Deep Learning Techniques to Bridge the Gap between 2D and 3D Ultrasound Imaging

Pak Hei Yeung

Pembroke College
University of Oxford

Submitted in partial fulfilment of the degree of Doctor of Philosophy

Trinity 2022

Abstract

Three-dimensional (3D) ultrasound imaging has contributed to our understanding of fetal developmental processes in the womb by providing rich contextual information of the inherently 3D anatomies. However, its use is limited in clinical settings, due to the high purchasing costs and limited diagnostic practicality. Freehand two-dimensional (2D) ultrasound imaging, in contrast, is routinely used in standard obstetric exams. The low cost and portability of 2D ultrasound render it uniquely suitable for use in low- and middle-income settings. However, high level of expertise is always involved and it inherently lacks a 3D representation of the anatomies, which limit its potential for more accessible and advanced assessment. Capitalizing on the flexibility offered by freehand 2D ultrasound acquisition, this thesis presents a deep learning-based framework for optimizing the utilization and diagnostic power of 2D freehand ultrasound in fetal brain imaging.

First, a localization model is presented to predict the location of 2D ultrasound fetal brain scans in the 3D brain atlas. It is trained by sampling 2D slices from aligned 3D fetal brain volumes, such that heavy annotations for each 2D scan are not required. This can be used for scanning guidance and standard plane localization.

An unsupervised methodology is further proposed to adapt a trained localization model to freehand 2D ultrasound images acquired from arbitrary domains, for example sonographers, manufacturers and acquisition protocols. This enables the model to be used at the bedside in practice, where it can be fine-tuned with just the images acquired in any arbitrary domains before inference.

Building upon the ability to localize 2D scans in the 3D brain atlas, a framework is further presented to reconstruct 3D volumes from non-sensor-tracked 2D ultrasound images using implicit representation. With this slice-to-volume reconstruction framework, additional 3D information can be extracted from the 2D freehand scans.

Finally, a semi-automatic model, trained only on raw 3D volumes without any manual annotation, is presented to segment any arbitrary structures of interest in 3D medical volumes, while only requiring manual annotation of a single slice during inference. The model is tested on wide variety of medical imaging datasets and anatomical structures, verifying its generalizability.

In the design of the framework presented in this thesis, three fundamental principles, namely minimal human annotation, generalizability and sensorless operation, are followed to optimize its seamless integration into the clinical workflow. This may modernize freehand routine scanning and enhance its accessibility, while maximizing the clinical information gained from routine scans acquired as part of the continuum of pregnancy care.

Deep Learning Techniques to Bridge the Gap between 2D and 3D Ultrasound Imaging

Pak Hei Yeung

Pembroke College

University of Oxford

Submitted in partial fulfilment of the degree of

Doctor of Philosophy

Trinity 2022

Acknowledgements

Personal

First of all, my greatest and sincerest thanks must be given to my supervisors, Prof. Ana Namburete and Dr. Weidi Xie, for their support and guidance throughout my DPhil in Oxford. Thank you for always supporting me and providing me with different opportunities to become a better person and researcher. I am privileged to stand on your shoulders, which allows me to see a completely new world in the last four years. My gratitude also goes to my collaborators, Dr. Monique Haak and Ms Moska Aliasi, for the clinical input and insight, and Prof. Alison Noble, for your precious comments and support.

I am very fortunate to be part of the amazing OMNI Lab. I would like to thank Nicola, Maddy and Linde for your supportive comments, feedback and corrections for my papers and thesis. I wish to put all your names on the first page of this thesis, if I am allowed to. Thank you, Felipe, for all your technical help and, more importantly, the countless after-work discussions and brainstorming of new ideas, which are definitely some of the most inspiring conversations I have during my DPhil. OMNI is the best group and my sincerest gratitude goes to everyone in the group.

It is one of the luckiest things in my DPhil to know my housemates and dearest friends, Joe, Bestin and Jose. Thank you, Joe and Bestin, for all the amazing food, drinks and games you prepared and organized in the house. Thank you, Jose and Juliana, for literally everything, including all your support and encouragement. It is always the best moment of the day when we are having dinner and a chat after work. Because of you, the house becomes home. I have met a lot of amazing people in Oxford, who have helped and guided me through my DPhil. I would like to express my sincere gratitude to every one of you.

Finally, and most importantly, I would like to thank my parents and family, who always support and believe in me unconditionally. This work is an important milestone of my life and I would like to dedicate it to you.

Funding

I am deeply grateful to the R C Lee Centenary Scholarship Commission for your generosity, which makes it possible for me to study in this amazing university for my DPhil. In particular, I would like to thank Dr Deanna Lee Rudgard and Mrs Clara Lee for all your support and encouragement.

Abstract

Three-dimensional (3D) ultrasound imaging has contributed to our understanding of fetal developmental processes in the womb by providing rich contextual information of the inherently 3D anatomies. However, its use is limited in clinical settings, due to the high purchasing costs and limited diagnostic practicality. Freehand two-dimensional (2D) ultrasound imaging, in contrast, is routinely used in standard obstetric exams. The low cost and portability of 2D ultrasound render it uniquely suitable for use in low- and middle-income settings. However, high level of expertise is always involved and it inherently lacks a 3D representation of the anatomies, which limit its potential for more accessible and advanced assessment. Capitalizing on the flexibility offered by freehand 2D ultrasound acquisition, this thesis presents a deep learning-based framework for optimizing the utilization and diagnostic power of 2D freehand ultrasound in fetal brain imaging.

First, a localization model is presented to predict the location of 2D ultrasound fetal brain scans in the 3D brain atlas. It is trained by sampling 2D slices from aligned 3D fetal brain volumes, such that heavy annotations for each 2D scan are not required. This can be used for scanning guidance and standard plane localization.

An unsupervised methodology is further proposed to adapt a trained localization model to freehand 2D ultrasound images acquired from arbitrary domains, for example sonographers, manufacturers and acquisition protocols. This enables the model to be used at the bedside in practice, where it can be fine-tuned with just the images acquired in any arbitrary domains before inference.

Building upon the ability to localize 2D scans in the 3D brain atlas, a framework is further presented to reconstruct 3D volumes from non-sensor-tracked 2D ultrasound images using implicit representation. With this slice-to-volume reconstruction framework, additional 3D information can be extracted from the 2D freehand scans.

Finally, a semi-automatic model, trained only on raw 3D volumes without any manual annotation, is presented to segment any arbitrary structures of interest in 3D medical volumes, while only requiring manual annotation of a single slice during inference. The model is tested on wide variety of medical imaging datasets and anatomical structures, verifying its generalizability.

In the design of the framework presented in this thesis, three fundamental principles, namely minimal human annotation, generalizability and sensorless

operation, are followed to optimize its seamless integration into the clinical workflow. This may modernize freehand routine scanning and enhance its accessibility, while maximizing the clinical information gained from routine scans acquired as part of the continuum of pregnancy care.

Contents

List of Figures

List of Tables

List of Abbreviations

2D, 3D . . . . Two- or three-dimensional

ANOVA . . . . Analysis of variance

BPD . . . . . Biparietal diameter

CNS . . . . . Central nervous system

CSP . . . . . Cavum septi pellucidi

CT . . . . . Computerised tomography

ConvNet . . . Convolutional Neural Network

GAN . . . . . Generative adversarial networks

G . . . . . Generalization

HC . . . . . Head circumference

LMIC . . . . . Low- and middle-income countries

MAE . . . . . Mean absolute error

MSE . . . . . Mean least-squared error

MHA . . . . . Minimal human annotation

MLP . . . . . Multi-layer perceptron

MRI . . . . . Magnetic resonance imaging

NCC . . . . . Normalized cross-correlation

RL . . . . . Reinforcement learning

SO . . . . . Sensorless operation

SOI . . . . . Structure of interest

SSIM . . . . . Structural similarity index measure

TCD . . . . . Transcerebellar diameter

TT . . . . . Transventricular

1 Introduction

Contents

1.1 Motivation

Central nervous system (CNS) abnormalities are one of the most common classes of congenital diseases (*i.e.* diseases that are present from birth) [1]. For example, in the UK, the prevalence of neural tube defects is 1-2 per 1000 births [2]. Many of the abnormalities can be diagnosed early by medical imaging, making it essential for monitoring fetal growth. Relying on pulses of high-frequency sound waves, two-dimensional (2D) ultrasound imaging is safe to use even for the fetuses, who are young and vulnerable. Together with its cost-effectiveness, real-time acquisition capabilities and portability, freehand 2D ultrasound is routinely used for monitoring fetal growth and assessing fetal anatomy. In the UK, at least two ultrasound scans are recommended during pregnancy, normally taking around half an hour per session [3]. In the scanning session, as illustrated in Fig. 1.1, the sonographer needs to

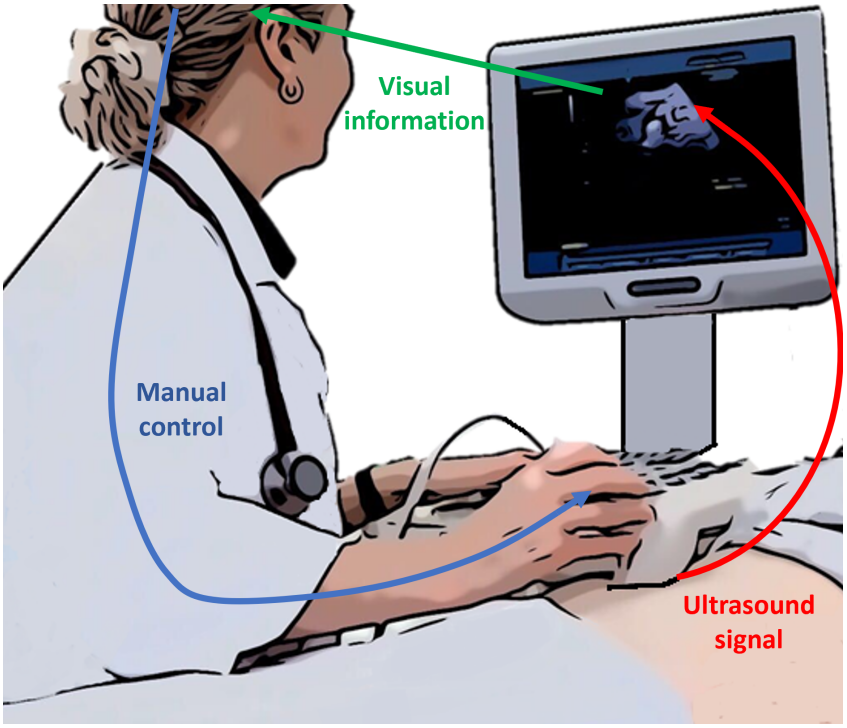


Figure 1.1: Illustration of freehand ultrasound scanning. The sonographer manually adjusts the position of the ultrasound probe (blue arrow) based on the feedback from the 2D scans (green arrow) displayed on the screen in real time (red arrow).

manually navigate the ultrasound probe. The screen will display the corresponding 2D scans in real time and he or she needs to identify different anatomic landmarks of the fetal brain from the planes and then use these landmarks to determine the standard planes of view [4]. Different biometric measurements are finally collected from these views to monitor the fetal brain development. The detailed protocol for fetal brain scanning is reviewed in Chapter 2.1.

Two-dimensional ultrasound is the preferred choice for fetal brain scanning in clinical practice. However, more sophisticated diagnosis still relies on magnetic resonance imaging (MRI), which is more expensive, resource-demanding and not suitable for all pregnancies, and hence limiting its utilization and accessibility in resources-constrained settings. Although some modern ultrasound equipment, such as 3D and 4D machines, may also provide more detailed diagnostic information, when compared to the basic 2D ultrasound, they are not usually available, especially in the low- and middle-income countries (LMIC) [5]. The goal of this thesis, therefore, is to capitalize on the flexibility offered by freehand 2D ultrasound and maximize its

diagnostic potential, which may benefit the obstetric services in different settings. In order to achieve that, several factors that limit the potential and use of 2D ultrasound need to be first identified:

(i) High level of expertise is always involved in the scanning, which relies on in-depth understanding of fetal anatomy and experience in ultrasound imaging. This requires a significant amount of training and shortage of adequately skilled personnel may form a barrier for its use [6].

(ii) Each 2D ultrasound image is limited to representing a cross-sectional view of the 3D anatomy, which inherently fails to capture rich contextual volumetric information.

Addressing these challenges is not trivial due to the unique properties of 2D ultrasound images, when compared to images acquired from other medical imaging modalities. For example, computerised tomography (CT), by design, captures the anatomy in 3D during scanning, while other 2D imaging modality, such as X-ray, captures a projectional view, but not a cross-sectional view, of the 3D body.

In order to address the aforementioned limitations and optimize the potential and utilization of 2D ultrasound for fetal brain imaging, this thesis presents several deep learning-based components, comprising of an end-to-end framework to assist the acquisition of 2D ultrasound fetal brain images and the extraction of volumetric and structural information from them. The proposed framework is able to localize 2D ultrasound images in the 3D brain atlas, which may assist the freehand navigation during the scanning. With the localized images, the proposed framework further reconstructs a 3D volume, from which different 3D structures and regions can be segmented semi-automatically.

To maximize the impact of the framework and its utilization at the bedside, three fundamental principles are followed:

Minimal Human Annotation (MHA). In the era of deep learning, big data is critical to the performance of the models. Specifically, training a deep neural

network requires a lot of data and their corresponding manual annotations for the task of interest, for example ImageNet [7]. However, in the medical imaging field, obtaining data manually annotated by domain experts is costly and challenging. Therefore, whenever possible, the training (and inference) pipelines of the work presented in this thesis are formulated as unsupervised or self-supervised [8] learning problems, substantially reducing the burden on expert annotators to manually produce labels, when compared to conventional supervised learning.

Generalizability (G). In practice, medical images are acquired from different scanners (*i.e.* cross-vendor variability) with different acquisition protocols. Together with other sources of variation, such as patients' demographics, the trained models usually fail to generalize to the diverse distributions, due to the limited coverage of the training data, and, hence, can suffer a catastrophic drop in performance when applied to data from a different domain. Therefore, this thesis focuses on the generalizability of the proposed models, to ensure that they have the potential to be utilized in practice, beyond medical research applications.

Sensorless Operation (SO). The major advantages of 2D ultrasound are its cost-effectiveness and portability. In order to retain these strengths, the proposed framework only requires image information, and does not need any positional information from an external sensor. This would facilitate their direct application to most of the existing 2D scanning devices, which are available in most obstetric care units, without modification. Specifically, the proposed framework should be genuinely software-based, where addition of external sensor is not prohibited and may contribute to further improvement. Nevertheless only image information is already sufficient to achieve satisfactory performance.

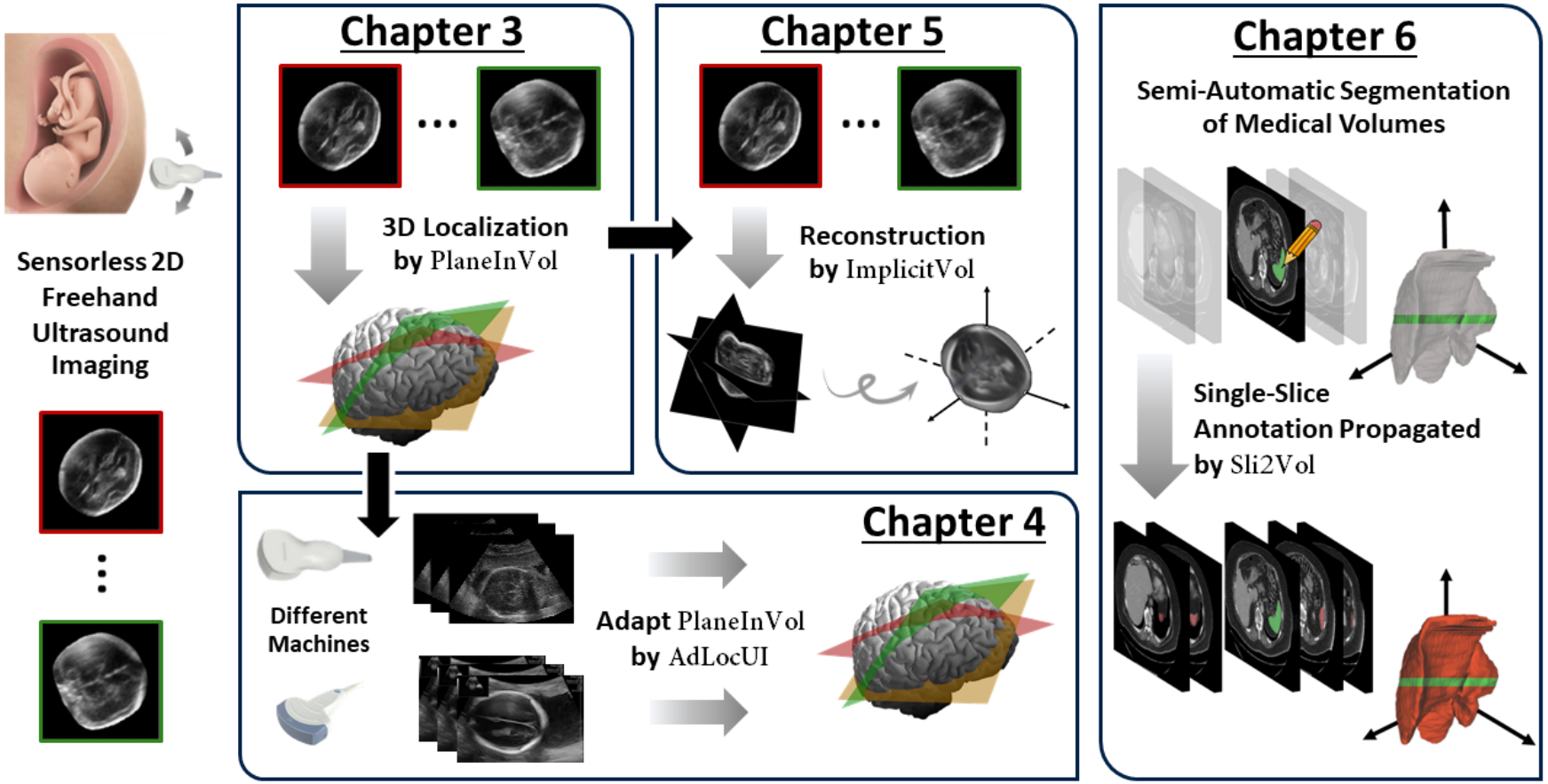


Figure 1.2: Schematic summarizing the framework presented in this thesis. With a sequence of 2D freehand ultrasound fetal brain scans, their respective locations in the 3D brain atlas are predicted by `PlaneInVol` (Chapter 3). A trained `PlaneInVol` model can be adapted to freehand 2D ultrasound images acquired from arbitrary machines with `AdLocUI` (Chapter 4). The localized sequence of 2D freehand ultrasound fetal brain scans can then be used to reconstruct a 3D volume by `ImplicitVol` (Chapter 5). Finally, a general semi-automatic segmentation framework, `Sli2Vol`, can be used to segment any arbitrary structures of interest (SOIs) in 3D medical volumes while only requiring manual annotation of a single slice during inference (Chapter 6).

1.2 Thesis Contribution

The main contribution of this thesis is a deep learning-based framework for optimizing the utilization and diagnostic power of 2D freehand ultrasound in fetal brain imaging. The framework, as summarized in Fig. 1.2 is composed of four components, namely plane localization (Chapter 3), unsupervised domain adaptation (Chapter 4), slice-to-volume reconstruction (Chapter 5) and semi-automatic segmentation (Chapter 6). This section outlines each of their major contributions and overviews their mutual relationships and dependencies on their roles in the framework. The contributions will also be linked to the three aforementioned fundamental principles (Section 1.1), as superscript$^{\mathbf{MHA}, \mathbf{G}, \mathbf{SO}}$.

Chapter 3 - 3D Localization of 2D Images. As the first component of the framework, `PlaneInVol` is proposed to predict the plane location of 2D ultrasound fetal brain scans in a 3D brain atlas. `PlaneInVol` can be used for scanning guidance and standard plane localization. It provides the following novel contributions: Firstly, `PlaneInVol` is genuinely sensorless$^{\mathbf{SO}}$ (*i.e.* in both training and inference stages), which is particularly important for fetal imaging. Fetal motion is independent and uncorrelated with the freehand positioning of the probe. A tracking sensor, therefore, can only record the probe position but not the plane position due to the relative motion between the fetus and the probe. Many related approaches (reviewed in Chapter 2.6) rely on sensor tracking in either training or inference stage, limiting their applicability in point-of-care setting. Secondly, `PlaneInVol` is trained by sampling 2D slices from aligned 3D fetal brain ultrasound volumes, resembling the idea of self-supervised learning, such that heavy annotations for each 2D scan are not required$^{\mathbf{MHA}}$.

Chapter 4 - Adaptive Localization of 2D Images. Since `PlaneInVol` is trained with 2D slices sampled from 3D ultrasound volumes, there is no guarantee that the trained model will have equally good performance on native freehand 2D ultrasound images, which can be acquired from machines from different manufacturers and with

different acquisition protocols. In this chapter, this limitation will be addressed by the proposed `AdLocUI`, a framework that **Ad**aptively **Loc**alizes 2D **U**ltrasound **I**mages in the 3D brain atlas. Specifically, a novel domain adaptation methodology is presented to adapt a trained localization model (*e.g.* `PlaneInVol`) to freehand 2D ultrasound images acquired from arbitrary machines[G]. Since the domains (*e.g.* sonographers, manufacturers and acquisition protocols) can have many different variations and it is impossible to collect manual annotation for each of them, an *unsupervised* domain adaptation framework[MHA] is proposed, where the trained network can be fine-tuned by `AdLocUI` with just the images of the target domain before inference[G].

Chapter 5 - Volumetric Reconstruction from 2D Images. In this chapter, `ImplicitVol` is proposed to reconstruct 3D volumes from non-sensor-tracked 2D ultrasound images. In Chapter 3, `PlaneInVol` is presented to predict the location of 2D ultrasound fetal brain scans in the 3D brain atlas, which is used by `ImplicitVol` for images' location initialization. `ImplicitVol` will then jointly refine the images' locations and learn the volumetric reconstruction. Most prior works focused on volumetric reconstruction from 2D ultrasound images rely on sensor tracking in either training or inference stages, while `ImplicitVol` is genuinely sensor-free[SO]. In addition, `ImplicitVol` is the first study that relies on *implicit representation* for the ultrasound volumetric reconstruction task.

Chapter 6 - Volumetric Segmentation from a Single Slice. Building upon the ability to reconstruct a 3D volume from 2D ultrasound fetal brain images by `ImplicitVol` as proposed in Chapter 5, a more general problem regarding 3D medical volumes will finally be investigated, namely semi-automatic segmentation. In this chapter, `Sli2Vol` is proposed to segment any *arbitrary* structure of interest (SOI) in 3D volumes while only requiring manual annotation of a *single* slice. Training `Sli2Vol` requires only raw volumes, but not any manual annotation, and during inference, any *arbitrary* SOIs in 3D volumes can be segmented by only

manually annotating a *single* slice within the volume[MHA]. It will be demonstrated that a *single* trained `Sli2Vol` model can work on wide variety of datasets and anatomical structures, without any parameter-tuning[G]. To the best of my knowledge, this is the first study to undertake cross-domain evaluation on such large-scale and diverse benchmarks for semi-automatic segmentation approaches, which shifts the focus to *generalizability* across different devices, clinical sites and anatomical SOIs.

