TMI.2010.2050897.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1218–1226, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/salimans15.html.

Irina Sanchez and Veronica Vilaplana. Brain mri super-resolution using 3d generative adversarial networks, 2018.

Tomovski Z. Sandev T. Generalized langevin equation. *Fractional Equations and Models. Developments in Mathematics*, 61, 2019. URL https://doi.org/10.1007/978-3-030-29614-8_6.

Stefan Schmidt, Jörg Kappes, Martin Bergtholdt, Vladimir Pekar, Sebastian Dries, Daniel Bystrov, and Christoph Schnörr. Spine detection and labeling using a parts-based graphical model. In Nico Karssemeijer and Boudewijn Lelieveldt, editors, *Information Processing in Medical Imaging*, pages 122–133, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-73273-0.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. 2017.

Ivor J.A. Simpson, Julia A. Schnabel, Adrian R. Groves, Jesper L.R. Andersson, and Mark W. Woolrich. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 59(3):2438–2451, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.09.002. URL https://www.sciencedirect.com/science/article/pii/S105381191101041X.

Jamshid Sourati, Ali Gholipour, Jennifer G. Dy, Xavier Tomas-Fernandez, Sila Kurugol, and Simon K. Warfield. Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE Transactions on Medical Imaging*, 38(11):2642–2653, 2019. doi: 10.1109/TMI.2019.2907805.

Andrew M. Stuart, Jochen Voss, and Petter Wilberg. Conditional path sampling of sdes and the langevin mcmc method. *Commun. Math. Sci.*, 2(4):685–697, 12 2004. URL https://projecteuclid.org:443/euclid.cms/1109885503.

Liu Y. Fels S. Rohling R.N. Abolmaesumi P. Suzani A., Seitel A. Fast automatic vertebrae detection and localization in pathological ct scans - a deep learning approach. In *In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, volume 9351, 2015.

Ingo *et al.* Todt. Evaluation of cochlear implant electrode position after a modified round window insertion by means of a 64-multislice ct. *Acta oto-laryngologica*, 129(9):966—970, September 2009. ISSN 0001-6489. doi: 10.1080/00016480802495388. URL https://doi.org/10.1080/00016480802495388.

Nicolas Toussaint, Jean-Christophe Souplet, and Pierre Fillard. MedINRIA: Medical Image Navigation and Research Tool by INRIA. In *Proc. of MICCAI'07 Workshop on Interaction in medical image analysis and visualization*, Brisbane, Australia, Australia, 2007. URL https://hal.inria.fr/inria-00616047.

Andy Tsai, Anthony Yezzi, WilliamM Wells III, ClareM Tempany, Dewey Tucker, Ayres Fan, WEricL Grimson, and AlanS Willsky. A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Transaction in Medical Imaging*, 22(2):137–54, 02 2003.

Richard Tsai and Stanley Osher. Level set methods and their applications in

image science. *Communications in mathematical sciences*, 1, 12 2003. doi: 10.4310/CMS.2003.v1.n4.a1.

K. Van Leemput. Encoding probabilistic brain atlases using bayesian inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837, June 2009. ISSN 1558-254X. doi: 10.1109/TMI.2008.2010434.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

J. M. Verburg and J. Seco. Ct metal artifact reduction method correcting for beam hardening and missing projections. In *Phys Med Biol*, 2012.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. In Nicholas Ayache, Sébastien Ourselin, and Anthony Maeder, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*, pages 319–326, Berlin, Heidelberg, 2007a. Springer Berlin Heidelberg. ISBN 978-3-540-75759-7.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic Demons Using ITK's Finite Difference Solver Hierarchy. In *Insight Journal – ISC/NA-MIC Workshop on Open Science at MICCAI 2007*, no address, Australia, 2007b. URL https://hal.inria.fr/inria-00616035. Source code available online.

Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21665–21674. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f6a8dd1c954c8506aadc764cc32b895e-Paper.pdf.

Jian Wang, William Wells, Polina Golland, and Miaomiao Zhang. Efficient

laplace approximation for bayesian registration uncertainty quantification. In *21st International Conference on Med Image Comput Comput Assist Interv.(MICCAI 2018)*, volume 11070, pages 880–888, Granada, Spain, 09 2018. ISBN 978-3-030-00927-4. doi: 10.1007/978-3-030-00928-1_99.

Jianing Wang, Jack H. Noble, and Benoit M. Dawant. 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, 2019a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101553. URL http://www.sciencedirect.com/science/article/pii/S1361841519300969.

Jianing Wang, Jack H. Noble, and Benoit M. Dawant. 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, 2019b. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101553. URL http://www.sciencedirect.com/science/article/pii/S1361841519300969.

Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal GAN for video generation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1149–1158, 2020a. doi: 10.1109/WACV45572.2020.9093492.

Z. Wang and *et al.* A Deep Learning based Fast Signed Distance Map Generation. In *Medical Imaging with Deep Learning*, Montréal, Canada, July 2020. URL https://hal.inria.fr/hal-02570026.

Zihao Wang and Hervé Delingette. Attention for Image Registration (AiR): A Transformer Approach. working paper or preprint, April 2021a.

Zihao Wang and Hervé Delingette. Quasi-symplectic langevin variational autoencoder, 2021b.

Zihao Wang, Clair Vandersteen, Thomas Demarcy, Dan Gnansia, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. Deep learning based metal artifacts reduction in post-operative cochlear implant ct imaging. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 121–129, Cham, 2019c. Springer International Publishing. ISBN 978-3-030-32226-7.