1.3 Thesis Structure

There are seven chapters in this thesis. In this chapter (Chapter 1), the motivation, contribution and structure of this thesis are introduced. Chapter 2 includes the literature review of the relevant topics. The contribution chapters (Chapter 3-6) present the components of the proposed framework for optimizing the utilization and diagnostic power of 2D freehand ultrasound in fetal brain imaging. Finally, the works are concluded in Chapter 7 and the limitations and potential future works are discussed.

In each contribution chapter (Chapter 3-6), the motivation and contribution of the proposed model will first be **introduced**, followed by the technical details of the **methods**. The **experimental details**, including the dataset and experimental setup, are then described, followed by the **experimental results**. Finally, the chapter will be **concluded** by summarizing the findings and relating them to the overall framework proposed in this thesis.

1.4 Publications

The works presented in this thesis have been published on (or submitted to) the following conferences and journals:

Chapter 3 **Yeung, P.H.**, Aliasi, M., Papageorghiou, A.T., Haak, M., Xie, W. and Namburete, A.I.,: Learning to Map 2D Ultrasound Images into 3D Space with Minimal Human Annotation., *Medical Image Analysis*, vol. 70, pp.72-86, May 2021.

Chapter 4 **Yeung, P.H.**, Aliasi, M., Haak, M., the INTERGROWTH-21st Consortium, Xie, W. and Namburete, A.I.,: Adaptive 3D Localization of 2D Freehand Ultrasound Brain Images., *International conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022.

Chapter 5 *(under review)* **Yeung, P.H.**, Hesse, L., Aliasi, M., Haak, M., the INTERGROWTH-21st Consortium, Xie, W. and Namburete, A.I.,: Sensorless Volumetric Reconstruction of Fetal Brain Freehand Ultrasound Scans with Deep Implicit Representation., *submitted to Medical Image Analysis*.

Chapter 6 **Yeung, P.H.**, Namburete, A.I., and Xie, W,: Sli2Vol: Annotate a 3D Volume from a Single Slice with Self-Supervised Learning., *International conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

2
Literature Review

In this chapter, the relevant literature on topics that are closely related to the work presented in this thesis is reviewed. It starts with summarizing obstetric ultrasound scanning and comparison between the 2D and 3D ultrasound. After that, several related topics of deep learning are reviewed. The methods for computer-aided ultrasound scanning and slice-to-volume registration for ultrasound and other medical imaging modalities are finally discussed .

Contents

2.1 Obstetric Ultrasound Scanning

Ultrasound has been widely used in obstetric care since the 1970s [9]. Unlike X-ray and CT, ultrasound is non-ionizing and much safer, making it well-suited for fetal scanning. Also, ultrasound is generally much cheaper and portable than MRI, which makes it more suitable to be used in different settings. Normally, minimum of two standard ultrasound scanning sessions, one performed between 11 weeks to 14 weeks of gestation and the other performed between 18 weeks to 20 weeks of gestation, are needed [3]. For the first session, the nuchal translucency and the crown rump length are measured to estimate the chance of the fetus being affected by anomaly, such as Down's syndrome and Edwards' syndrome. The second session involves more detailed examination of fetal anatomy, including the head, abdomen, femur and spine, to monitor the fetal growth and assess the health condition of the fetus [10].

2.1.1 Ultrasound Scanning of Fetal Brain

A major part of the scanning session focuses on fetal brain examination, which will be discussed in detail in this section. In order to ensure that the screening results are consistent and reproducible, some organizations, such as Public Health England, have published guidelines for fetal ultrasound scanning [3]. Normally, during 18 to 20 weeks of pregnancy, a session of transabdominal ultrasound scanning is undertaken and several biometric parameters related to fetal head/brain, namely head circumference (HC), atrium of the lateral ventricle and transcerebellar diameter (TCD) are manually annotated and measured with digital callipers [3]. The standard approach is to first identify different anatomic landmarks of the fetal brain. The sonographer will then use these landmarks to determine the standard planes of

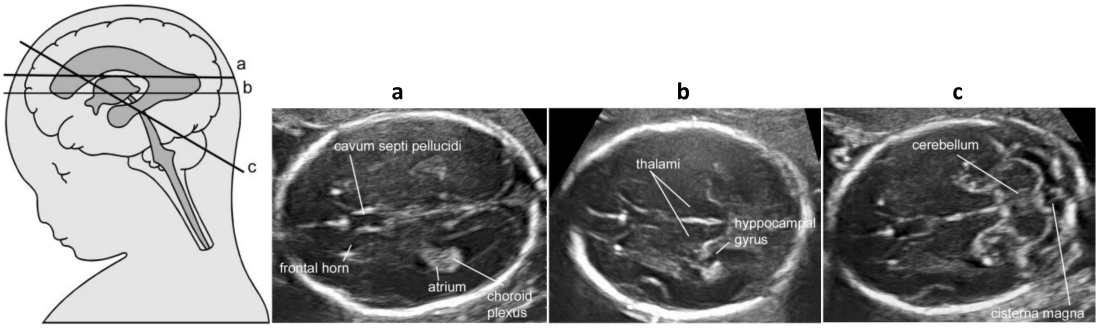


Figure 2.1: Location of the standard planes of view of the fetal brain. (a) Biparietal diameter (BPD) plane; (b) transthalamic (TT) plane; (c) transcerebellar (TC) plane. Adopted from [1].

view (Fig. 2.1), namely the transventricular (TT) plane, the transcerebellar (TC) plane and the biparietal diameter (BPD) plane [11]. The biometric measurements collected from these standard planes are correlated with brain development and potential anomalies by checking the normal range of values at a particular age on the fetal growth chart or tables [12]. HC is normally used for gestational age estimation during 13-25 weeks [12] and studies [13, 14] have suggested that TCD is also a reliable biomarker for such task. Besides, Vinkesteijn et al. [15] suggested that reduced TCD was observed in fetuses with chromosomal abnormalities. In addition, ventriculomegaly (*i.e.* enlargement of the lateral ventricle), which is potentially associated with variety of neuropsychiatric disorders, can be diagnosed by measuring the atrium of the lateral ventricle during ultrasound scanning [16, 17].

Besides those quantitative measurements, sonographers or clinicians may also rely on qualitative evaluation by inspecting different anatomic structures, such as the lateral ventricles, cavum septi pellucidi (CSP), cerebellum and cisterna magna, within the standard planes to check for anomalies [1]. For instance, the CSP can be identified in the TT plane from as early as 15 weeks of gestation and its absence or enlargement shown in the ultrasound images may indicate abnormal brain development and diseases, such as septo-optic dysplasia, holoprosencephaly and middle interhemispheric variant [18–20]. The cerebellum and cisterna magna can be visualized in the TC plane and CNS abnormalities, for example meningomyelocele and encephalocele, may be reflected by the absence or abnormal size or shape of

these two structures [21]. These are all qualitative screenings that are suitable for most pregnancy cases, Paladini et al. [1] highlighted the need for more detailed fetal neurosonography for diagnosis of complex malformations. Detailed fetal neurosonography focuses on the qualitative assessment of a set of additional planes of view, for example the transfrontal plane, the transcaudate plane and the parasagittal plane. Basically, sonographers will align the ultrasound probe to various biomarkers of the fetal head to identify those planes and then assess the anatomic structures and their characteristics, including their presence, shape, size and texture, in these planes.

Although the detailed fetal neurosonography may provide richer diagnostic information to the clinicians, the principle of the scanning is the same as the basic screenings. They both rely on (i) moving the ultrasound probe to identify biomarkers of the brain and hence the approximate location of the desired plane of view and (ii) fine-tuning the probe position to obtain the optimal view for diagnosis. In other words, finding the desired and accurate plane during scanning is an important part in clinical ultrasound fetal brain imaging. However, this process requires adequate understanding of fetal anatomy as well as knowledge and experience of ultrasound imaging, which may involve significant amount of training and may not be available in many settings. In addition, although criteria for manually locating the standard planes are broadly defined [11], determining whether a particular view is optimal involves subjective judgement and, hence, inter-operator variability is inevitable. These limitations may be overcome by developing an automated tool for localizing different planes in the 3D space, which is one of the contributions of this thesis.

2.1.2 Obstetric Ultrasound in Low- and Middle-Income Countries

The cost-effectiveness and portability of ultrasound imaging enable it to be used in low- and middle-income countries (LMIC). When compared to high-income countries, hardware maintenance and repair are more difficult, posing extra challenges for more bulky imaging devices, such as CT and MRI [22]. World Health Organization

recognized the importance and potential of ultrasound imaging in primary healthcare services by aiming to meet 90% of global needs of medical imagining with ultrasound and X-ray [23]. While X-ray is ionizing and not suitable for most obstetric needs, ultrasound imaging becomes an indispensable option. Application, such as gestational age prediction, is particularly useful in LMIC [5], especially when the last menstrual period is unknown or unreliable, which cannot be used to derive the age. Accurate gestation age prediction is important for reducing inductions for post term labor [24] and distinguishing between pre-term babies and babies with low birth weight [25].

Despite the safety and potential advantages of ultrasound screening, it is still a debatable topic for its *routine* usage in LMIC, mainly due to the resources concern [26]. Specifically, having routine obstetric ultrasound may distract resources from other healthcare services and, hence, causes disruption. For example, radiographers are needed when performing ultrasound scanning, which may disrupt their operation of X-ray scanning [27] One solution is to train lay providers, such as midwives and nurses, for using ultrasound. Numerous studies [25, 28, 29] have suggested that those health providers are capable to deliver obstetric ultrasound scanning after intense training. Nevertheless, Henwood et al. [30] emphasized that continual education and training should be followed, which can be resource-demanding in LMIC. Remote learning is, therefore, a potential means to reduce the cost of educational programs and training supervision [31]. Telecommunications may also facilitate teleconsultation and remote ultrasound imaging, which may further ease the shortage of human resources in LMIC [22, 32].

One of the potential contributions of this thesis is to localize 2D ultrasound images in a 3D fetal brain atlas. It can be easily developed as an interactive training framework to provide guidance and feedback to lay providers when they are performing freehand ultrasound scanning. The other contribution, namely volumetric reconstruction from 2D ultrasound images, may also facilitate offline and secondary examination [33, 34], which is suitable for teleconsultation.

2.2 Comparison between 2D and 3D Scans

While the scanning protocols introduced in Section 2.1 are all 2D-based, 3D ultrasound imaging has emerged in some commercial ultrasound system, which makes 3D obstetric scanning possible. In this section, the current clinical applications, limitations and potential of these two modes of ultrasound imaging are compared.

2.2.1 Current Clinical Application for Fetal Brain Imaging

Most of the current standard clinical ultrasound tests rely on 2D imaging [3, 10]. As reviewed in Section 2.1, 2D fetal ultrasound is used for measuring biometric parameters, such as the HC and TCD, as well as facilitating qualitative assessment of different cross-sectional views of the fetal brain. On the other hand, although 3D ultrasound images of the head may be collected in some of the scanning sessions, they are mainly used for non-diagnostic purposes, including visualization of fetus' appearance for psychological purposes, such as improving maternal-fetal bonding [35]. Some studies [36–38] suggested that clinically, 3D ultrasound of the fetal head may also identify fetal facial abnormalities, for example cleft lip and palate. Nevertheless, the application of 3D ultrasound for fetal brain imaging is still in the research stage [39, 40] and it is not widely accepted for clinical and diagnostic purposes.

2.2.2 Limitations and Potential

For 2D ultrasound scanning, a substantial amount of time is spent on finding the standard planes and desired views in 3D space for evaluation and biometrics measurement. Pistorius et al. [41] pointed out that scanning and assessment time could be significantly reduced by $\frac{2}{3}$ by using 3D ultrasound. Even though analysis of 3D ultrasound images may take longer time, the overall scanning and analysis is still faster for 3D ultrasound. This is an important consideration when ultrasound is used in busy hospitals and clinics with high patient throughput [42].

(a) Butterfly iQ **(b)** GE Voluson E8

Figure 2.2: Examples of some commercial ultrasound systems. (a) Two-dimensional ultrasound images can be acquired by portable ultrasound probe and analysed on a phone. (b) Acquiring and analysing 3D ultrasound images usually require bulkier and more complicated hardware system.

This may also be useful in LMIC, as introduced in Section 2.1.2, where different resources, including time, are limited.

During 2D ultrasound scanning, clinicians can only save a limited number of fetal brain images. On the other hand, 3D ultrasound allows them to store the whole brain volume and navigate through any random planes of view for offline analysis [33]. There are several advantages of having access to 3D images. Firstly, clinicians can re-interpret the image at any time and view a structure at different orientations by navigating through oblique planes. Dückelmann et al. [43] suggested that this allowed the clinicians to look at different views in different modes, such as thick slice 3D rendering, when necessary, and may provide much richer information for analysing the brain structures. Secondly, compared to 2D images, saved 3D images are easier to be used for secondary examination by a clinical colleague not involving in the scanning, which may lead to more objective and reliable diagnosis [33, 34].

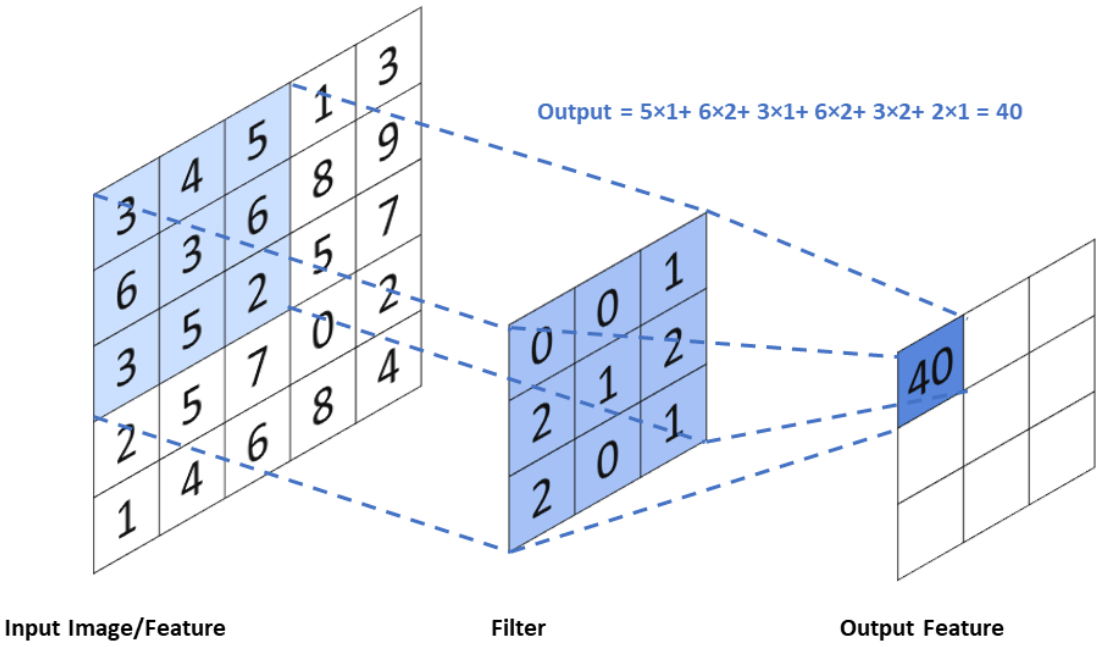
There are also different studies suggesting the diagnostic potential of 3D ultrasound. Merz et al. [39] found that 3D ultrasound was a more useful tool than

2D ultrasound in detecting and evaluating the severity of fetal defects, such as agenesis of the corpus callosum. 3D ultrasound can also facilitate extraction of volumetric measurements, which may be used for growth analysis and assessing abnormalities [40]. Vinals et al. [44] used 3D ultrasound to evaluate posterior fossa and obtained different biometrics, such as the surface area of cerebellar vermis. They suggested that such evaluation may potentially be used for early diagnosis of vermian anomalies.

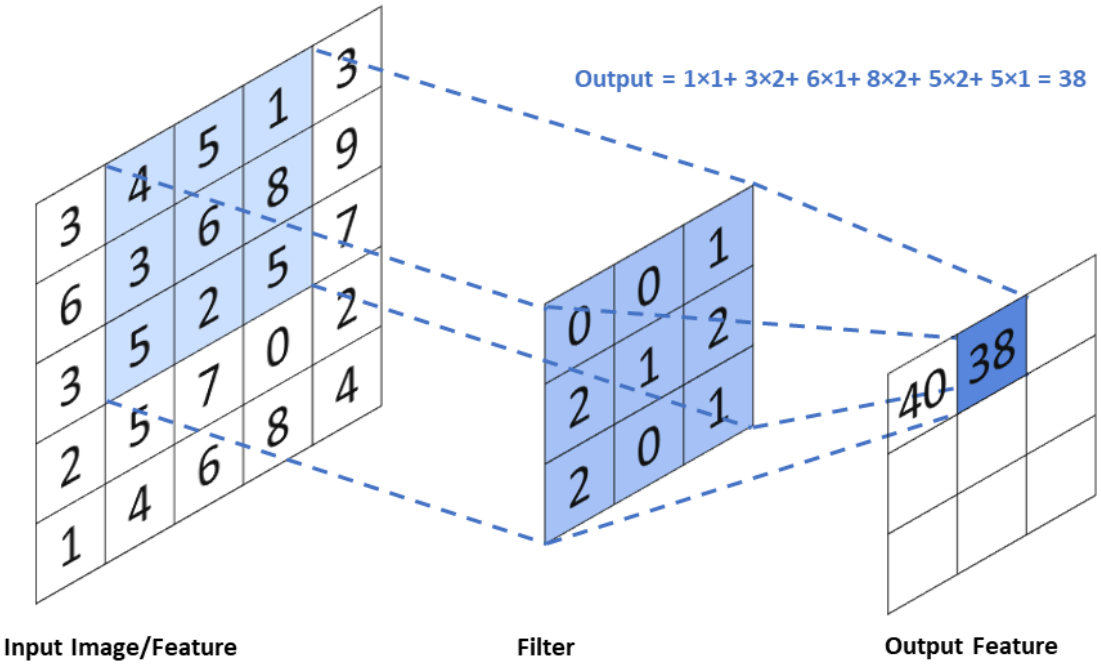
In addition, automated analysis techniques targeting 3D ultrasound fetal brain images are being developed, for instance, models to predict gestational age [45, 46] and extract standard planes from 3D images [47]. In summary, 3D ultrasound has great potential to provide higher degree of flexibility and more information to clinicians during diagnosis.

Cost and size are another important consideration when comparing 2D and 3D ultrasound. The cost of a medical ultrasound system varies according to its functionalities. A commercial system from quality brands with 3D scanning capabilities normally costs more than £10,000. High-tier systems, such as GE Voluson E10 and Siemens S2000, with more advanced features may cost more than £100,000 [48]. With much simpler hardware requirement and probe design, some handheld medical 2D ultrasound systems, such as Butterfly iQ and GE Vscan, cost around £1,500 to £3,000 [49]. These handheld systems are about the size and weight of a mobile phone or tablet computer, which makes them highly portable (Fig. 2.2a). On the contrary, 3D ultrasound system, such as the GE Voluson E8 (Fig. 2.2b), is 1.3-meter tall with the weight of 120kg. The much higher cost and bulkier size may limit the uptake of 3D ultrasound in some settings, for example small hospital in remote regions.

One of the potential contributions of this thesis is to reconstruct 3D ultrasound image of fetal brain from a sequence of 2D images acquired at different orientations. This may bridge the advantages of both 2D and 3D ultrasound, by reserving the cost-effectiveness, portability and routine usage of 2D ultrasound, while possessing the aforementioned potential and flexibility of 3D ultrasound.



(a) Step 1



(b) Step 2, stride equals 1

Figure 2.3: Computation of convolutional layer. (a) and (b) show two consecutive computational steps, with stride equals 1.

2.3 Convolutional Neural Network

This thesis is about computer-aided medical image analysis and processing, which can be categorized as a sub-field of computer vision. Since 2012, the field of computer vision has been revolutionized by convolutional neural network (ConvNet). It is a subclass of multi-layer perceptron (MLP), where the sharing of kernel parameters enables ConvNet to be robust to translation of objects in the image and less prone to overfitting when compared to MLP, gaining it great success.

The first ConvNet-typed network was neocognitron [50], a hierarchical network with multiple layers. It was inspired by human visual system, where local features were extracted at earlier stages, which were gradually integrated into global features at later stages.

ConvNet is most commonly used for processing data in the form of regular grid structure, for example images (*i.e.* 2D grid of pixels) and volumes (*i.e.* 3D grid of voxels). With an input image of dimensions of height H, width W and channel C, it will pass through the convolutional layers, where matrix multiplication is taking place, followed by non-linear activation, normalization and pooling. Normally, a series (*e.g.* N) of such computation will result in a feature map of dimensions of height $\frac{H}{2^N}$, width $\frac{W}{2^N}$ and number of features F, which can be used for different downstream tasks, for example classification and segmentation.

In this section, the basic building blocks of ConvNet will be introduced, followed by different representative ConvNet architectures. The supervised training of ConvNet will be presented finally.

2.3.1 Basic Building Blocks

Convolutional Layer. The convolutional layer is the core component of ConvNet. When an input image is fed into a ConvNet, it will pass through a series of convolutional layers (and other layers). By convention, the output from a convolutional layer is called the **features**. The convolutional layer is composed of a set of learnable filters of dimensions $k \times k \times d_i \times d_o$, where k is the spatial dimensions of the filters

(*i.e.* **receptive field**), d_i and d_o are the input and output depth, respectively. For example, a convolutional layer of dimensions $3 \times 3 \times 16 \times 64$ will process input features with depth 16 to output features with depth 64. When an input image or its features pass through a convolutional layer, the filters slide through the whole input image/features, where scalar product is computed to output the features' value at each pixel location. Two hyperparameters, namely **stride** c_s and **padding** c_p, need to be defined by the user. Stride refers to the step size of the filters. If the stride is small, the receptive field will significantly overlap. Adjusting the stride can control the degree of overlapping. Padding refers to concatenation of some values at each border of the input. For example, zero-padding of 3 means concatenating the input with 3 rows/columns of zeros at each border. The spatial dimensions of the output features are controlled by the spatial dimension of the input, c_s and c_p, using the following formulae:

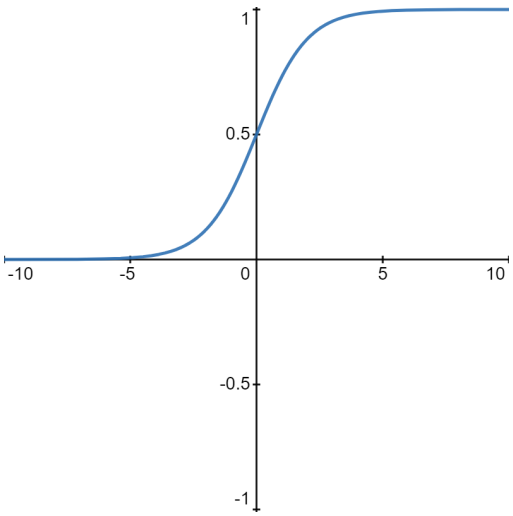
$$H_{out} = \frac{H_{in} + 2c_p - k}{c_s} + 1 \tag{2.1}$$

$$W_{out} = \frac{W_{in} + 2c_p - k}{c_s} + 1 \tag{2.2}$$

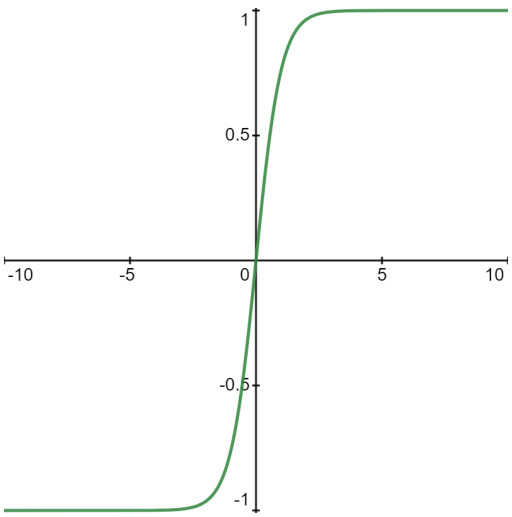
where H_{in}, H_{out}, W_{in}, W_{out} are the input and output height and width, respectively.

Fig. 2.3 demonstrates the computation of a convolutional layer. The input dimension is $5 \times 5 \times 1$. With the filters of dimension $3 \times 3 \times 1 \times 1$, according to Eqs. (2.1) and (2.2), the output features will have the dimensions of $3 \times 3 \times 1$. In practice, d_i and d_o are usually much larger than 1 and increase across layers, which enables the extraction of increasing amount of features. When compared to MLP, the number of filter weights is much lower, due to parameter sharing. This is based on the assumption that different regions at the image share some similarities, which can be identified with the same filter. Therefore, the filters can be reused, simply by sliding them through the whole image. During training, the weights of the filters are updated through backpropagation, which will be introduced later in this chapter.

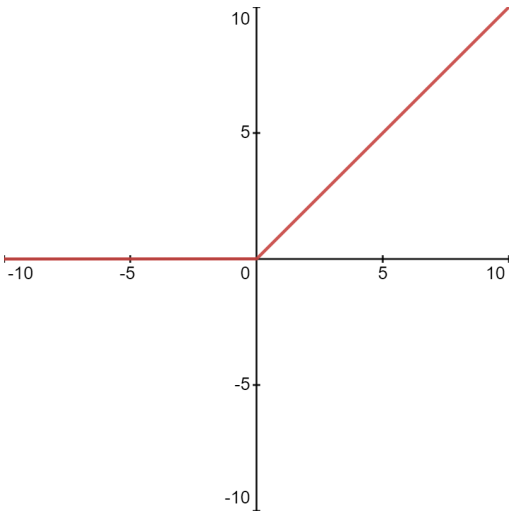
Non-linear activation. The computation of convolutional layer only linearly transforms the input. To learn the more complex non-linear relationships between



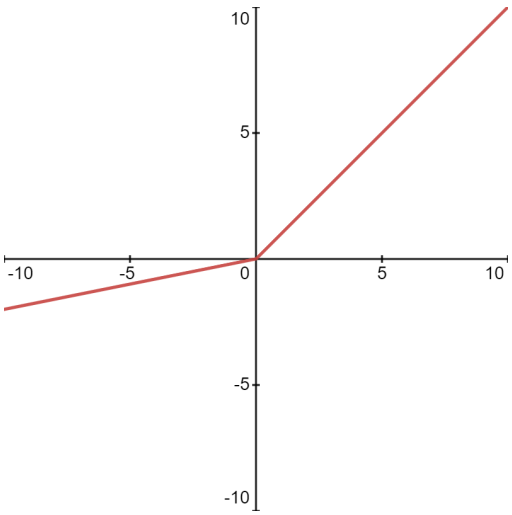
(a) Sigmoid, $\frac{1}{1+e^{-x}}$



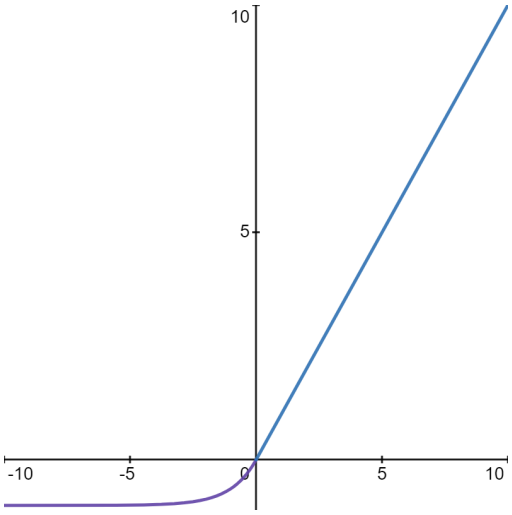
(b) Tanh



(c) ReLU, $\max(0, x)$



(d) Leaky ReLU, $\max(\alpha x, x)$



(e) ELU, $\beta(e^x - 1)$ for $x < 0$ and x for $x \geq 0$

Figure 2.4: Examples of common non-linear activation functions.

the input and the output, non-linear activation usually follows the convolutional layer. The family of non-linear activation functions is big and still expanding. Some common ones are plotted in Fig. 2.4 and introduced as follows:

- **Sigmoid** (Fig. 2.4a): Sigmoid is usually used in the output layer of ConvNet when the output is probability-based (*i.e.* from 0 to 1), given that:

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

  is within the range of 0 and 1. The major drawback of Sigmoid is the vanishing gradient problem as x is getting closer to 1 or 0, the gradient of the function will approach 0, making the training much slower and more difficult, especially if the ConvNet is deep. Also, Sigmoid is non-zero centered, meaning that the output from it is always moved to either positive or negative values. This makes the training of the network more computationally expensive [51]. Therefore, Sigmoid is seldom used with the hidden convolutional layers nowadays.

- **Tanh** (Fig. 2.4b): Hyperbolic tangent (Tanh) is similar to Sigmoid, except that it ranges from -1 to 1. Although it is zero centered, it also suffers from the vanishing gradient problem as Sigmoid does. Therefore, although Tanh was used with the hidden convolutional layers in early version of ConvNet, such as LeNet [52], it is rarely used for this purpose nowadays.

- **ReLU [53]** (Fig. 2.4c): Rectified linear unit (ReLU) is one of the most popular activation functions used with the hidden convolutional layers. Despite its simplicity,

$$\text{ReLU}(x) = \max(0, x) \tag{2.4}$$

  the vanishing gradient problem, as encountered by Sigmoid and Tanh, is much less severe for ReLU because the gradient is always 1 when x is larger than 0. In addition, its simple computation, without calculating divisions or exponential, speeds up the whole training process.

- **Leaky ReLU** (Fig. 2.4d): One problem of ReLU is that when x is negative, the gradient is always 0, which will affect the training of the ConvNet. Leaky ReLU retains small negative output values when x is negative by

$$\text{Leaky ReLU}(x) = \max(\alpha x, x) \tag{2.5}$$

where $0 < \alpha < 1$ is a hyperparameter. This allows non-zero gradients even when $x < 0$.

- **ELU [54]** (Fig. 2.4e): Similar to leaky ReLU, exponential linear unit (ELU) also allows non-zero gradients when x is negative:

$$\text{ELU}(x) = \begin{cases} \beta(e^x - 1), & \text{if } x < 0 \\ x, & \text{otherwise} \end{cases} \tag{2.6}$$

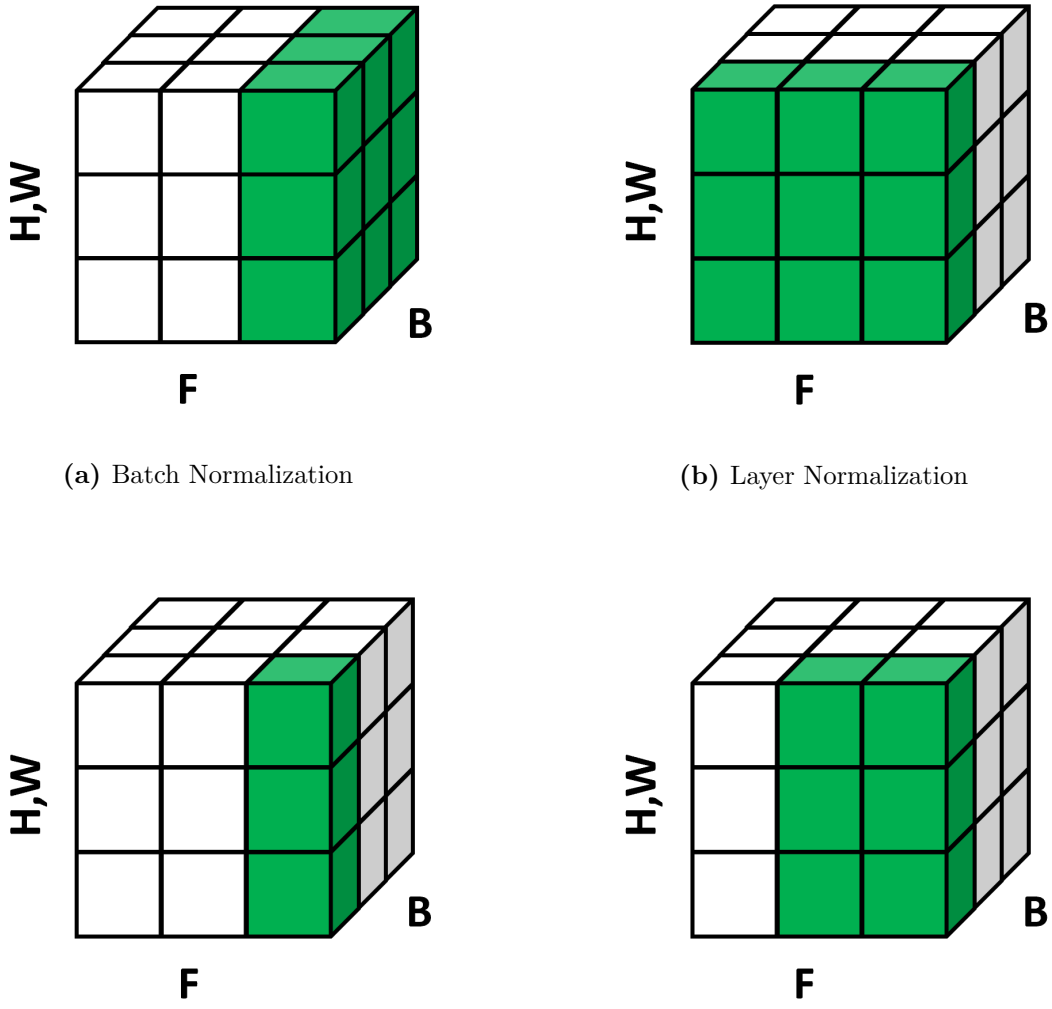
where β is a hyperparameter. When compared to ReLU, the negative part of ELU may push the mean output values closer to 0, which can speed up the training and decrease the bias shifts.

- **Sinusoidal**: Sinusoidal (*i.e.* periodic) is not a common activation function of ConvNet. Nevertheless, it achieves much better performance when it is used for implicit representation (Chapter 5) as it is more capable of modeling complex natural signals with fine details and, hence, their spatial and temporal derivatives [55].

- **Softmax**: Softmax is usually used in the output layer of ConvNet for multi-class classification problems. With an input of a vector of n elements, softmax will output a value from 0 to 1 for each input element, which can be interpreted as its probability, by:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2.7}$$

where x_i is the i^{th} element of the input vector. The sum of the output values of all elements equals to 1.



(a) Batch Normalization

(b) Layer Normalization

(c) Instance Normalization

(d) Group Normalization

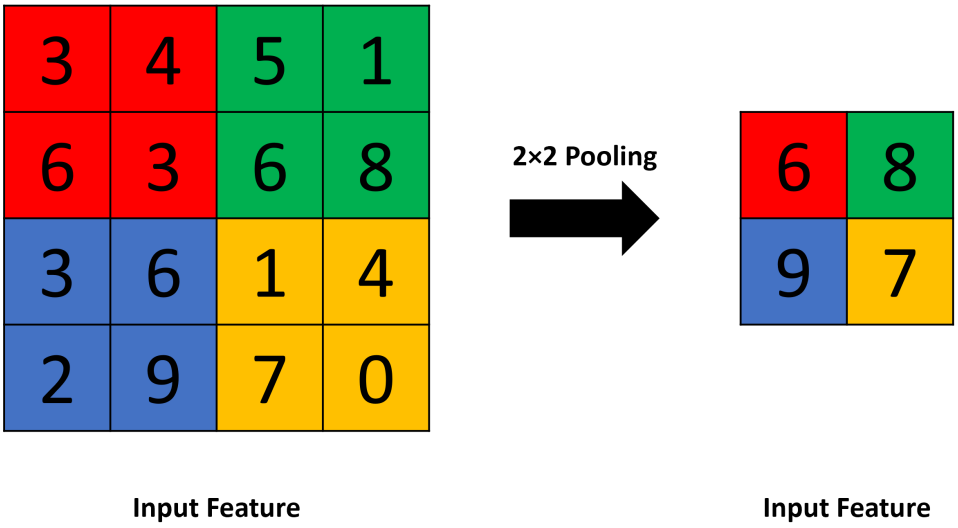
Figure 2.5: Examples of common normalization layers. H, W, F and B represent the height, width, number of features and batch, respectively. Figures recreated from [56]

Normalization. In 2015, Ioffe et al. [57] proposed **batch normalization** (Fig. 2.5a) for accelerating deep network training by reducing internal covariate shift. It was suggested that during training, the continuous update and, hence, shift of distribution of each layer's parameters make the training more difficult. By standardizing the mini-batch inputs of each layer to have approximately zero mean and unit variance, higher learning rates and less deliberately-designed initialization of parameters can be used for training deep ConvNet, accelerating the training and leading to convergence of better solutions [56]. Debate about the contributions of

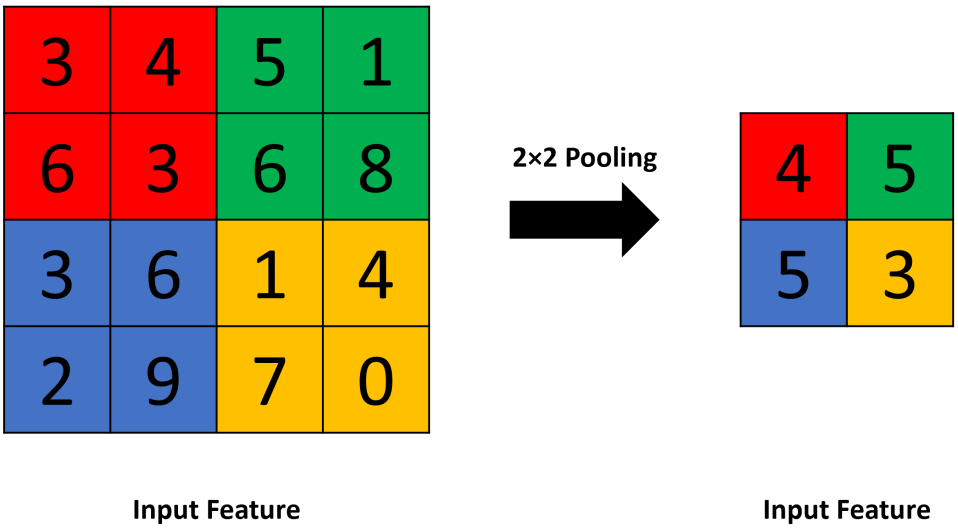
batch normalization is still going on, but there are several more popular assumptions trying to explain the success of it. Santurkar et al. [58] suggested that batch normalization smooths the optimization landscape during the training, which makes the gradient more predictive and, hence, leads to faster and more stable training. Qiao et al. [59] studied it from the angle of elimination singularities. It was suggested that batch normalization prevents the model from approaching elimination singularities (*i.e.* points along the training trajectory where neurons become deactivated). Getting too close to them may affect the training time and cause adverse effects to the model performances.

Inspired by batch normalization, Ba et al. [60] proposed **layer normalization** (Fig. 2.5b). Instead of computing the mean and variance over a mini-batch of inputs, they are computed along the feature dimension, such that the normalization is independent on the mini-batch size. It was originally proposed for recurrent networks and is now used in most Transformer-typed networks [61]. Similarly, Wu et al. [62] further proposed **group normalization** (Fig. 2.5d), which is also applied along the feature dimension. The main difference is that features are divided into g (*i.e.* a hyperparameter) groups, where each group is normalized separately. It was designed for computer vision tasks, such as object detection and segmentation, where small batch size is usually used due to memory constraint. For a special case when $g = 1$, it is called **instance normalization** (Fig. 2.5c). Fig. 2.5 summarizes different types of normalization and how the means and variance are computed.

Pooling. Pooling can reduce the spatial dimensions of the features. With fixed and limited computational memory, this may create budget for increasing the depth of the features. Pooling scales the input features by different functions, for example *max* (Fig. 2.6a) or *average* (Fig. 2.6b). The spatial dimension of the output features can be controlled by the **kernel size** and **stride** of the pooling operation. In Fig. 2.6, a pooling of kernel size of 2×2 and stride of 2 is demonstrated. Since information will be lost after pooling, having large kernel size or stride is not recommended and may lead to degradation of performance.



(a) Max Pooling



(b) Average Pooling

Figure 2.6: Examples of common pooling computation.

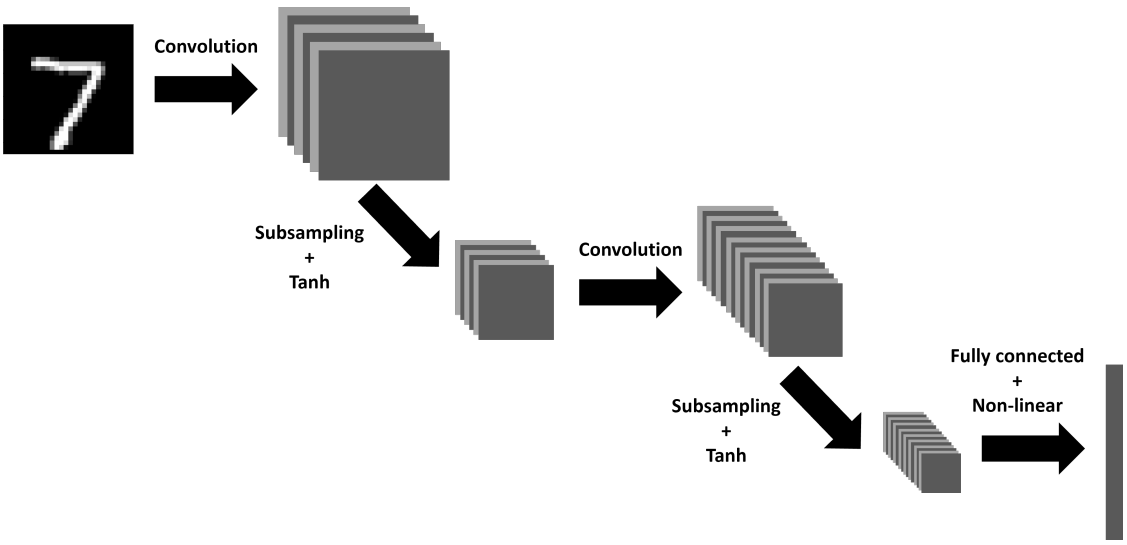


Figure 2.7: The architecture of LeNet [52].

2.3.2 Popular ConvNet Architectures

Since the first emergence of ConvNet, many different architectures have been proposed. Nevertheless, there are some popular architectures that are the milestones of the development of ConvNet, which will be introduced as follows:

- **LeNet [52]**: LeNet, as proposed in the 1990s, marked the first success of ConvNet. As shown in Fig. 2.7, the input image is first passed through convolutional layers, followed by subsampling (*i.e.* similar to pooling operation to decrease the spatial dimension of the feature maps) and Tanh non-linearities. Similar operation repeats again, before the feature maps are flattened and passed through some fully connected layers for classification. LeNet was tested on MNIST, a handwritten digits dataset. Despite the fact that it was not tested on more real-world settings due to the computation limitations at that time, LeNet formed a basis for the design of modern ConvNet architectures.

- **AlexNet [64], VGGNet [65] and GoogLeNet [63]**: AlexNet marked the next breakthrough of ConvNet, where it achieved a significant margin of improvement on the ImageNet Challenges [7], thanks to the boost of computation power enabled by hardware (*i.e.* GPU) advancement and the availability of large amount of labelled data. When compared to LeNet,

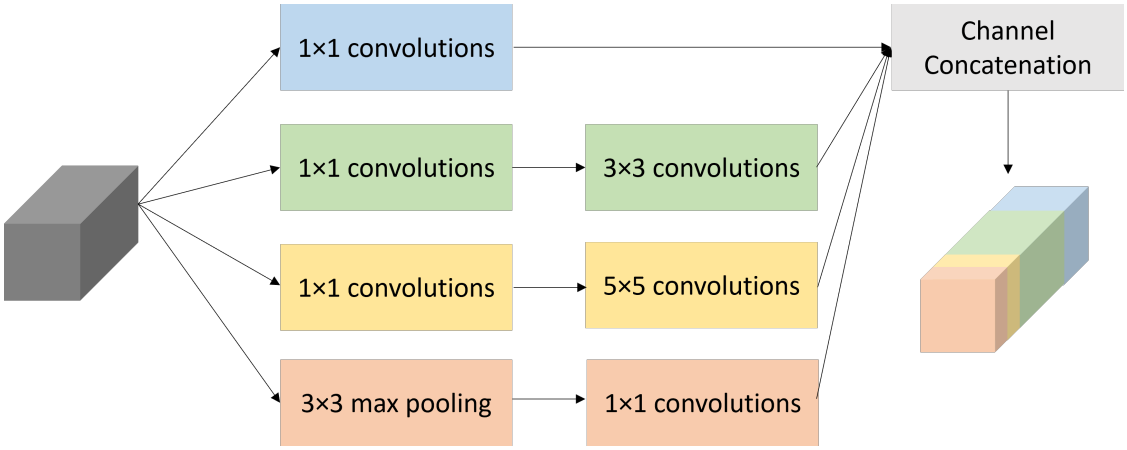


Figure 2.8: Inception module of GoogLeNet [63].

there are several modifications on the architecture, including having more convolutional layers; having a much larger kernel size (11×11) and strides of 4 at the first convolutional layer; using ReLU instead of Tanh as the hidden non-linearities and replacing subsampling with max pooling. Building upon AlexNet, various networks have been proposed, which achieved even better performance on the ImageNet Challenges. VGGNet [65] replaces convolutional kernels of variable sizes (*i.e.* 3×3, 5×5 and 11×11 kernels) in the AlexNet with 2 or 3 convolutional layers of fixed size (*i.e.* 3×3 kernels) as the building block. Max pooling follows each building block to decrease the spatial dimensions of the feature maps by half. Using this systematic design, different versions of VGGNet with different depths have been proposed, which also reduces the number of trainable parameters when compared to AlexNet due to the smaller convolutional kernel size. GoogLeNet [63], on the other hand, uses variable convolutional kernel size as AlexNet. However, the convolutional kernels of variable sizes are used within the proposed Inception Module (Fig. 2.8). It was suggested that such design is better to capture features of different sizes. Intuitively, when compared to AlexNet, VGGNet is deeper, while GoogLeNet is wider.

- **ResNet [66]**: If the depth of VGGNet keeps increasing, the training may become unstable and eventually cause adverse effect to the network's performance.