Zihao Wang, Clair Vandersteen, Thomas Demarcy, Dan Gnansia, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. Deep learning based metal artifacts reduction in post-operative cochlear implant ct imaging. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 121–129, Cham, 2019d. Springer International Publishing. ISBN 978-3-030-32226-7.

Zihao Wang, Clair Vandersteen, Thomas Demarcy, Dan Gnansia, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. Deep learning based metal artifacts reduction in post-operative cochlear implant ct imaging. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 121–129, Cham, 2019e. Springer International Publishing. ISBN 978-3-030-32226-7.

Zihao Wang, Clair Vandersteen, Thomas Demarcy, Dan Gnansia, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. A Deep Learning based Fast Signed Distance Map Generation. In *Medical Imaging with Deep Learning*, Montréal, Canada, July 2020b.

Zihao Wang, Clair Vandersteen, Thomas Demarcy, Dan Gnansia, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. A Deep Learning based Fast Signed Distance Map Generation. In *MIDL 2020 - Medical Imaging with Deep Learning*, Montréal, Canada, July 2020c. URL https://hal.inria.fr/hal-02570026.

Zihao Wang, Clair Vandersteen, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. One-shot Learning Landmarks Detection. working paper or preprint, November 2020d. URL https://hal.inria.fr/hal-03024759.

Zihao Wang, Clair Vandersteen, Charles Raffaelli, Nicolas Guevara, and Hervé Delingette. One-shot Learning Landmarks Detection. working paper or preprint, November 2020e. URL https://hal.inria.fr/hal-03024759.

Zihao Wang, Zhifei Xu, Jiayi He, Chulsoon Hwang, Jun Fan, and Hervé Delingette. Long short-term memory neuron equalizer, 2020f.

Zihao Wang, Thomas Demarcy, Clair Vandersteen, Dan Gnansia, Charles Raffaelli,

Nicolas Guevara, and Hervé Delingette. Bayesian logistic shape model inference: application to cochlea image segmentation, 2021a.

Zihao Wang, Clair Vandersteen, Thomas Demarcy, Dan Gnansia, Charles Raffaelli, Nicolas Guevara, and Herve Delingette. Inner-ear augmented metal artifact reduction with simulation-based 3d generative adversarial networks, 2021b.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.

Wikipedia contributors. Timeline of machine learning — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Timeline_of_machine_learning&oldid=1012627434, 2021. [Online; accessed 29-March-2021].

Wilhelm Wimmer, Lukas Anschuetz, Stefan Weder, Franca Wagner, Hervé Delingette, and Marco Caversaccio. Human bony labyrinth dataset: Co-registered ct and micro-ct images, surface models and anatomical landmarks. *Data in Brief*, 27:104782, 2019. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2019.104782. URL https://www.sciencedirect.com/science/article/pii/S2352340919311370.

Christopher Wolf, Maximilian Karl, and Patrick van der Smagt. Variational inference with hamiltonian monte carlo, 2016.

D. Wu, F. Zhu, and L. Shao. One shot learning gesture recognition from rgbd images. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–12, 2012.

Jianhua Wu and Leif Kobbelt. Piecewise linear approximation of signed distance fields. In *Vision, Modeling and Visualization*, pages 513–520, 09 2003.

Yizi Wu, Jiaju Man, and Ziqing Xie. A double layer method for constructing signed distance fields from triangle meshes. *Graphical Models*, 76(4):214 – 223, 2014. ISSN 1524-0703.

Stefan Wörz and Karl Rohr. Localization of anatomical point landmarks in 3d medical images by fitting 3d parametric intensity models. *Medical Image Analysis*, 10(1):41 – 58, 2006. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2005.02.003.

Zhifei Xu, Zihao Wang, Yin Sun, Chulsoon Hwang, Hervé Delingette, and Jun Fan. Jitter Aware Economic PDN Optimization with a Genetic Algorithm. working paper or preprint, May 2021. URL https://hal.archives-ouvertes.fr/hal-03219748.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021.

Devira Zahara, Rima Diana Dewi, Askaroellah Aboet, Fikri Mirza Putranto, Netty Delvrita Lubis, and Taufik Ashar. Variations in Cochlear Size of Cochlear Implant Candidates. *International Archives of Otorhinolaryngology*, 23:184 – 190, 06 2019. ISSN 1809-4864.

Irene *et al.* Zanette. Trimodal low-dose x-ray tomography. *Proceedings of the National Academy of Sciences of the United States of America*, 109:10199–204, 06 2012. doi: 10.1073/pnas.1117861109.

J. Zhang, M. Liu, and D. Shen. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Transactions on Image Processing*, 26(10):4753–4764, 2017.

Ruibo Zhang, Yali Huang, and Zhao Zhen. A ultrasound liver image enhancement algorithm based on multi-scale retinex theory. In *Bioinformatics and Biomedical Engineering, (iCBBE)*, pages 1–3, 05 2011. doi: 10.1109/icbbe.2011.5780462.

Y. Zhang and H. Yu. Convolutional neural network based metal artifact reduction in x-ray computed tomography. *IEEE Transactions on Medical Imaging*, 37(6): 1370–1381, 2018.

J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

I. George Zubal, Charles R. Harrell, Eileen O. Smith, Zachary Rattner, Gene Gindi, and Paul B. Hoffer. Computerized three-dimensional segmented human anatomy. *Medical Physics*, 21(2):299–302, 1994. doi: https://doi.org/10.1118/1.597290. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.597290.