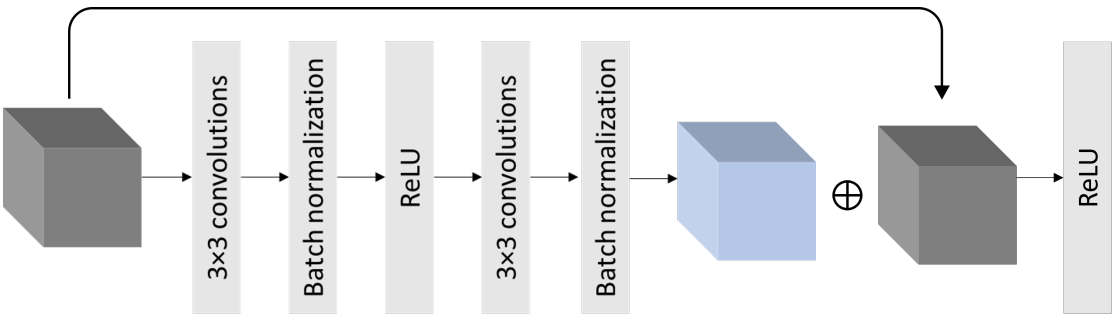


Figure 2.9: Residual unit of ResNet [66].

He et al. [66] proposed residual unit (Fig. 2.9), where skip connection is placed across every two convolutional layers to bypass convolutional computations. By using residual units as the building block, gradients can be back-propagated more easily by easing the vanishing gradient problem and the resulting ConvNet (*i.e.* ResNet) can reach the depth of hundreds of convolutional layers. ResNet led to another boost of performance of ConvNet and residual unit is a standard component of most of the ConvNet architectures nowadays.

- **Fully Convolutional Network (FCN) [52] and UNet [67]**: While the aforementioned ConvNets were proposed for classification tasks, pixel-level prediction, for example semantic segmentation and image reconstruction, is also very common. To serve this purpose, FCN [52] was proposed, which is composed of two parts. The first part, similar to the aforementioned ConvNets, downsamples the input image through convolutional and pooling operations, resulting in feature maps with much smaller spatial dimensions. Instead of passing them through fully connected layers, the feature maps are passed through series of upsampling (or transposed convolutional) and convolutional layers, gradually restoring the spatial dimensions of the feature maps. Finally, they are passed through a 1×1 convolutional layer to output per pixel prediction. Building upon FCN, Ronneberger et al. [67] proposed UNet (Fig. 2.10) for biomedical image segmentation. When compared to FCN, features from different levels of the downsampling and upsampling parts are concatenated through multiple skip connections. Low- and high-level

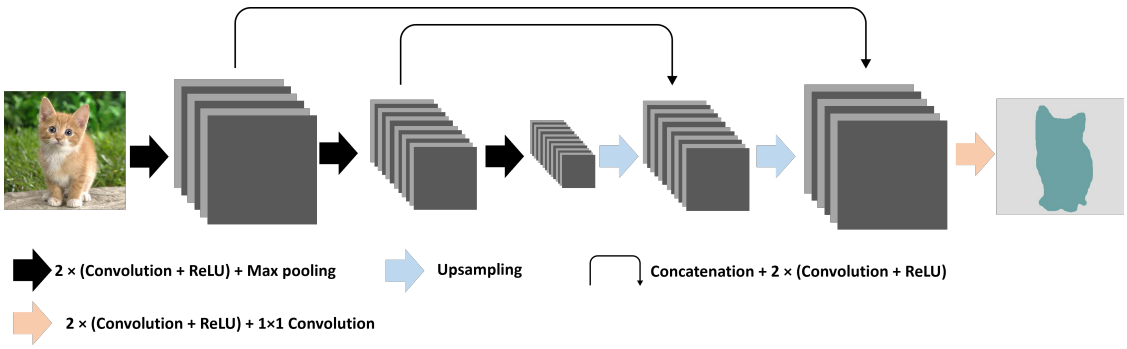


Figure 2.10: A simplified schematic of the architecture of UNet [67].

information can therefore be fused better before making the prediction, which generally leads to better performance in semantic segmentation. UNet is very popular in the field of medical image analysis and is one of the most commonly used baseline models for comparison.

2.3.3 Supervised Training of ConvNet

Intuitively, training a ConvNet in a supervised manner is an iterative process. For every iteration, the network tries to make predictions on the training inputs. The **loss function** informs the network how "wrong" the predictions are, which can then guide the update of the network's parameters through **back propagation**. In this section, different components involved in supervised training of ConvNet will be introduced.

Loss Function. Supervised learning requires paired training data in the form of $\{x_i, y_i\}$, where x_i is the input data and y_i is the corresponding label, which is usually obtained by manual annotation. When x_i is passed through a ConvNet, $\psi(\cdot;\theta)$, parameterized by θ, the prediction is defined as $\hat{y}_i$:

$$\hat{y}_i = \psi(x_i;\theta) \tag{2.8}$$

The goal is to minimize the prediction error, namely the difference between y_i and $\hat{y}_i$, which is quantify by the loss function, $\mathfrak{L}$. Depending on the task, $\mathfrak{L}$ can be of different

forms. For example, for regression problem, mean least-square error cab be used:

$$\mathfrak{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{\mathbf{y}}_i - \mathbf{y_i})^2 \tag{2.9}$$

where N is the number of data of the whole dataset or within a mini-batch, depending on the type of optimization used, which will be introduced in the following section. If the loss is big, the difference between the ground-truth label and prediction is big, meaning that the performance of the network is not good. This not only provides an indicator of how good the ConvNet performs, but also helps the optimization of the network by providing a direction for its parameters to be updated, which will be introduced in the following section.

Optimization. With the loss function, each ConvNet parameter, θ_j, can be updated by gradient descent:

$$\theta_j^{new} = \theta_j^{old} - \alpha \frac{\partial \mathfrak{L}}{\partial \theta_j} \tag{2.10}$$

where α is the learning rate, which is a hyperparameter. For naïve gradient descent, $\mathfrak{L}$ is calculated on the entire dataset for one optimization step, which is not efficient. Therefore, stochastic gradient descent for mini-batch is proposed, where $\mathfrak{L}$ is calculated on a mini-batch of data for one optimization step. However, due to the much smaller sample size of a mini-batch, the variance may become higher. Therefore, momentum, γ is proposed to dampen the fluctuation of the update of the parameters during optimization. Specifically, v is defined as

$$v_t = \gamma v_{t-1} + \alpha \frac{\partial \mathfrak{L}}{\partial \theta_j} \tag{2.11}$$

where θ_j is updated by

$$\theta_j^{new} = \theta_j^{old} - v_t \tag{2.12}$$

Optimization is a very active research area. In recent years, different optimization techniques, such as Adam [68] and AdamW [69], have been proposed and used for optimizing ConvNet.

(a) Original **(b)** Grayscale **(c)** Gaussian noise

(d) Blur **(e)** Flip **(f)** Contrast

(g) Crop **(h)** Rotate

Figure 2.11: Common image augmentation. Examples created by Albumentations[70]

Data Augmentation. One of the drawbacks of training a deep ConvNet is the large amount of training data required. However, in reality, it is very costly and sometimes impossible to obtain large amount of data. In practice, one technique that is often employed when training a deep ConvNet is image augmentation. Specifically, an image is artificially modified to create a new variation. As illustrated in Fig. 2.11, some common modifications include geometric augmentations (*e.g.* rotation and flipping), color and contrast augmentations (*e.g.* RGB to grayscale) and noise and blurring augmentations (*e.g.* Gaussian noise and blur). Several augmentations can be combined, which may theoretically create infinitely different variations of the training images. Nevertheless, excessive augmentations may harm the final performance. For example, if an image is too blurry or rotated to an extent that is not realistic, using it to train a ConvNet may pose adverse effect to it when it is

tested on other non-augmented images. Therefore, in practice, augmentations are some hyperparameters to be tuned during training.

This section presented the standard supervised training pipeline of ConvNet. In this thesis, other forms of training and learning are involved, which vary from the standard supervised pipeline. For example, in Chapter 3, 4 and 6, the label, y, is not obtained from manual annotation, but generated from the data itself. In Chapter 5, not only the weights of the network, but also the input will be optimized through back propagation. The details will be presented in the corresponding chapters.

2.3.4 Self-Attention

Self-attention is getting more popular and commonly used with ConvNet recently. It is a kind of learnable weighting mechanism, which was first proposed for the task of neural machine translation [71, 72]. Before self-attention was proposed, machine translation was achieved by sequence-to-sequence learning [73], where an encoder network processes the input sentence into a fixed-length feature vector and an decoder network processes this feature vector to output a sentence in the target language. Normally, recurrent neural networks (RNN), for example long short-term memory (LSTM) [74] or gated recurrent units (GRU) [75], are used as the encoder and decoder network. One of the drawbacks is that the fixed-length feature vectors generated by the encoder may not be a good representation of the whole input sentence, especially when it is long. Instead of just inputting a fixed-length feature vector to the decoder, Bahdanau et al. [71] proposed to use feature vectors of every word of the input sentence during decoding. When the decoder outputs a word, the input it receives would be a previous hidden state plus the weighted sum of the feature vectors of every word of the input sentence. A learnable attention layer is used to assign weights to feature vectors of different words of the input sentence according to the previous hidden state of output and current position in the output sentence. Vaswani et al. [72] incorporated an attention mechanism

in the encoder part for neural machine translation. This eliminates the need of RNN-typed encoder and may enable the learning of more complicated relationships between words in the input sentence, which is important for translation. The proposed model generates 3 feature vectors, **key**, **query** and **value**, to represent each input word/vector. By computing a scaled dot-product of these three sets of vectors, it is equivalent to generating an updated feature vector for each input word/vector by combining weighted information of every input word/vector. Wang et al. [76] applied a similar idea in computer vision problems by treating images and videos as spatial and temporal sequences of pixels. The proposed network extracts the feature of a location in an image or video by computing a weighted sum of features at all positions, which is the same as computing a scaled dot-product of the key, query and value vectors as proposed in [72].

Besides neural translation, the attention mechanism has been applied to different computer vision and medical imaging studies. Jetley et al. [77] proposed to use global features (i.e. output of the fully connected layer after the feature extractor network) as the query vector and compute an attention map by comparing with local features (i.e. different positions in the feature maps generated at different levels of the feature extractor network). The attention map generated in this way may indicate which part of the image contributes more significantly to the prediction, which is similar to object detection even though only image-wise labels are used as supervision. Schlemper et al. [78, 79] applied similar concepts in training a model for ultrasound standard plane detection and segmentation, which may be able to focus on and indicate regions with representative anatomical landmarks during prediction. In this thesis, self-attention and related mechanisms are manipulated in Chapter 3 and 6.

2.4 Deep Learning - Self-Supervised Learning

One of the main reasons for the current success of deep ConvNet in computer vision is the availability of large labelled datasets, such as ImageNet [7] and COCO

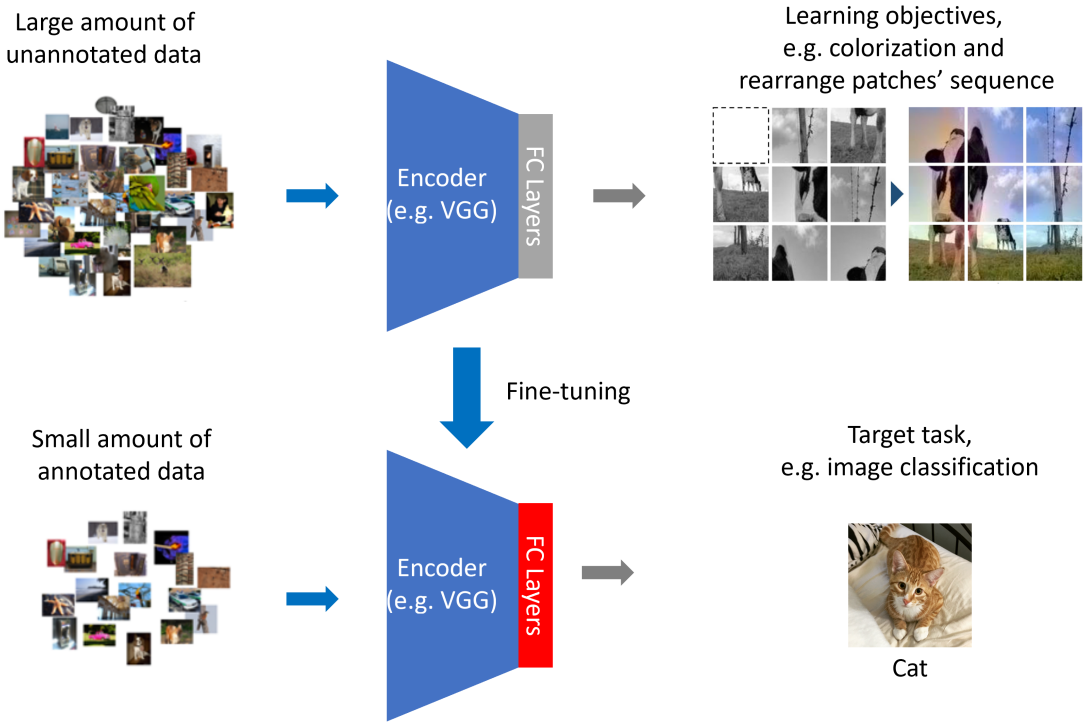


Figure 2.12: The general pipeline and learning objective of self-supervised learning. Using large amount of unannotated data to train a model using some proxy tasks, the model can then be fine-tuned by small amount of annotated data for the target task, such as image classification.

[80], for supervised learning. However, human annotations are often expensive. This is even more problematic in the field of medical imaging, where annotating images requires extensive anatomical knowledge. Therefore, as explained in Chapter 1.1, one of the fundamental principles that motivates the works proposed in this thesis is to use minimal human annotation. One way to achieve this is to employ self-supervised learning, where the supervision is coming from the data itself. As shown in Fig. 2.12, the fundamental idea of self-supervised learning is to train a model, on some **proxy tasks**, to learn a good intermediate representation of an image and its objects by using unannotated data. The model can then be fine-tuned with a small amount of annotated data and adopted in the **target tasks**.

The most straightforward example of a self-supervised learning technique is an autoencoder [81], where an encoder-decoder network is trained to regenerate the input Fig. 2.13. The encoder will first convert the high-dimensional input to a low-

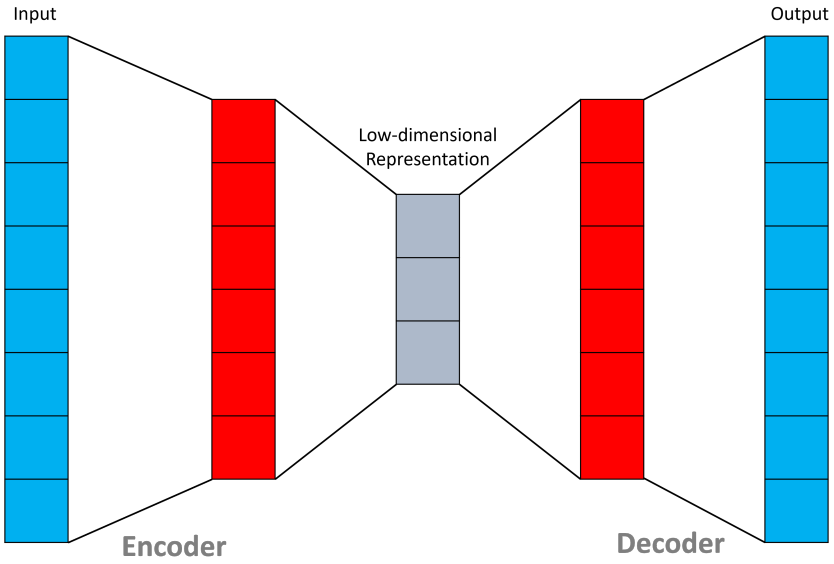


Figure 2.13: The architecture of an autoencoder. The encoder will first convert the high-dimensional input to a low-dimensional representation, which will then be used by the decoder to reconstruct the input.

dimensional representation, which will then be used by the decoder to reconstruct the input. Since information would be lost during encoding, the encoder needs to learn the most representable features of the input such that these low-dimensional features can still be used to reconstruct the high-dimensional input by the decoder. The learned encoder can be used to initialize another model for related tasks. Since the output of the autoencoder is the input itself, no human annotation is needed. Vincent et al. [82] extended this idea by inputting data corrupted by noise to the autoencoder and it was trained to output the uncorrupted input data. The aim is to let the model learn a better representation of the data that captures more useful and meaningful information. While autoencoder-based techniques rely on data reconstruction as the learning objective, a number of self-supervised learning studies investigated its use for context prediction [83–85], which was inspired by the skip-gram model [86] in natural language processing, as the learning objective. By using a specific word in the sentence as the input, skip-gram model [86] is trained to predict the words around it in the sentence. Applying this idea to computer vision, Doersch et al. [83] trained a model to predict the relative location of a pair of patches sampled from an image. The idea is that in order to correctly predict

the relative position of different parts of an object, for example eyes and mouth on a face, the model must first know what a face looks like, which is equivalent to learning a good representation of the face. Kim et al. [84] made the task even more challenging by training a model to recover the arrangement of shuffled patches sampled from an image, which was similar to solving jigsaw puzzles. Gidaris et al. [85] randomly rotated the input image and trained a model to predict the rotation. Besides spatial information, colorization, a process that adds color to an image, can also be used to supervise learning of image features [84, 87, 88]. By inputting a grayscale image or single-channel image to the model and predicting the colorization of the image, the model may be able to learn features, such as blue sky and brown wood, along with the class labels.

Nevertheless, predicting low-level features (*i.e.* colorization or pixels' value) may be easily affected by the low-level noise of the image, which is undesirable because the goal of self-supervised learning is to learn a high-level representation of the input. Instead of learning to predict features in the low-level image space, Oord et al. [89] formulated self-supervised learning as making prediction in the high-level latent space. By inputting a part of the image to the model, it needs to learn to differentiate features of the remaining part of the image from features extracted from other images [90]. Since the learning and loss are based on high-level features, the learned representation would be more robust to low-level perturbations and artifacts. One of the main contributions of those studies [89, 90] is the proposal of using contrastive loss for self-supervised learning. Specifically, the contrastive loss "pull together" the features of an image and its augmented version (*i.e.* positive pair), while "push apart" the features of that image and other different images. Using similar idea, different frameworks, for example SimCLR [91], MoCo [92] and PIRL [93], were proposed and achieved state-of-the-art self-supervised learning performance.

With additional temporal information, video data (a set of 2D images in a sequence) may provide even richer source of supervision when compared to a static 2D image. Misra et al. [94] trained a model to predict whether the input video clip was in correct temporal order by sampling tuples of frames from videos and shuffled

some of them to form positive and negative training data. Lee et al. [95] extended the idea of perturbing the video sequence by shuffling the frames and training a model to sort the sequence of the input frames that were temporally shuffled. Besides frames sequence, visual tracking can be another form of supervision. Wang et al's work [96] built upon the idea that tracking two patches within the same video, for example track of a cat, should have a closer representation in the deep feature space than two random patches from two different videos, for example a cat and a bicycle. This representational difference of distance in the feature space can be used to supervise the learning.

Consistency of prediction can be used as a source of supervision as well. Zhou et al. [97] proposed a method using cycle consistency to supervise a model that learns to predict dense visual correspondence across different object instances. With two synthetic object instances, s_1 and s_2 (synthetic images taken at different orientation of a 3D CAD car), a model can be trained to predict the dense visual correspondence of two real object instances, r_1 and r_2 (real images of different cars at different orientation), using the condition of consistency that the correspondence flow from s_1 to r_1 to r_2 to s_2 is the same as that from s_1 to s_2, which is known when s_1 and s_2 are synthesized. Chen et al. [98] used cycle-consistency loss to train a model to align images (*i.e.* geospatial images collected from satellites) to noisy annotations, such as misaligned annotations of buildings and railway. One of the ideas of the proposed method is that correct annotation is unique and the transformation from a noisy annotation to any further perturbed annotations, back to the correct annotation should be consistent, regardless of the perturbation applied. As the noise of the annotations is random, the model may gradually learn how to align the noisy annotations to the correct objects in the satellite images, which may show unique features.

Self-supervised learning is heavily manipulated in this thesis. In Chapter 3 and 6, the proposed frameworks are trained with supervision generated from the raw data, resembling the idea of self-supervised learning. In Chapter 4, the proposed unsupervised cycle consistency is inspired by Zhou et al.'s work [97] as introduced

earlier.

2.5 Deep Learning - Unsupervised Domain Adaptation

As explained in Chapter 1.1, two of the fundamental principles that motivate this thesis are to use minimal human annotation and enhance generalizability. One closely related area of research is unsupervised domain adaptation. The objective of unsupervised domain adaptation is to train a model using data from a labelled source domain and apply the model to test on data from an unlabelled target domain. In many scenarios, collecting annotations of one domain is much easier than collecting annotations of another related domain. One classical example is to train a model using photo-realistic simulations [99], which may be easily generated and labelled. The model trained by just the simulated images may not generalize well to corresponding real images, due to the different data distribution in the source domain (synthetic images) and target domain (real images) [100]. In the field of medical imaging, domain shift can be observed in several forms, including cross-modality variability due to the innate difference between different imaging modalities, and cross-vendor variability due to different post-processing protocols venders may apply [101]. As manual annotation of medical images is an expensive process, requiring a significant amount of time and expertise, there is strong incentive for investigating unsupervised domain adaptation. These methods aim to utilize existing annotated medical images to train a model on tasks involving unlabelled target images.

The goal of unsupervised domain adaptation in deep learning is to learn data features that are domain-invariant [102]. With the aligned feature space of both source and target domains, a classifier or decoder trained with labelled data of the source domain can be used on data of the target domain. Many of the state-of-the-art works applied this concept through adversarial learning by minimizing both the loss of the task, for example the classification or segmentation loss, and the domain divergence error [103]. Ganin et al. [104] proposed a gradient reversal layer, which

is used to reverse the gradient of the domain discriminator during training so that it is maximized while the label classifier loss is minimized as usual. This fools the model and encourages the feature extractor to generate domain invariant features during optimization. While [104] used the same feature extractor for both source and target domains, Tzeng et al. [105] proposed training the source and target feature extractor separately. A feature extractor and classifier were trained using the labelled source data. Then a target feature extractor was trained such that a domain discriminator could not differentiate between features extracted by the source feature extractor and features extracted by the target feature extractor. Finally, the target feature extractor and source classifier were combined to test on data from the target domain. Although these two works adopted different approaches during training, the fundamental principle of learning feature extractor(s) that extract features from source and target domain to a shared feature space is the same.

Another closely related approach incorporates the generative component of generative adversarial networks (GAN) [106]. In [107], a generator conditioned on the source images and noise vector generated synthetic target images. A classifier was then trained to predict class labels of both source and synthetic images, while the discriminator was trained to predict the domain labels of target and synthetic images. Hoffman et al. [108] extended the concept of the CycleGAN [109] to unsupervised domain adaptation. Instead of just training one generator to synthesize target images from source images, another generator which generated source images from (synthetic) target images was also learned. By doing so, cycle consistency can be employed as a loss to regularize the training. Recent studies [110, 111] adopted similar ideas to unsupervised domain adaptation. Nevertheless, the core idea is similar that synthetic images generated from a generator are used to confuse the domain discriminator so that features extracted from the source and target domain are indiscriminate with respect to domains but discriminative with respect to the main task.

Most of the aforementioned pipelines, especially the ones involve GAN, are complicated and difficult to train and reproduce. In addition, most of them were

proposed for and demonstrated on classification task only. In Chapter 4, a novel unsupervised domain adaptation mechanism, based on cycle consistency, is proposed, which achieves significantly better performance than other baseline unsupervised domain adaptation methods on a regression (*i.e.* 3D localization) task.

2.6 Computer-aided Ultrasound Scanning

As introduced in Section 2.1, the standard approach of fetal ultrasound scanning is to first determine the standard planes of view, which is not a trivial task. Therefore, numerous methods have been proposed for automated standard plane detection for 2D fetal ultrasound. Earlier studies [112–114] proposed to use the Adaboost classifier or support vector machine classifier to detect key anatomical landmarks in a sequence of 2D ultrasound images. Presence and orientation of the detected landmarks were used to identify an image as either a standard or non-standard plane of view.

Recently proposed methods employed ConvNet for standard plane detection. Chen et al. [115] fine-tuned a pretrained CaffeNet model [7] to detect the standard planes in ultrasound fetal abdominal images. Baumgartner et al. [116] further trained a ConvNet model to classify fetal ultrasound images into 14 categories, including different types of standard plane images and background images. Using an attention mechanism, Schlemper et al. [78] proposed a ConvNet model that may simultaneously perform standard plane detection and weakly supervised structure localization using only image-level class label for training. Some studies further utilized reinforcement learning for the task, which may better simulate the search process of manual scanning [117, 118]. A recent study [119] extended standard plane detection to a guidance system using an external motion sensor.

Spatio-temporal information of 2D ultrasound videos has also been explored for standard plane detection. Chen et al. [120] and Huang et al. [121] presented different multi-task recurrent neural network models that can utilize the temporal information of consecutive sequences in ultrasound videos to provide extra contextual clues for the detection task. Gao and Noble [122] used image-level labels to train a

two-stream spatio-temporal ConvNet to recognize fetal heart frames and localize the heart in freehand fetal ultrasound videos.

When 3D ultrasound scanning is available, if conventional biometrics, such as HC and TCD (Section 2.1), needs to be measured, one of the most straightforward ways is to identify the standard planes of view (*i.e.* cross-sectional views) from the acquired 3D volumes. Hence, a slightly different task from the aforementioned standard plane detection for 2D fetal ultrasound is standard plane localization in 3D volumes, which aims at identifying the standard planes within a given volume. Several studies have suggested different methods for this task. Ryou et al. [123] proposed to exploit sharp boundary information in the 3D ultrasound volume to detect the fetal region-of-interest (ROI) and then classify head and body slices within the ROI using a transfer learning ConvNet. The standard head and abdominal planes are automatically selected by incorporating prior clinical knowledge about the position of the standard plane within the two structures. Li et al. [47] presented a ConvNet that is able to output the transformation required to move the plane of the input 2D cross-sectional image of a 3D fetal brain ultrasound volume towards the standard plane of view. Such prediction is computed iteratively during inference. Recent studies [118, 124] proposed different reinforcement learning frameworks for standard plane localization in 3D MRI and ultrasound volumes. These reinforcement learning frameworks provide feedback from the environment (*i.e.* the 3D volume) during the search for the standard planes, which mimics the navigation performed by experienced operators when they are locating the target view planes in the volumes.

In this thesis (Chapter 3 and 4), a more general task is investigated, where the plane location of a 2D ultrasound scan is predicted in a predefined 3D anatomical atlas. This can be easily adapted to standard plane detection by simply identifying the standard planes in the predefined 3D anatomical atlas.

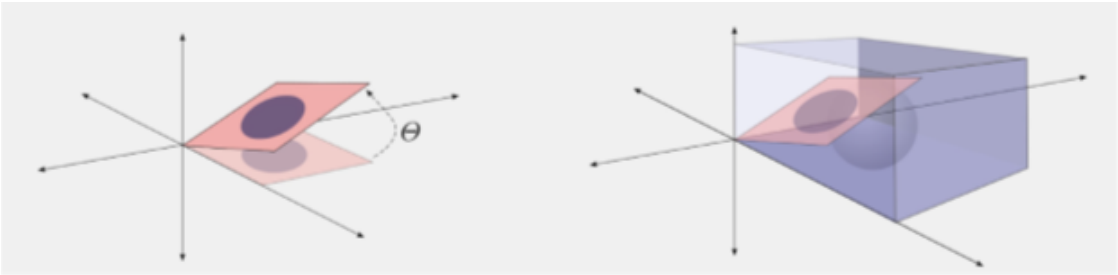


Figure 2.14: Illustration of slice-to-volume registration. Slice-to-volume registration is the process of aligning 2D images into a shared 3D coordinate system. Adopted from [125].

2.7 Slice-to-Volume Registration

Image registration is the process of aligning two or more images into a shared coordinate system. Slice-to-volume registration is a sub-class of this problem, where the images to be registered are 2D and the target coordinate system is 3D, which is directly related to the topic of this thesis, namely bridging the gap between 2D and 3D ultrasound imaging. As shown in Fig. 2.14, slice-to-volume registration is achieved by estimating the transformation, Θ, of a reference plane to its actual position in the 3D space. Several applications in medical imaging rely on slice-to-volume registration which have been extensively reviewed in [125]. For the scope of this thesis, this section will focus on tomographic 2D image (*i.e.* ultrasound or slice from MRI) in the application of motion correction and volume reconstruction.

2.7.1 Conventional Approaches

One of the major applications of slice-to-volume registration in medical imaging is motion correction of fetal MRI. Due to the movement of the fetus during image acquisition, inter-slice motion artifacts may corrupt the overall quality of the acquired 3D MRI volume. Alansary et al. [126] summarized a general framework for slice-to-volume registration for this specific task. Motion correction of a set of motion-corrupted images is the first step of this framework, where one image is selected as the initial reference and slices of the remaining images in the set are aligned to it. The reference is updated incrementally. Similarity between

the reference image and motion-corrupted images is measured by an objective function and its purpose is to guide the optimization of transformation parameters through gradient descent. Different objective function, including mutual information [127], cross-correlation [128, 129] and mean square difference [130], were used in different studies for registration.

With the motion-corrected images, volume reconstruction is achieved through a series of post-processing steps, for example data interpolation to fill the missing voxels, reducing blurring [127, 128], outlier removal [131] and bias correction [129]. Similar approaches have also been applied for 3D ultrasound volume reconstruction from 2D freehand scanned slices, Wen et al [132] further incorporated positional information from a motion sensor attached to the ultrasound probe to slice-to-volume registration. When 3D positional information is available, volumetric reconstruction is usually much simpler and is referred as compounding. Karamalis et al. [133] performed that of a freehand ultrasound sweep by finding line-plane intersection between ultrasound slices. Kainz et al. [131] developed an algorithm for fast volume reconstruction using GPU acceleration, which made online application of 3D ultrasound volume reconstruction possible during clinical scanning. Despite the effectiveness reported in these studies, the proposed approaches usually involve complicated pipelines, which are difficult to reproduce and highly specific to the task and data concerned [125]. This makes the comparison of these techniques very difficult.

2.7.2 Deep Learning Approaches

Thanks to the current success of deep learning in computer vision, some studies proposed the use of ConvNet for slice-to-volume registration. Hou et al. [134, 135] attempted to train a deep model to predict the rotations and translations of 2D slices sampled from a 3D MRI volume, which was aligned to an atlas coordinate system. They firstly corrected manually a set of motion-corrupted MRI volumes. Then, these volumes were aligned to an atlas so that they were in the same coordinate system.

Slices were randomly sampled from these aligned volumes and a model was trained to predict their position in the atlas coordinate system. Geometric loss, which minimizes both rotation and translation, was used in the training. With this learned model, slices from motion corrupted MRI volume can be registered to an aligned space. They further utilized slice-to-volume registration from [131] to process the output from the learned model to give the final reconstructed image. A similar deep learning pipeline was utilized for standard plane localization in 3D ultrasound [47]. During the training, in addition to geometric loss, image loss was used. Image loss captures the difference between slices sampled from the groundtruth position and predicted position. During inference, a slice sampled from the predicted location was fed to the model iteratively until the prediction converged.

In this thesis, Chapter 3 is inspired by [134, 135] to propose a network trained by 2D slices sampled from 3D ultrasound volumes. Chapter 4 adapts the trained network to native 2D ultrasound images by proposing a novel unsupervised domain adaptation mechanism.

3

3D Localization of 2D Ultrasound Images

This first contribution chapter proposes `PlaneInVol`, a ConvNet for predicting the position of 2D scans in a 3D brain atlas for fetal brain ultrasound. `PlaneInVol` is trained by sampling 2D slices from aligned 3D fetal brain volumes, resembling the idea of self-supervised learning, such that heavy annotations for each 2D scan are not required. The trained network takes a set of arbitrary number of images as input, and output the predicted 3D location of each individual 2D scan. The work presented in this chapter has been published in:

Yeung, P.H., Aliasi, M., Papageorghiou, A.T., Haak, M., Xie, W. and Namburete, A.I.,: Learning to Map 2D Ultrasound Images into 3D Space with Minimal Human Annotation., *Medical Image Analysis*, vol. 70, pp.72-86, May 2021.

Contents

3.1 Introduction

The goal of medical image analysis is to extract anatomical information from the images, which can be used for diagnosis, monitoring and treatment planning. In order to achieve this, the very first task is often ensuring that the image captures the regions and structures that are going to be analyzed and their positions and orientations are known. This generally applies to most medical image analysis tasks of different imaging modalities. For example, in MRI neuroimaging, the brain volumes are generally registered to a predefined 3D anatomical atlas before analysis and comparison, while in CT abdominal imaging, this can be easily controlled by adjusting the position of the machines and the patients before scanning. Comparatively, it is more challenging for the task investigated in this thesis, namely 2D ultrasound fetal neuroimaging, given its uniqueness:

- Unlike CT and MRI images that include the complete 3D anatomy, each 2D ultrasound image only captures 2D cross-sectional view of an inherently 3D anatomy. Identification and localization of the views and structures are, hence, more challenging.

- The orientation and position of the fetus is relatively arbitrary and unknown to the observer before scanning. The potential arbitrary fetal motion during the scanning may further complicate the whole process.

Conventionally, sonographers are trained to identify and locate the cross-sectional views of 2D ultrasound images by mentally reconstructing the 3D anatomy [42]. However, this requires a significant amount of training and involves subjective judgement and, hence, inter-operator variability is inevitable, which may further affect the subsequent analysis and diagnosis that directly relies on the acquired 2D images.

In this chapter, a ConvNet, `PlaneInVol`, is proposed to predict the corresponding location of 2D ultrasound fetal brain images, including both standard and non-standard planes, in a pre-defined 3D reference coordinate system (*i.e.* 3D fetal brain atlas). As such, the following contributions are presented:

(i) The localization of 2D ultrasound images of fetal brain in 3D space is defined as a self-supervised learning problem. Using 2D slices sampled from only a small number (*i.e.* 50) of aligned 3D volumes as training data, which are processed by the proposed preprocessing pipeline (*i.e.* Section 3.2.1), it is demonstrated that `PlaneInVol` trained in such manner may generalize to actual 2D ultrasound images and videos (Fig. 3.1).

(ii) Inspired by the idea of relation networks [136, 137], a new ConvNet architecture that takes an arbitrary number of input images as a set is proposed. It is demonstrated that this is a better utilization of available information and leads to improved performance. This setting is particularly suitable for 2D freehand ultrasound scanning, where a large but indeterminate number of 2D images are usually available.

(iii) `PlaneInVol` is first benchmarked on a synthetic dataset, where 2D slices are sampled from 3D volumes, and hence the groundtruth location of these slices is known. It is shown that the proposed model consistently outperforms a strong baseline described in [134, 135]. In addition, `PlaneInVol` is tested on

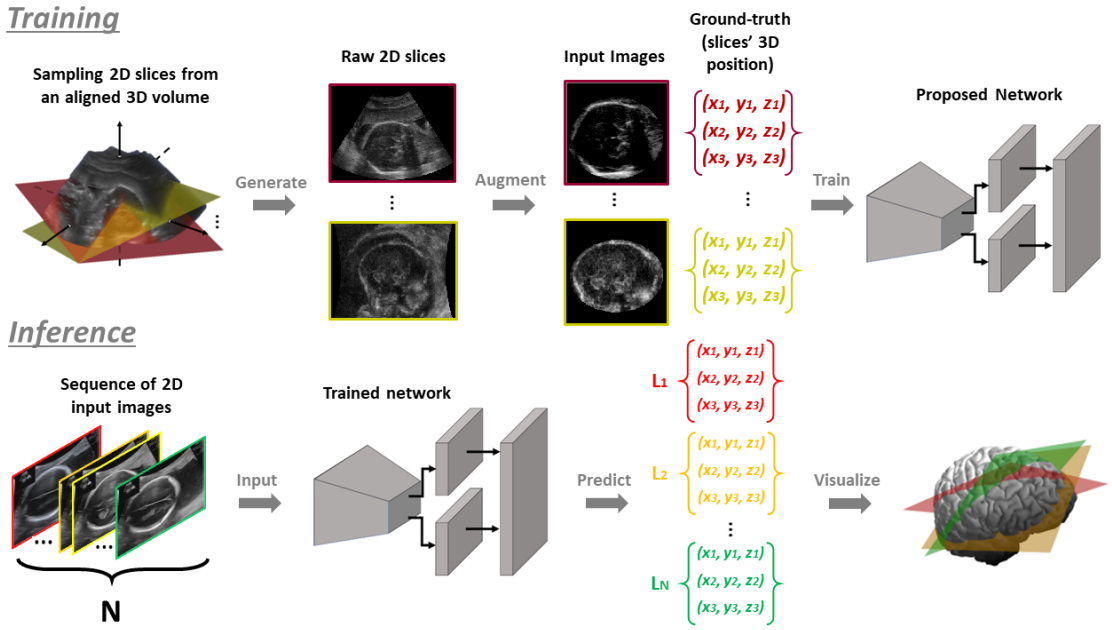


Figure 3.1: Pipeline of the proposed work. During training, 2D slices sampled from aligned 3D volumes are augmented and used to train the proposed ConvNet, `PlaneInVol`. The trained network can be used to predict the 3D location of arbitrary number of 2D images.

real 2D ultrasound images and videos with annotations from two experienced clinicians and medical professionals. The proposed model also outperforms the strong baseline [134, 135] and achieves comparable performance to human annotation.

3.1.1 Standard Planes Detection and Localization

In Chapter 2.6, the works about automated standard plane detection for 2D fetal ultrasound have been reviewed. Despite their effectiveness in detecting standard plane images, all of the above methods can only predict whether the image is acquired at a standard plane, but not the exact location of the image in the corresponding 3D space. Furthermore, a large amount of annotated data is required to train the model. Instead of training a classification model, in this chapter, a regression model is learned to predict the location of 2D ultrasound images of the fetal brain in a predefined 3D anatomical atlas. This is a more general task, which can be easily adapted to standard plane detection by simply identifying the

standard planes in the predefined 3D anatomical atlas. The model can further provide information about the relative position between the current plane and any standard or oblique planes of interest. Also, 2D images sampled from 3D volumes that are aligned to a predefined 3D anatomical atlas [138] are used so that the locations of images are automatically known and no further human annotation is needed, which is originally the major obstacles for training a localization network.

A slightly different task, namely standard plane localization in 3D volumes, has also been discussed in Chapter 2.6. Despite the excellent performance achieved by the reviewed methods, they require 3D volumes as the input, either directly or by having information extracted from the volume as a feedback during the localization process. This may limit their practical application because as reviewed in Chapter 2.2, most of the current standard clinical tests rely on only 2D ultrasound, and 3D ultrasound is not always available in many settings due to its cost and clinicians' preference [1]. On the other hand, the proposed method, `PlaneInVol`, just relies on 2D ultrasound images and it can be easily used with 2D ultrasound scanning to localize any standard or oblique planes of interest.

3.2 Methods

3.2.1 Training Data Generation

As introduced in Chapter 2.3.3, supervised learning requires paired training data in the form of $\{x_i, y_i\}$, where x_i is the input data point (*i.e.* 2D ultrasound image of fetal brain) and y_i is the label (*i.e.* the 3D location of the input image). Conventionally, the label is obtained by manual annotation, and the goal is usually to learn a function that maps the input sample x to a corresponding output label y. However, annotating the location of a random 2D ultrasound image of the fetal brain in the 3D space is very challenging. Therefore, 2D slices are artificially sampled from aligned 3D ultrasound volumes of the fetal brain, results in theoretically infinite number of data pairs $\{x_i, y_i\}$. Despite the volume alignment is semi-automatic (minimal effort is required from manual correction), the training for the proposed model resembles

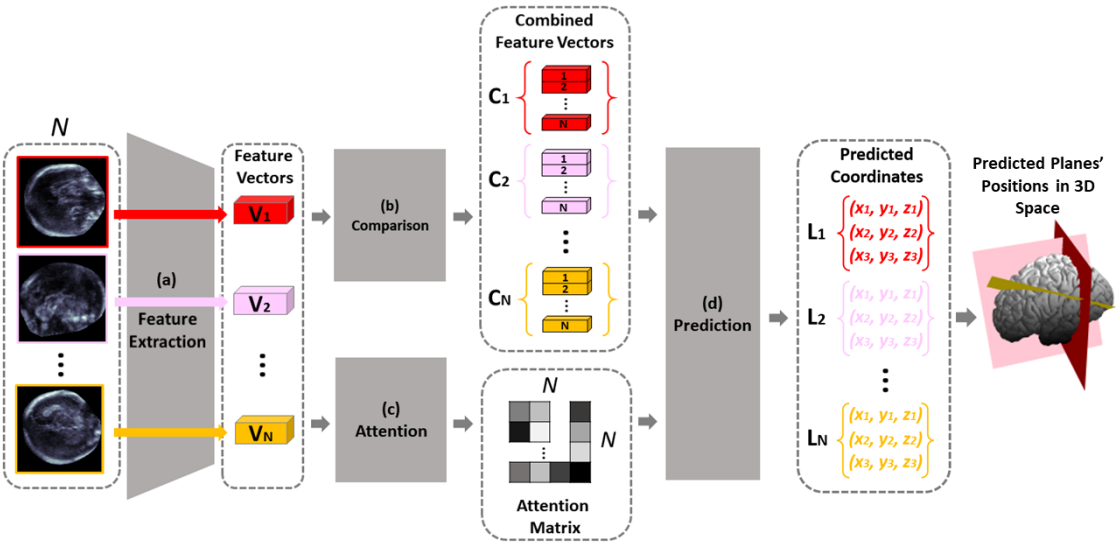


Figure 3.2: Overview of the proposed network, `PlaneInVol`. It consists of 4 sequential modules, namely (a) Feature Extraction, (b) Comparison, (c) Attention and (d) Prediction, which are represented by the grey blocks in the figure.

self-supervised learning, in the sense that training labels can be generated from the data itself. Three main steps are involved in generating the training data, $\{x_i, y_i\}$:

(I) The raw 3D volumes are firstly aligned to a predefined 3D anatomical atlas, $\mathbb{R}^3_{atlas}$, with the method proposed in [138], followed by a manual correction step. For every aligned 3D volume, $\mathbf{V} \in \mathbb{R}^{h \times w \times d}$, where h, w and d are the height, width and depth of the 3D volume respectively, there is an associated binary mask of skull, $\mathbf{B} \in \{0, 1\}^{h \times w \times d}$. The masks are generated by the ConvNet model proposed in [139].

(II) Following the sampling scheme adopted by [134, 135], 2D images and their corresponding 2D binary masks are sampled from the aligned 3D volumes, $\mathbf{V}$, and 3D binary mask, $\mathbf{B}$. In order to generate 2D images that are evenly distributed in a 3D volume, the surface normal of the sampling planes should be evenly spaced on the surface of a unit sphere [134], and this can be achieved by Fibonacci sphere sampling of polar coordinates, $p(\phi, \theta)$, where ϕ and θ are the azimuth and elevation angles respectively. Assuming m surface normals

are sampled, $\{\phi_i\}_{i=1}^m$ and $\{\theta_i\}_{i=1}^m$ can be calculated by:

$$\phi_i = \frac{2\pi(i-1)}{(\sqrt{5}+1)/2} \tag{3.1}$$

$$\theta_i = \cos^{-1}\left(\frac{2(1-i)}{m}\right) \tag{3.2}$$

By defining the surface normal by Eq. 3.1 and 3.2, the coordinate of the centre point of the sampling plane as well as the in-plane rotation (*i.e.* plane rotation about its surface normal), 2D images can be sampled from the 3D volumes.

(III) The sampled 2D images are randomly processed by one of the three proposed ways during training, namely (i) masking the 2D images by the convex hull of the associated sampled 2D binary masks to remove *most* of the extracranial contents, (ii) masking the 2D images by 2D circular masks with arbitrary size larger than the associated sampled 2D binary masks to remove *part* of the extracranial contents or (iii) not masking the 2D images at all to keep *all* the extracranial contents. While (i) and (ii) prevent the model from making predictions based on the background (*i.e.* extracranial structures) of the images, (iii) tries to minimize the influence of the shape and size of the binary masks, which are normally unavailable during inference, towards the prediction. Also, since 2D images are artificially sampled from 3D fetal brain volumes, they may look differently compared to the actual 2D images, in terms of resolution, intensity and noise. Extensive data augmentation is used to make the model more generalizable, including geometrical transformation, scaling, contrast modification, and addition of random noise.

These three pre-processing steps are only required during training, but not for inference when the trained network is being employed to actual 2D images.

3.2.2 Model Architecture

The input to the proposed network, `PlaneInVol`, is a set containing an arbitrary number of 2D images, $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}, \mathbf{I}_i \in \mathbb{R}^{h \times w}$, where N, h and w are the number of images, height and width of image, respectively. The output is the set of corresponding predicted locations $\mathcal{L} = \{\mathbf{L}_i\}_{i=1}^{N}$, where $\mathbf{L}_i \in \mathbb{R}^{3 \times 3}$, referring to the 3 Cartesian coordinates (*i.e.* x, y, z) of the 3 landmarks that define the predicted plane. Following the approach proposed in [134, 135], the centre, the bottom right and left corners of a plane are used as the landmarks to define the predicted plane. In order to simulate the motion of an actual ultrasound scan, during training, a constraint is imposed to the N input images such that the distance between two consecutively sampled slices (*i.e.* $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$) should be smaller than a predefined value, which is 20 pixels.

`PlaneInVol` consists of 4 sequential modules: *Feature extraction (Fig. 3.2a)*: a feature encoder (*i.e.* a shared ConvNet backbone) is used to generate a fixed-length feature vector, $\mathbf{v}_i$, to represent each input image. *Comparison (Fig. 3.2b)*: the feature vectors for each image are compared pairwise to compute the relationship between every pair of input images, which is further represented by the set of comparison feature vectors, $\{\mathbf{c}_{ij}\}_{i=1,j=1}^{N,N}$. *Attention (Fig. 3.2c)*: an attention mechanism is applied on the set of feature vectors to weight the contribution of each pairwise relationship. *Prediction (Fig. 3.2d)*: while generating a summarization feature vector of every input image for prediction of position in 3D space, the affinity matrix, is used to weight the comparison feature vectors. Each module is described in more detail below.

Feature extraction (Fig. 3.2a)

A feature extractor (*i.e.* a shared ConvNet backbone) is used to generate a fixed-length feature vector, $\mathbf{v}_i \in \mathbb{R}^{1 \times 512}$, for each input image, $\mathbf{I}_i$. A common feature encoder (*i.e.* shared weights) is used for all input images, such that the feature extraction is invariant to the permutation and number of input images. This

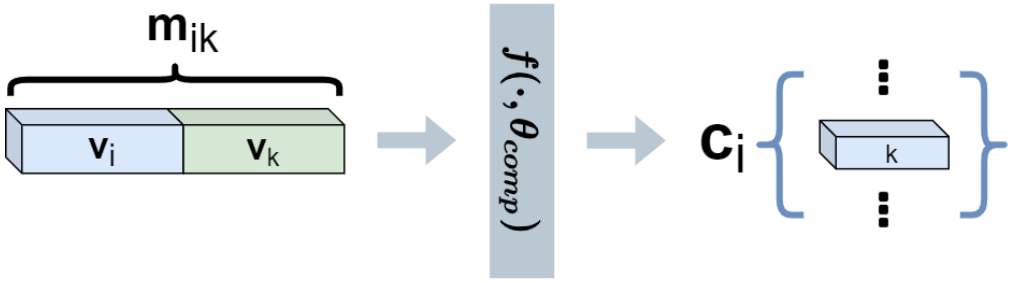


Figure 3.3: The processing unit of the *Comparison* module (Fig. 3.2b). A pair of feature vectors, $\mathbf{v}_i$ and $\mathbf{v}_k$, are concatenated and passed to the comparison network, $f(\cdot;\theta_{comp})$, to output a comparative feature vector, $\mathbf{c}_{ik}$.

is a desirable property for ultrasound images analysis due to the randomness of freehand image acquisition.

Here, the feature extractor, $e(\cdot;\theta_{feat})$, parameterized by θ_{feat}, is based on the VGG-16 network architecture [65]. With an arbitrary number, N, of 2D input images, $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}, \mathbf{I}_i \in \mathbb{R}^{h\times w}$, the output from this module is:

$$[\mathbf{v}_1,\ \mathbf{v}_2, ...,\ \mathbf{v}_N] = [e(\mathbf{I}_1;\theta_{feat}),\ e(\mathbf{I}_2;\theta_{feat}),\ ...,\ e(\mathbf{I}_N;\theta_{feat})] \tag{3.3}$$

where $e(\cdot;\theta_{feat})$ is shared between different $\mathbf{I}_i$.

Comparison (Fig. 3.2b)

The set of feature vectors, $\{\mathbf{v}_i\}_{i=1}^{N}$, is compared pairwise in this module. Instead of directly predicting the location of the image from its corresponding feature vector (*i.e.* each image position is predicted independently of all others) as proposed in [134, 135], it is proposed in this thesis that it will be beneficial for each input image to also consider its relative position with respect to other images, as all images are different planes of the brain of the same fetus and, hence, likely to be inter-correlated. This is achieved by combining the feature vector to generate a comparative feature vector, $\mathbf{c}_{ij}$, of every input image pair, $\mathbf{I}_i$ and $\mathbf{I}_j$.

This comparison is implemented in two steps, which are summarized by the processing unit as shown in Fig. 3.3. Firstly, concatenation between vector pairs is computed, which can be formally expressed as:

$$[\mathbf{m}_{11},\ \mathbf{m}_{12}, ...,\ \mathbf{m}_{NN}] = [(\mathbf{v}_1 \parallel \mathbf{v}_1),\ (\mathbf{v}_1 \parallel \mathbf{v}_2), ...,\ (\mathbf{v}_N \parallel \mathbf{v}_N)] \tag{3.4}$$

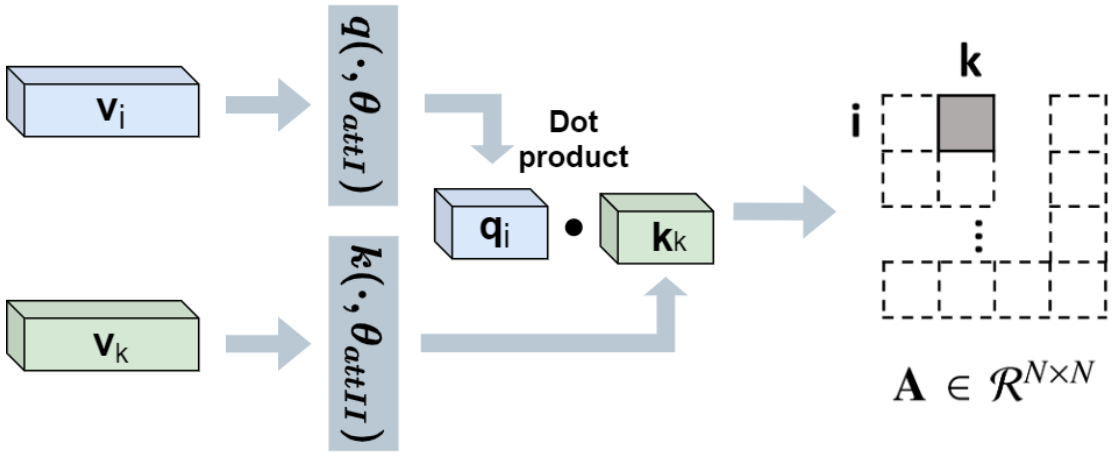


Figure 3.4: The processing unit of the *Attention* module (Fig. 3.2c). The dot product between a pair of embedded feature vectors, $\mathbf{q}_i$ and $\mathbf{k}_k$, gives rise to $\mathbf{A}_{ik}$.

where $\|$ is the concatenation operator and $\{\mathbf{m}_{ij}\}_{i=1,j=1}^{N,N}$, $\mathbf{m}_{ij} \in \mathbb{R}^{1\times1024}$ is the set of concatenated feature vectors.

Secondly, the set of concatenated feature vectors is passed as input to the comparison network, $f(\cdot;\theta_{comp})$, parameterized by θ_{comp}. The comparison network is a fully connected layer that merges the information of the two feature vectors into a comparative feature vector:

$$[\mathbf{c}_{11},\ \mathbf{c}_{12}, ...,\ \mathbf{c}_{NN}] = [f(\mathbf{m}_{11};\theta_{comp}),\ f(\mathbf{m}_{12};\theta_{comp}),\ ...,\ f(\mathbf{m}_{NN};\theta_{comp})] \quad (3.5)$$

where $\{\mathbf{c}_{ij}\}_{i=1,j=1}^{N,N}, \mathbf{c}_{ij} \in \mathbb{R}^{1\times512}$ is the set of comparative feature vectors.

Attention (Fig. 3.2c)

Different comparative feature vectors, $\mathbf{c}_{ij}$, may contribute differently to the final prediction of plane position. It is proposed to compute the relative contribution of each pairwise comparison by using an attention module [72]. The module will learn to assign more attention (*i.e.* a higher scalar weight) to comparisons with higher relational contribution and vice-versa. Contribution means the extent of any type of relationship, for example the similarity between a pair of images, which is related to the final prediction and hence can be learned by the model from the loss. The output of this attention module will be an affinity matrix, $\mathbf{A}$.

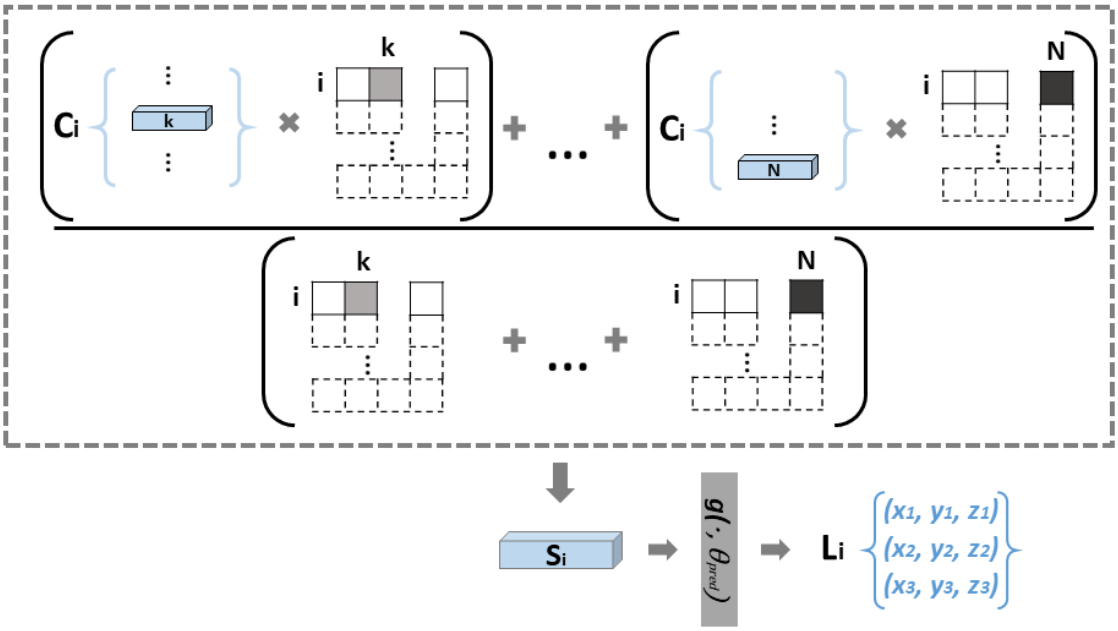


Figure 3.5: The processing unit of the *Prediction* module (Fig. 3.2d). The summarization feature vector, $\mathbf{s}_i$, is computed by $\{\mathbf{A}_{ik}\}_{k=1}^{N}$ and $\{\mathbf{c}_{ik}\}_{k=1}^{N}$. It is then processed by the prediction network, $g(\cdot;\theta_{pred})$, to output the set of predicted locations $\mathbf{L}_i$.

Fig. 3.4 displays the processing unit of the *Attention* module. To compute the affinity matrix, $\mathbf{A} \in \mathbb{R}^{N\times N}$, the dot product between pairs of feature vectors, $\{\mathbf{v}_i\}_{i=1}^{N}$, in an embedding space is computed as follows:

$$\mathbf{A}\,(i,j) = q(\mathbf{v}_i;\theta_{attI})k(\mathbf{v}_j;\theta_{attII})^T \tag{3.6}$$

where $q(\cdot;\theta_{attI})$ and $k(\cdot;\theta_{attII})$ are embedding networks (*i.e.* multilayer perceptrons), parameterized by θ_{attI} and θ_{attII}, respectively, that map the feature vectors into an embedding space, $\mathbb{R}^{1\times 256}$.

Prediction (Fig. 3.2d)

Fig. 3.5 shows the processing unit of the *Prediction* module. To compute the final prediction of each input image, the prediction module uses the affinity matrix, $\mathbf{A}$, to weight the comparative feature vectors, $\{\mathbf{c}_{ij}\}_{i=1,j=1}^{N,N}$, to compute a summarization feature vector, $\mathbf{s}_i \in \mathbb{R}^{1\times 512}$, for every input image, $\mathbf{I}_i$. The summarization feature vector, $\mathbf{s}_i$, gathers information from all images, weighted by the learned contribution

towards the prediction of $\mathbf{I}_i$ and is computed as follows:

$$\mathbf{s}_i = \frac{\sum_{j=1}^{N} \mathbf{A}(i,j)\mathbf{c}_{ij}}{\sum_{j=1}^{N} \mathbf{A}(i,j)} \tag{3.7}$$

The set of predicted locations $\{\mathbf{L}_i\}_{i=1}^{N}, \mathbf{L}_i \in \mathbb{R}^{3\times3}$, is obtained by passing the set of summarization feature vectors, $\{\mathbf{s}_i\}_{i=1}^{N}$, to the prediction network, $g(\cdot;\theta_{pred})$, parameterized by θ_{pred}:

$$[\mathbf{L}_1,\ \mathbf{L}_2, ...,\ \mathbf{L}_N] = [g(\mathbf{s}_1;\theta_{pred}),\ g(\mathbf{s}_2;\theta_{pred}),\ ...,\ g(\mathbf{s}_N;\theta_{pred})] \tag{3.8}$$

In summary, the predicted location, $\mathbf{L}_i$, of image, $\mathbf{I}_i$, is derived from $\mathbf{s}_i$ and hence the weighted sum of $\mathbf{c}_{ij}$ for all j. In other words, when predicting the location of image $\mathbf{I}_i$, information of all images within the same space, $\{\mathbf{I}_j\}_{j=1}^{N}$, will be considered. Furthermore, their relative contribution and degree of relationships with $\mathbf{I}_i$ will be taken into account by the affinity matrix, $\mathbf{A}$.

3.2.3 Loss Function

During training, the mean least-square error is applied as the loss function ($\mathfrak{L}$):

$$\mathfrak{L}\left(\hat{\mathbf{L}}, \mathbf{L}\right) = \frac{1}{N}\sum_{i\ =\ 1}^{N} \left(\hat{\mathbf{L}}_i - \mathbf{L_i}\right)^2 \tag{3.9}$$

where $\hat{\mathbf{L}}$ and $\mathbf{L}$ are the ground-truth and predicted locations, respectively.

3.3 Experiment

3.3.1 Dataset

The 3D ultrasound fetal brain volumes ($160 \times 160 \times 160$ voxels at a resolution of $0.6 \times 0.6 \times 0.6$ mm^3) were obtained as part of the INTERGROWTH-21st study [140], which were collected using a Philips HD9 curvilinear probe at a 2–5 MHz wave frequency. For each 3D volume, there is at least one associated 2D image taken at the standard TT plane routinely used for biometric and structural assessment.

Experiment \ Dataset	INTERGROWTH (3D Volumes)	INTERGROWTH (2D TT Plane Images)	Video of Freehand Brain Scanning
Training	✓(21 weeks & 18-22 weeks)		
Testing			
Sections 3.3.4 and 3.4.1	✓(21 weeks)		
Sections 3.3.4 and 3.4.1	✓(18-22 weeks)		
Sections 3.3.5 and 3.4.2	✓		
Sections 3.3.6 and 3.4.3		✓	
Sections 3.3.7 and 3.4.4			✓
Sections 3.3.8 and 3.4.5	✓(21 weeks)		

Table 3.1: Summary of different experiments and the corresponding dataset.

Both the 2D images and 3D volumes were acquired following strict requirements to ensure that the image quality satisfied pre-defined criteria [141]. For instance, the fetal skull occupied at least 50% of the image, and the image was not affected by fetal or maternal movements. Fetal anomaly ultrasound scan is recommended to be undertaken between 18 to 21 gestational weeks and some flexibilities are allowed for this age range [3]. In this study, images were selected from fetuses with gestational age ranging from 18 to 22 gestational weeks. Each image was masked and aligned to a coordinate space as described in Section 3.2.1. A summary of training and different experiments and their corresponding dataset is presented in Table 3.1.

3.3.2 Training Details

In this study, a baseline model (only slight modification based on network architecture proposed by [134, 135]) was re-implemented to compare its performance to the proposed model, `PlaneInVol`. The exact network architectures of the baseline model and `PlaneInVol` are presented in Table 3.2. Optimization was achieved using the ADAM algorithm [68] with mini-batches of size 32. The initial learning rate was set to 10^{-4}, which was decreased by half when errors plateaued.

Fifty and fifteen 3D volumes acquired at 21 gestational weeks were selected for training and validation, respectively. For each 3D volume in each training epoch, 50 evenly distributed normals were sampled using the Fibonacci Sphere Sampling

Module	Baseline Model	Proposed Model	Output Size
Feature Extraction	Input Layer		$N \times 160 \times 160 \times 1$
	conv, 3×3, 64 conv, 3×3, 64 max pool, 2×2, stride 2		$N \times 80 \times 80 \times 64$
	conv, 3×3, 128 conv, 3×3, 128 max pool, 2×2, stride 2		$N \times 40 \times 40 \times 128$
	conv, 3×3, 256 conv, 3×3, 256 conv, 3×3, 256 max pool, 2×2, stride 2		$N \times 20 \times 20 \times 256$
	conv, 3×3, 512 conv, 3×3, 512 conv, 3×3, 512 max pool, 2×2, stride 2		$N \times 10 \times 10 \times 512$
	conv, 3×3, 512 conv, 3×3, 512 conv, 3×3, 512 max pool, 2×2, stride 2		$N \times 5 \times 5 \times 512$
	Fully Connected Layer		$N \times 512$ $(\{\mathbf{v}_i\}_{i=1}^N)$
Comparison	-	Pairwise Feature Concatenation	$N \times N \times 1024$ $\{\mathbf{m}_{ij}\}_{i=1,j=1}^{N,N}$
	-	Fully Connected Layer	$N \times N \times 512$ $\{\mathbf{c}_{ij}\}_{i=1,j=1}^{N,N}$
Attention	-	Embedding Networks $\times$ 2 (*i.e.* Fully Connected Layers)	$N \times 256$ $(q(\mathbf{v}_i; \theta_{attI}))$ $N \times 256$ $(k(\mathbf{v}_j; \theta_{attII}))$
	-	Dot Product	$N \times N$ $(\mathbf{A})$
Prediction	-	Weighted Average	$N \times 512$ $(\{\mathbf{s}_i\}_{i=1}^N)$
	Fully Connected Layer		$N \times 9$ (Resize to $N \times 3 \times 3$, $\{\mathbf{L}_i\}_{i=1}^N$)

Table 3.2: Network architectures of the baseline model and the proposed model, `PlaneInVol`. For the feature extraction module and the final layer of the prediction module, the baseline model and `PlaneInVol` have the same architecture, but they do not share weights (*i.e.* they are trained separately).

method as described in Section 3.2.1. Along each normal, 15 planes perpendicular to the normal, with average spacing of 2.4 mm were chosen (Fig. 3.10a). For each plane, four 2D slices (160×160 pixels), with random in-plane rotation were sampled. Therefore, in total, there were $150,000$ and $45,000$ images for each

training and validation epoch, respectively.

Since an infinite number of different 2D slices can be sampled from a 3D volume in principle, this feature was used to introduce random variation to the sampling parameters for each training epoch. Therefore, the $150,000$ training images were expected to be different for every epoch. This kept the number of training data for each training epoch relatively small as compared to [134, 135], while the number of different images used for the whole training was much larger. This was regarded as a type of data augmentation, which may prevent the model from overfitting while having a reasonable amount of varied training data for each epoch.

3.3.3 Evaluation metrics

Three evaluation metrics were used to evaluate and compare the performance of the models. First, Euclidean distance (ED) between all the coordinates of the predicted and ground-truth planes is computed as follows:

$$\mathrm{ED}(\hat{\mathbf{P}}, \mathbf{P}) \;=\; \frac{\sum_{i=1,\ j=1}^{h,\ w} dist(\hat{\mathbf{p}}_{ij},\ \mathbf{p}_{ij})}{h \cdot w} \tag{3.10}$$

where $\hat{\mathbf{P}}$ and $\mathbf{P}$ are the predicted and ground-truth planes and $dist(\hat{\mathbf{p}}_{ij},\ \mathbf{p}_{ij}) = |\hat{\mathbf{p}}_{ij} - \mathbf{p}_{ij}|^2$ computes the Euclidean distance between the two points, $\hat{\mathbf{p}}_{ij}$ and $\mathbf{p}_{ij}$, where $\hat{\mathbf{p}}_{ij}$ and $\mathbf{p}_{ij}$ are the (x, y, z) coordinates of the pixel ij on $\hat{\mathbf{P}}$ and $\mathbf{P}$ respectively.

Secondly, plane angle (PA) between the predicted and ground-truth planes are computed as follows:

$$\mathrm{PA} = cos^{-1}(\hat{\mathbf{n}} \cdot \mathbf{n}) \tag{3.11}$$

where $\hat{\mathbf{n}}$ and $\mathbf{n}$ are the surface normals of the predicted and ground-truth planes, respectively. Smaller ED and PA suggest that the ground-truth and predicted planes locate closely to each other, which may represent more accurate prediction.

Thirdly, normalized cross-correlation (NCC) [142] between the input image and image sampled from the predicted plane is computed. Larger values suggest higher similarity between the two images and more accurate prediction of plane position.

3.3.4 Comparison with Baseline Model

Images sampled from 3D volumes were used to quantitatively evaluate the performance of different models. The proposed model, `PlaneInVol`, and the baseline model were compared using the evaluation metrics introduced in Section 3.3.3. In addition, in order to investigate the individual contribution of the newly proposed modules, namely the *Comparison* module (Section 3.2.2) and the *Attention* module (Section 3.2.2), ablation study has been conducted by removing the *Attention* module of the proposed network and applying equal weighting to every comparative feature vector, $\mathbf{c}_{ij}$ (*i.e.* replacing the affinity matrix, $\mathbf{A}$, with a matrix of ones).

Sensitivity to Input Image Support

Since `PlaneInVol` makes a prediction for each image by grouping information of all input images, prediction accuracy may be sensitive to the number of input images. Therefore, different numbers, $N \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, of input images were tested to investigate on how changing the number of input images may affect the prediction of the proposed model.

Application to Broader Gestational Age Range

Trained on images at 21 gestational weeks, images within a broader gestational age range (*i.e.* 18 to 22 gestational weeks) were tested to evaluate the generalizability of the models to different ages. For different ages, a slight change of brain anatomical structure is expected [41].

As a comparison, images of the whole gestational age range (*i.e.* 18 to 22 gestational weeks) were used to train a different set of models to verify if a single model can be used on a broad gestational age range.

3.3.5 Relationship between Plane Location and Accuracy of Prediction

In Section 3.3.4, 2D images have been sampled at different locations in each 3D volume. The results of the images sampled from the 15 fetal brain volumes in Section 3.3.4 were further analyzed to investigate how accurate the proposed model, `PlaneInVol`, is in predicting images in different regions of the 3D brain. Specifically, the accuracy of prediction of images sampled along different directions and at different distances from the centre of the brain were studied. Fig. 3.10a shows how planes were sampled from the 3D volume along one normal using the Fibonacci Sphere Sampling method as described in Section 3.2.1.

3.3.6 Real 2D Image Acquisition of Standard TT Plane

Real 2D images taken at the standard TT plane were tested. These images were acquired with the 15 3D test volumes in Section 3.3.4. For each 2D image, plane location was predicted by `PlaneInVol` and annotated by 2 individual experts separately. Using the predicted plane locations, the corresponding 2D images were sampled from the associated 3D volume. Variations, measured by the evaluation metrics introduced in Section 3.2.1, were estimated between the 3 different sets of predictions and annotations. They were further analyzed by one-way analysis of variance (ANOVA) to verify whether the difference between them is statistically significant or not.

3.3.7 Video of Freehand Fetal Brain Scanning

In addition to the single standard plane images as described in Section 3.3.6, 4 videos of 2D scans acquired from 4 subjects with gestational age between 19 to 21 weeks during fetal exams of the brain were also tested. The videos were acquired by sweeping the ultrasound probe along different directions during scanning. Therefore, the videos were composed of 2D views corresponding to different locations of the

fetal brain, which may or may not be a standard plane. Every video was treated as a set of 2D images for testing. Using the predicted plane locations, 2D slices were sampled from the 3D atlas volume. The video frames and the corresponding sampled 2D atlas slices were compared qualitatively, in terms of structures present and image orientation.

3.3.8 Impact of Learned Attention

As mentioned in Section 3.2.2, the affinity matrix, $\mathbf{A}$, weights the contribution of each pairwise comparison of the set of input images. To verify that the *Attention* module (Fig. 3.2c) actually learns to assign meaningful weights, the results of the slices sampled from the 15 fetal brain volumes in Section 3.3.4 were further analyzed. Using $N = 4$ input images (for easier comparison and visualization), the normalized attention, $\frac{\sum_{j=1}^{4} \mathbf{A}(i,j)}{\sum_{i=1}^{4}\sum_{j=1}^{4} \mathbf{A}(i,j)}$, associated to each input image was investigated.

3.4 Results

3.4.1 Comparison with Baseline Model

This section includes the results of the two experimental settings (Section 3.3.4 and 3.3.4). For both settings, all three evaluation metrics indicated that the performance of the proposed models surpassed that of the baseline model.

For each 3D volume, 3000 2D images were sampled in the same way as described in Section 3.3.2. Two settings were investigated, namely variation on number of input image (Section 3.3.4) and generalization to a broader gestational age range (Section 3.3.4).

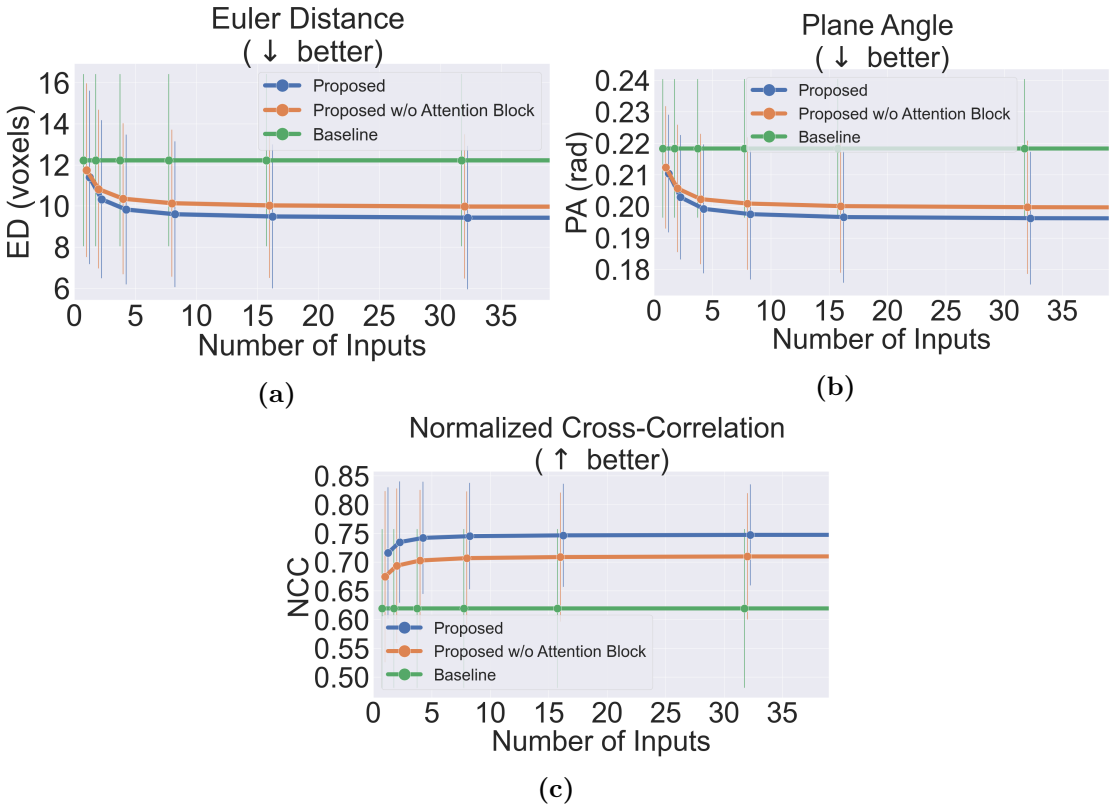


Figure 3.6: Results of sensitivity to input image support. The accuracy of the baseline model (green), the proposed model, `PlaneInVol` (blue), and the proposed model without *Attention* module (orange) is shown. The three graphs show the mean results (± standard deviation) of (a) Euclidean distance, (b) plane angle and (c) normalized cross correlation between groundtruth and prediction for different numbers of input images.

Sensitivity to Input Image Support

Fifteen 3D fetal brain volumes with gestational age of 21 gestational weeks were used for evaluation, yielding at total of 45,000 2D test images. Different numbers, $N \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, of input images were tested.

The results of this experiment are presented in Figs. 3.6a to 3.6c. Since the number of images would not affect the prediction of the baseline model, results of the baseline model were the same for different number of input images.

For the proposed models, both with and without the *Attention* module, performance increased with the number of input images by as much as 17%, 7% and 5% as indicated by ED, PA and NCC, respectively. This may be reasonable because the proposed models make a prediction for each image by grouping information of all input images. More input images may provide more information for the prediction.

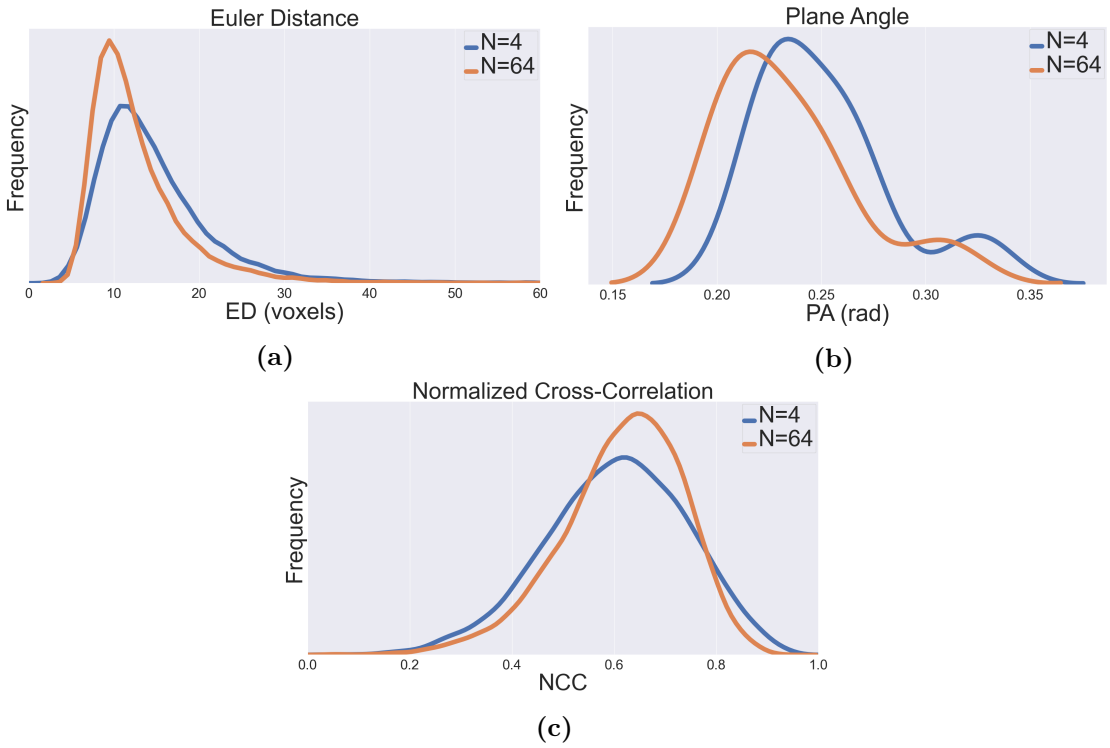


Figure 3.7: The result distribution of `PlaneInVol`. Result distribution of (a) Euclidean distance, (b) plane angle and (c) normalized cross correlation at 21 gestational weeks with $N = 4$ and $N = 64$ are shown.

Also, the ablation study suggested that the *Comparison* and *Prediction* modules, which are responsible for grouping information of all input images, may primarily lead to improvement when compared to the baseline model by around 19% (ED), 8% (PA) and 15% (NCC). The addition of the *Attention* module, which assigns weights to the grouping of information, contributed to further improvement by an extra 5% (ED), 2% (PA) and 5% (NCC). Although such further improvement may appear to be marginal as shown in Figs. 3.6a to 3.6c, it is statistically significant for every number of input images and evaluation metric concerned ($p<0.05$, t-test).

In addition, all three evaluation metrics showed that the performance of the proposed models surpassed that of the baseline model by as much as 23% (ED), 11% (PA) and 21% (NCC) and when the number of input images increased, the improvement was more significant. It was observed that the gain in accuracy nearly saturated when the number of inputs exceeds 32 and therefore in Figs. 3.6a to 3.6c, the results for $N \in \{64, 128\}$ were omitted for clearer visualization. The result

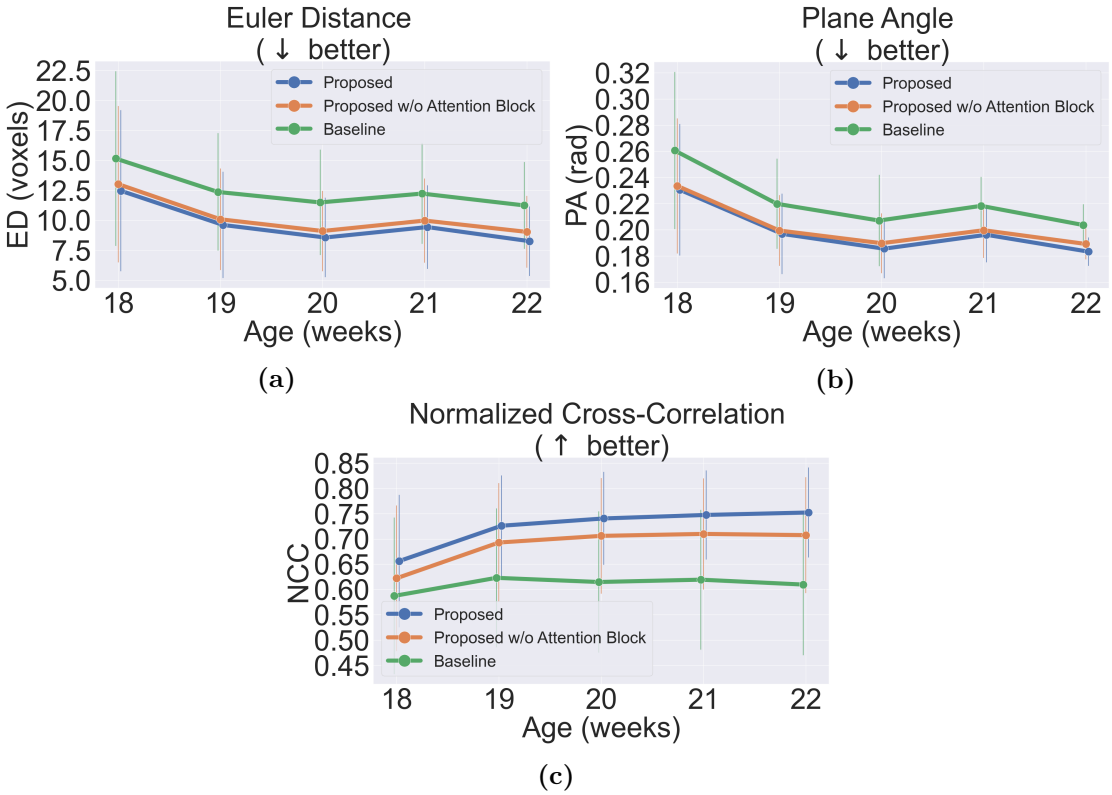


Figure 3.8: Results of application to broader gestational age range. The accuracy of the baseline model (green), the proposed model, `PlaneInVol` (blue), and the proposed model without *Attention* module (orange) is shown. The three graphs show the mean results (± standard deviation) of (a) Euclidean distance, (b) plane angle and (c) normalized cross correlation between groundtruth and prediction for different gestational ages.

distribution of $N = 4$ and $N = 64$ is further displayed in Fig. 3.7, which shows that although increasing the number of input images may not have a significant impact on reducing outliers, it shifted the distribution towards better performance.

Application to Broader Gestational Age Range

Fetal brain volumes with gestational age of 18 gestational weeks (50 volumes), 19 gestational weeks (34 volumes), 20 gestational weeks (57 volumes), 21 gestational weeks (15 volumes) and 22 gestational weeks (9 volumes) were used for testing in this experiment.

The results of the first part of this experiment are summarized in Figs. 3.8a to 3.8c. Using models trained on images with gestational age of 21 weeks, The

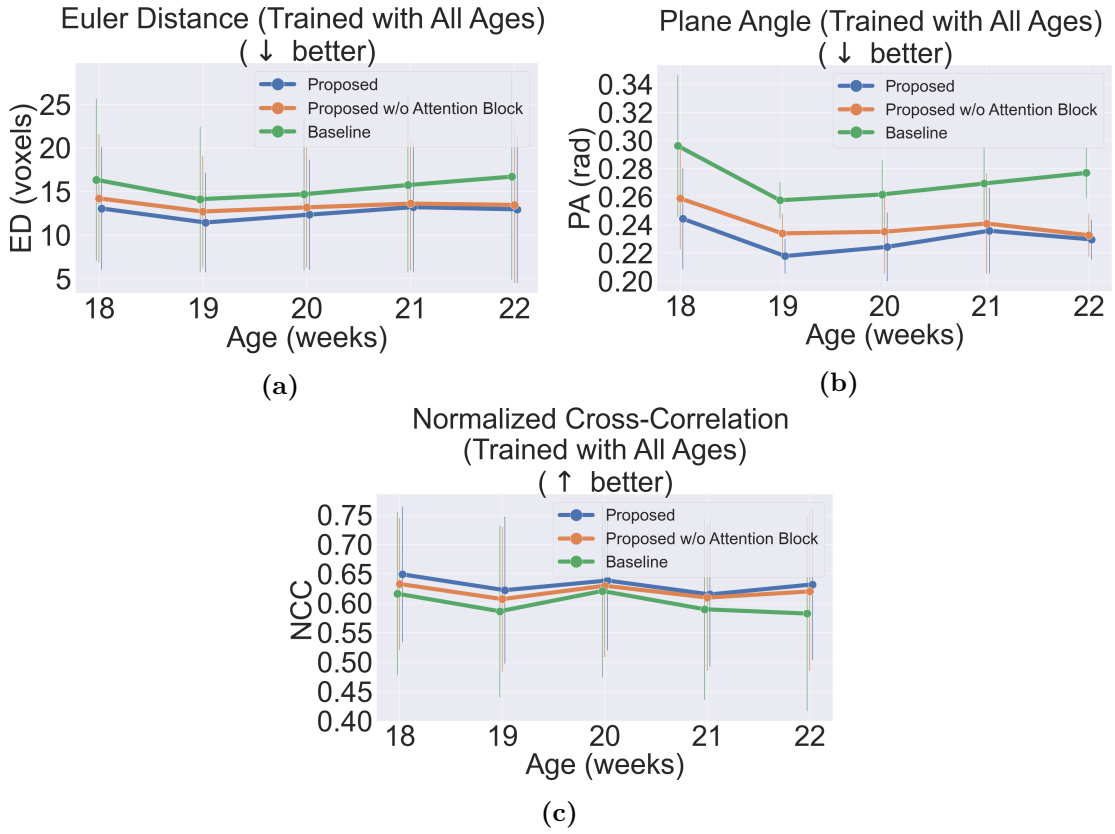


Figure 3.9: Results of application to broader gestational age range by models trained with all gestational ages. The accuracy of the baseline model (green), the proposed model, `PlaneInVol` (blue), and the proposed model without *Attention* module (orange) is shown. The three graphs show the mean results (± standard deviation) of (a) Euclidean distance, (b) plane angle and (c) normalized cross correlation between groundtruth and prediction by models trained with images of all gestational ages for different gestational ages.

models were tested on images with gestational age ranging from 18 to 22 weeks.

Two observations can be obtained: firstly, for all ages, predictions made by the proposed models were more accurate than those made by the baseline model by as much as 23% (ED), 11% (PA) and 21% (NCC). Also, predictions made by the proposed model without the *Attention* module were slightly less accurate than the complete version of the proposed model, `PlaneInVol`. The slight improvement caused by the incorporation of the *Attention* module is statistically significant for every age and evaluation metric concerned ($p<0.05$, t-test). Secondly, it was observed that in general, predictions on images at younger gestational ages were less accurate by as much as 51% (ED), 26% (PA) and 13% (NCC). A potential explanation is that fetuses during the second trimester are undergoing rapid neuro-

development [41]. Therefore, brain structures of fetuses at younger gestational ages may look quite different from those of fetuses of gestational age of 21 weeks, which are the images that the models were trained on.

The results of the second part of this experiment are summarized in Figs. 3.9a to 3.9c, where the models have been trained and tested on images of the whole gestational age range (*i.e.* 18 to 22 gestational weeks). When compared to the results of models trained on images of a single age, two observations can be obtained: firstly, predictions made by the proposed models were more accurate than those made by the baseline model by as much as 21% (ED), 16% (PA) and 9% (NCC). Also, predictions made by the proposed model without the *Attention* module were slightly less accurate than the complete version of the proposed model, `PlaneInVol`. The slight improvement caused by the incorporation of the *Attention* module is statistically significant for every age and evaluation metric concerned ($p<0.05$, t-test). Secondly, predictions made by models trained on images of the whole gestational age range were less accurate when compared to those made by models trained on images of just 21 weeks. This may be reasonable because images of different gestational ages were registered to different atlases as brain structures presented at different gestational ages may look quite different. For a single age, every plane location in the atlas space corresponds to a unique set of 2D image features. However, when a single model is trained with images of different gestational ages, it is equivalent to combining different unique atlas spaces into one and every plane location in this combined atlas space corresponds to multiple sets of 2D image features, each belongs to a specific age and hence they can be quite different to each other. This may be a more difficult and ambiguous learning task when compared to training models on images of a single age. Therefore, one single model trained on images of a broad gestational age range may have poorer performance when compared to models trained on a single age.

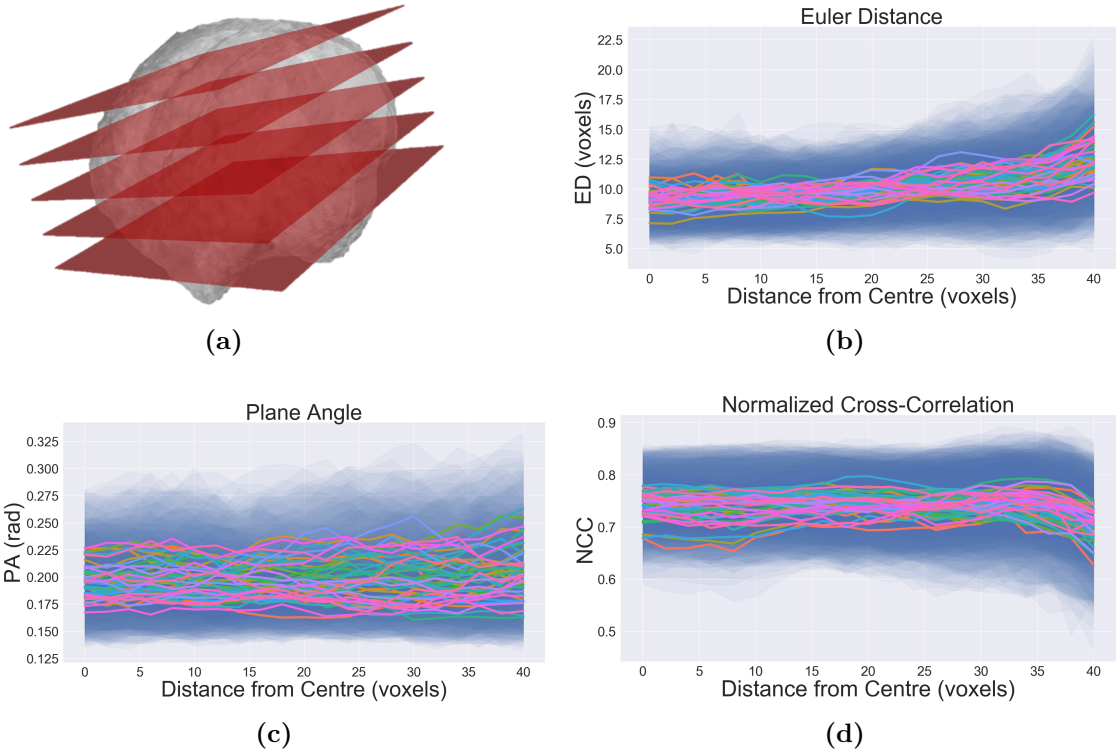


Figure 3.10: Plane location and accuracy of prediction. (a) shows how planes were sampled from the 3D volume along one normal of the unit sphere. Mean results (± standard deviation) of (b) Euclidean distance, (c) plane angle and (d) normalized cross correlation between groundtruth and prediction for images sampled from different locations of the 3D brain volumes are computed. Each curve in the figure indicates the mean results of one normal of the unit sphere and the blue shadow around it is the standard deviation of the results. Slices perpendicular to it and at different distance away from the centre of the 3D brain volumes were sampled and tested.

3.4.2 Relationship between Plane Location and Accuracy of Prediction

The results of finding the relationship between plane location and accuracy of prediction are presented in Fig. 3.10. Similar to the sampling procedure as introduced in Section 3.3.2, for each 3D volume, 50 normals evenly distributed on the unit sphere were chosen and each of them was represented by a colored curve in Fig. 3.10. In Fig. 3.10, values on each colored curve indicate the mean results, while the blue shadow around the curve is the standard deviation of the results. Along each normal and at different distance away from the centre of the 3D brain volumes, planes perpendicular to the normal were sampled.

Firstly, suggested by all three evaluation metrics, the performance of the proposed

	ED (voxels)	PA (rad)	NCC
M1 *v.s.* M2	9.12 ± 4.01	0.126 ± 0.055	0.867 ± 0.093
M1 *v.s.* Model	11.36 ± 3.26	0.179 ± 0.095	0.841 ± 0.096
M2 *v.s.* Model	11.44 ± 5.02	0.180 ± 0.120	0.837 ± 0.080
P value (one-way ANOVA)	0.257	0.227	0.639

Table 3.3: Comparison with manual annotation on real 2D images taken at the standard TT plane. Mean results (± standard deviation) and one-way ANOVA results between first manual annotation (M1), second manual annotation (M2) and prediction by the proposed model, `PlaneInVol`, are displayed. P values of the one-way ANOVA suggests the comparable performance by `PlaneInVol` and human annotations.

model in predicting images sampled along different directions (*i.e.* different lines in Fig. 3.10) were similar. The Euler distance, plane angle and normalized cross correlation were around 10 voxels, 0.20 rad and 0.75 respectively, which were similar to the overall result presented in Figs. 3.8a to 3.8c. In other words, the performance of `PlaneInVol` does not depend on the geometric orientation of the images sampled, which is desirable because during 2D freehand ultrasound scanning, images along different directions may be acquired.

Secondly, as suggested by ED (Fig. 3.10b) and NCC (Fig. 3.10d), when the images were farther away from the centre of the 3D brain volumes, the accuracy of the prediction dropped. This is reasonable because in general, images farther away from the centre, especially those near the edge of the brain, contain fewer indicative structures and hence are less informative and it is more difficult to predict their 3D location [134, 135].

3.4.3 Real 2D Image Acquisition of Standard TT Plane

Real 2D images taken at the standard TT plane were tested. Table 3.3 summarizes the variations between the plane locations predicted by the proposed model, `PlaneInVol`, and manually annotated by 2 individual experts. Although the mean values of the three evaluation metrics may suggest that the variation between the

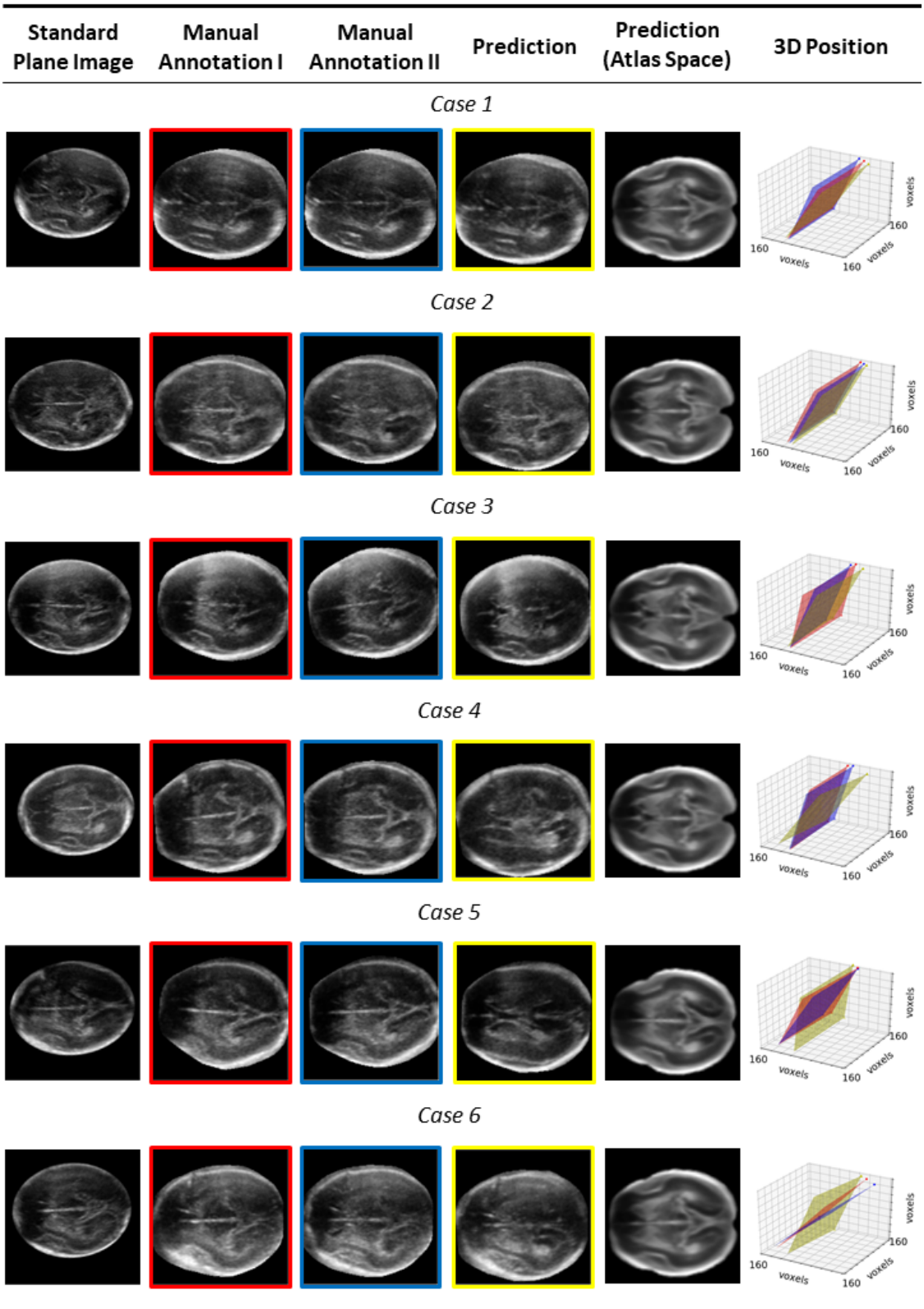


Figure 3.11: Visualization of manual annotation comparison. Six examples of native and masked 2D scans taken at the standard TT plane (*first column*); slices sampled from the corresponding 3D volume using the first manual annotation (*second column*), second manual annotation (*third column*) and `PlaneInVol`'s prediction (*fourth column*); slices sampled from the 3D atlas using `PlaneInVol`'s prediction (*fifth column*) and the position of the aforementioned slices in the 3D atlas space (*sixth column*). Frame color of the images (*second to fifth column*) corresponds to the planes as shown in the 3D atlas space (*sixth column*).

two sets of manual annotation is smaller than that between the model prediction and the manual annotations, p values of 0.257, 0.227 and 0.639 as calculated by the one-way ANOVA pointed out that it was failed to reject the null hypothesis, and there is no difference between the three groups of comparison, suggesting the comparable performance by `PlaneInVol` and human annotations.

It is understandable that the result obtained by the one-way ANOVA may not be convincing enough due to the limited amount of test images. Therefore, the 15 test images were further analyzed independently. While for most cases, the model prediction closely matched both (case 1 and 2 in Fig. 3.11) or either (case 3 in Fig. 3.11) set(s) of the manual annotation, it was found out that only three cases exhibited significant difference (*i.e.* more than 30% difference) between the model prediction and both sets of the manual annotation. They are presented as cases 4 to 6 in Fig. 3.11. It is evident that both the appearance (*fourth column*) and 3D location (*sixth column*) of the sampled slices using the proposed model's prediction differ significantly with those sampled from the manual annotations (*second* and *third column*). However, the slices sampled from the 3D atlas using the prediction by `PlaneInVol` (*fifth column*) actually look much more similar to the input standard plane image (*first column*) than the slices sampled from the 3D volume (*fourth column*). In other words, the large variation between the model prediction and the manual annotations in these three cases is mainly due to the misalignment between the three volumes and the atlas. The three volumes were checked again and it was verified that the poor volume quality makes perfect alignment to the atlas space extremely challenging.

3.4.4 Video of Freehand Fetal Brain Scanning

Fig. 3.12 shows the results of four video examples. It can be observed that the video frames and the corresponding slices sampled from the atlas present similar anatomical structures in the same orientation. Also, the predicted plane locations generally match with the motion of the probe when acquiring the videos, which

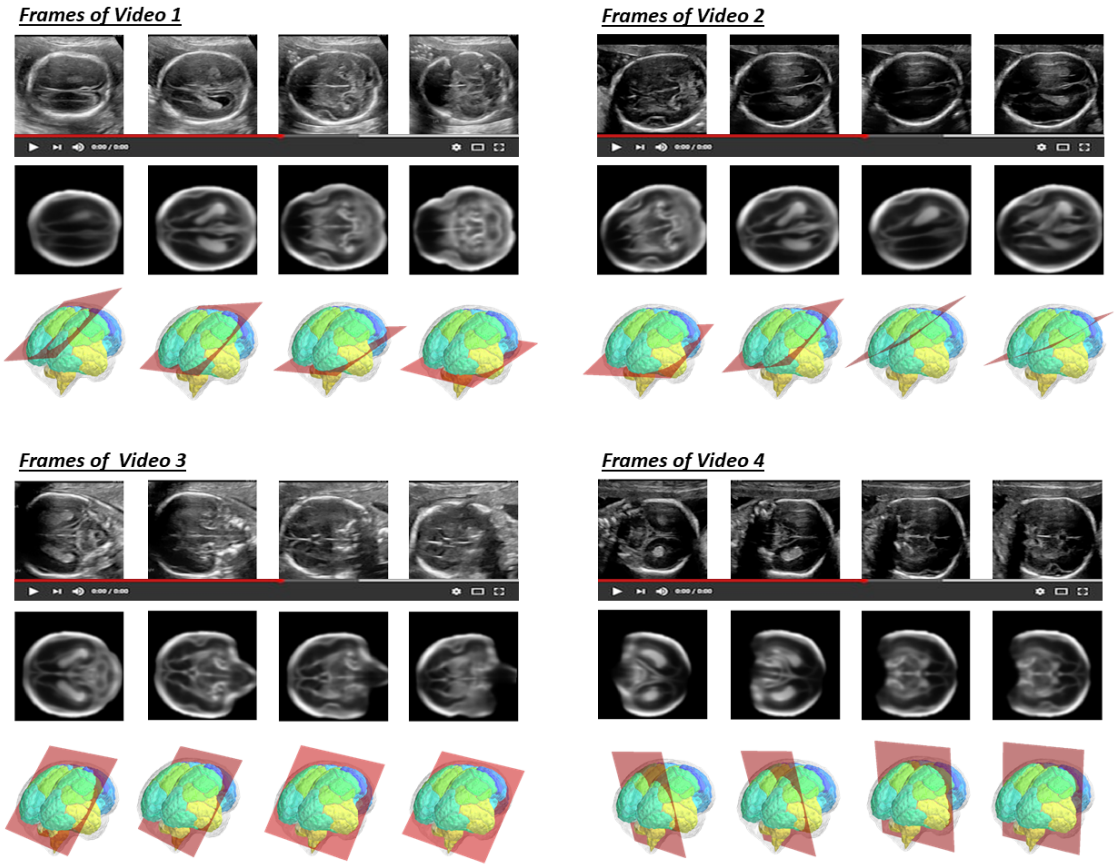


Figure 3.12: Results of four video examples. For each example, the upper row shows multiple frames of the video, which were input to the proposed model, `PlaneInVol`. Using the predicted plane locations, corresponding slices were sampled from the 3D atlas, which were shown in the middle row. The prediction plane location of each input video frame in the 3D atlas was displayed in the bottom row.

were roughly along the longitudinal axis (videos 1 and 2 of Fig. 3.12) and the sagittal axis (videos 3 and 4 of Fig. 3.12), respectively.

It may be noted that the anatomical structures in the upper hemisphere of some of the video frames (*e.g.* video 2 of Fig. 3.12) are not clearly discernible. This is due to the interaction between the ultrasound wave and concave fetal skull, which results in the anatomical structures presented in the hemisphere near the ultrasound probe (*i.e.* upper part of the video frames) generally being less visible [46]. On the other hand, the atlas represents both hemispheres. Therefore, the upper hemisphere of some of the video frames and that of their corresponding atlas slices may look different.

It was demonstrated in Section 3.4.2 that the accuracy of prediction decreases

when the input 2D image locates farther away from the centre of the brain. Fig. 3.13 shows two sets of consecutive frames which capture the external areas of the supratentorial region of the fetal brain. It can be observed that the slices sampled from the 3D atlas using the predicted plane location show completely different structures from their corresponding input frames and the predicted locations for consecutive frames do not show a smooth transition, which both further verify that the performance of `PlaneInVol` would decline when the input 2D images capture areas farther away from the centre of the brain, which present very limited structural features.

3.4.5 Impact of Learned Attention

The results of the slices sampled from the 15 fetal brain volumes in Section 3.4.1 were further analyzed to verify that the *Attention* module (Fig. 3.2c) actually learns to assign meaningful weights. Fig. 3.14a shows that the learned attention decreases with the increasing sampled slices' distance from the centre of the fetal brain. If the learned attention is interpreted as the weighting of contribution of the pairwise comparison of input images, as mentioned in Section 3.2.2, Fig. 3.14a may verify that the *Attention* module of the proposed model, `PlaneInVol`, actually learns to assign meaningful weights, because in general, regions closer to the centre of are more likely to contain richer structural information, and hence more indicative towards the final prediction. This can be visualized in Figs. 3.14b and 3.14c, where the images with blue and gray frames (*i.e.* sampled farther away from the centre) present less indicative structural information than the images with red and green frames (*i.e.* sampled closer to the centre). Therefore, the attention weight assigned to the images sampled farther away from the centre of the 3D volume, which quantify their degree of contribution towards the final prediction, is smaller in general.

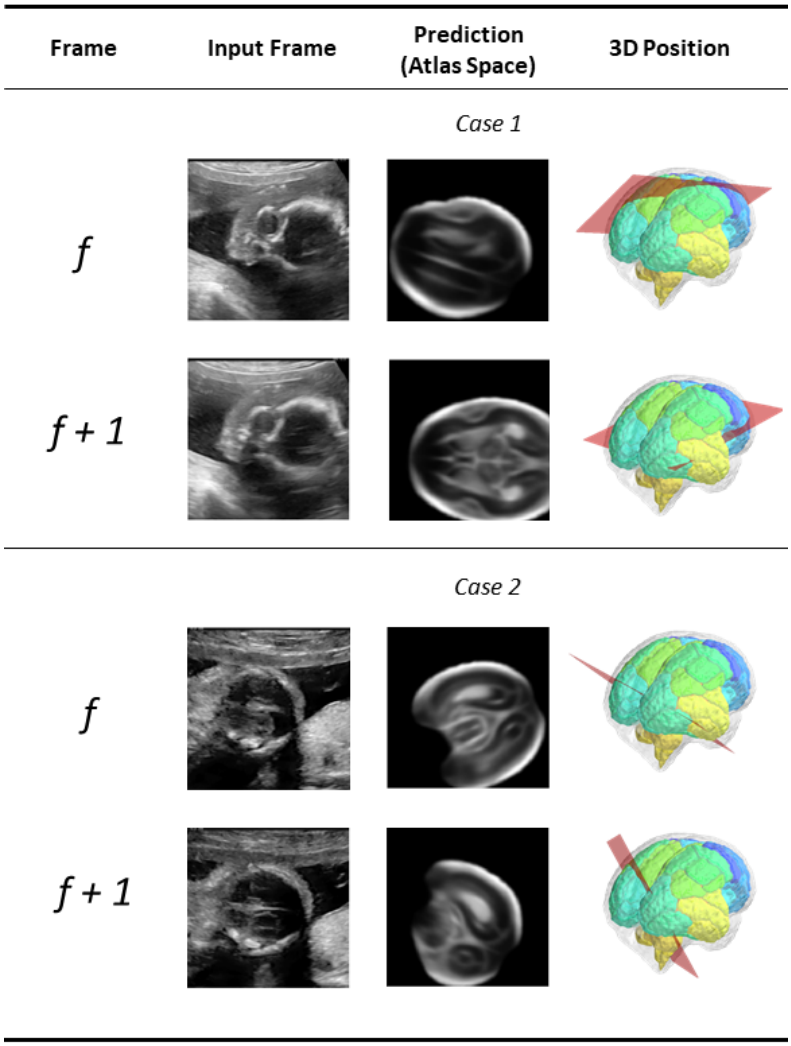


Figure 3.13: Examples of suboptimal prediction. Predicted locations of consecutive frames, which capture the external areas of the supratentorial region of the fetal brain, are completely different and do not show a smooth transition.

3.5 Conclusion

In this chapter, `PlaneInVol` is proposed to predict the position of 2D ultrasound fetal brain scans in 3D atlas space. Instead of purely supervised learning that requires heavy annotations for each 2D scan, the model is trained by sampling 2D slices from 3D fetal brain volumes, and target the model to predict the inverse of the sampling process, resembling the idea of self-supervised learning.

`PlaneInVol` is benchmarked on 2D slices sampled from 3D fetal brain volumes at 18-22 weeks of gestational age. Using three evaluation metrics, namely, Euclidean

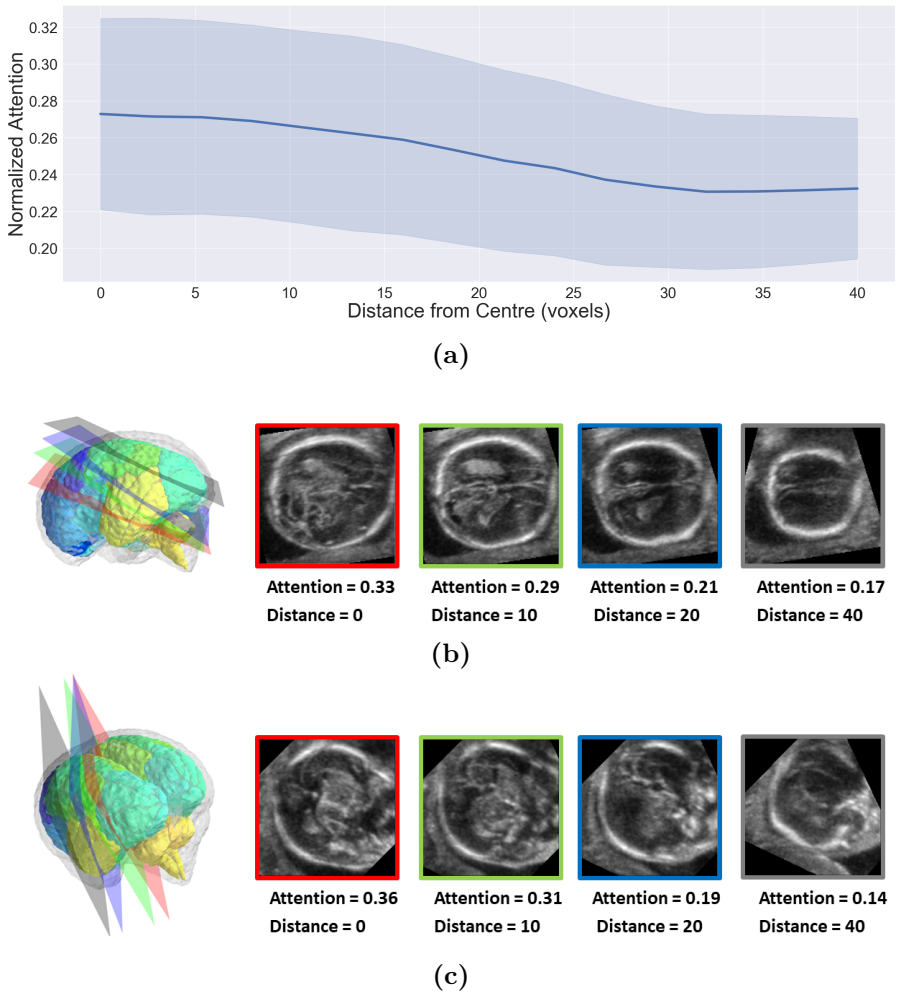


Figure 3.14: Attention visualization. (a) displays the relationship between the mean normalized learned attention (± standard deviation) and the sampled slices' distance from the centre of the fetal brain. (b) and (c) are two sets of attention visualization example. The 3D positions of the images are shown in the 3D fetal brain simulation on the left.

distance, plane angles and normalized cross correlation, which account for both the geometrical and appearance discrepancy between the groundtruth and prediction, in all these metrics, `PlaneInVol` outperforms a baseline model by as much as 23%, when the number of input images increases. It is further demonstrated that the proposed model generalizes to (i) real 2D standard transthalamic plane images, achieving comparable performance as human annotations, as well as (ii) videos of 2D freehand fetal brain scan.

The methodology presented in this chapter may facilitate better identification and localization of different ultrasound scans clinically, and hence lead to more accurate

and objective image acquisition and the analysis of fetal growth and development. The proposed model, `PlaneInVol`, will be further applied and studied in the coming chapters, specifically for unsupervised adaptation to different ultrasound machines (Chapter 4) and volumetric reconstruction from 2D ultrasound images (Chapter 5).

4

Adaptive 3D Localization of 2D Ultrasound Images

Following from the previous chapter, in which `PlaneInVol` is proposed to predict the position of 2D ultrasound fetal brain scans in the 3D brain atlas, this chapter extends that by presenting `AdLocUI`, a framework that **Ad**aptively **Loc**alizes 2D **U**ltrasound **I**mages in the 3D brain atlas. Specifically, `AdLocUI` adapts a trained localization model (*e.g.* `PlaneInVol`) to freehand 2D ultrasound images acquired from arbitrary domains (*e.g.* sonographers, manufacturers and acquisition protocols) in an *unsupervised* manner. The work presented in this chapter has been published in:

Yeung, P.H., Aliasi, M., Haak, M., the INTERGROWTH-21st Consortium, Xie, W. and Namburete, A.I.,: Adaptive 3D Localization of 2D Freehand Ultrasound Brain Images., *International conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022.

Contents

4.1 Introduction

In practice, ultrasound images are acquired by different machines and with different acquisition parameters at different clinical sites. Since there is no worldwide standardization on the quality assurance of ultrasound images [143], different manufacturers may adopt their own and patented image formation and post-processing algorithms when designing their machines, leading to different quality and characteristics of the acquired images. During clinical scanning, different sonographers may have their preferred acquisition protocols and parameters, further enlarging the variations between different acquired ultrasound images, even though they are capturing the same cross-sectional plane of interest.

In Chapter 3, `PlaneInVol` is presented to predict the position of 2D ultrasound fetal brain scans in the 3D brain atlas. `PlaneInVol` is trained on 2D slices sampled from aligned 3D fetal brain volumes, which may look differently from the actual 2D images, in terms of resolution, intensity and noise. Although extensive data augmentation (Chapter 3.2.1) is used to make the model more generalizable and it is demonstrated qualitatively that `PlaneInVol` also work on several native freehand 2D ultrasound sequences (Chapter 3.4.4), there is no guarantee that `PlaneInVol` is able to generalize to images acquired from arbitrary domains (*e.g.* sonographers, manufacturers and acquisition protocols), given that the data augmentation applied is impossible to capture all the variations in every domain. This significantly limits the application of `PlaneInVol` at the bedside.

In order to optimize its potential and applicability in practice, a 2D ultrasound image localization network (or `PlaneInVol`) should satisfy the following criteria:

(i) *Ease of training*: Training `PlaneInVol` should require as little manual annotation as possible, due to the practical challenges of collecting large amount of manually annotated data with high quality from medical professionals.

(ii) *Ease of generalization*: The trained `PlaneInVol` should be able to generalize to images acquired from arbitrary domains easily. If extra steps are needed to adapt the trained `PlaneInVol` to those images, the time and data required for the adaptation should be minimal.

In this chapter, the goal is to achieve both criteria by presenting `AdLocUI`, a framework that **Ad**aptively **Loc**alizes 2D **U**ltrasound **I**mages in the 3D brain atlas. This chapter makes the following contributions: *first*, a framework for the aforementioned localization task is proposed. It is demonstrated that a *single* model, trained with minimal manual annotation (*i.e.* co-alignment of a set of 3D volumes), can be fine-tuned in an **unsupervised** manner. This is verified by demonstrating the approach through adapting the model to 3 different datasets of ultrasound images, acquired from diverse machines and acquisition protocols differing from those of the training data. *Second*, a novel way to fine-tune the trained model to adapt to the target domain 2D ultrasound images is proposed, which utilizes the fact that the overall displacement of a sequence of images in the 3D anatomical atlas is equal to the displacement from the first image to the last in that sequence. As the *third* contribution, it is shown, with ablation studies, that the introduction of the proposed fine-tuning step leads to a significant improvement on localization accuracy when compared to naïve `PlaneInVol`, and `PlaneInVol` fine-tuned by popular unsupervised domain adaptation algorithms [104, 144, 145].

4.1.1 Unsupervised Domain Adaptation

The task considered in this chapter is a typical *domain adaptation* problem, namely adapting a model trained on *source* domain data (*i.e.* 2D slices sampled from 3D volumes) to *target* domain images (*i.e.* native 2D freehand images acquired from different machines). In practice, it is very difficult to obtain the ground-truth plane location for the target domain images to fine-tune the model in a supervised manner. Therefore, the task is formulated as an *unsupervised* domain adaptation problem.

As reviewed in Chapter 2.5, unsupervised domain adaptation considers the setting where no ground-truth task label is available for the target domain data. A few representative unsupervised domain adaptation works are introduced in this section, which will be used as the baselines to compare with `AdLocUI` in this chapter:

- **Multiple Kernel Maximum Mean Discrepancies (MK-MMD) [145]**: A ConvNet is first trained with the source domain data. It is then fine-tuned with the source and target domain data. The ConvNet architecture is divided into three parts. The first three *convolutional* layers extract low-level and, hence, more general features, which should be common across different domains. Therefore, those layers are frozen during fine-tuning. The remaining *convolutional* layers extract higher-level features and, hence, are less transferable. Those layers are updated during fine-tuning. For the following *fully connected* layers, they should be specific for different domains. The Hilbert space embeddings of representations of those layers are, therefore, matched by MK-MMD. Intuitively, although there is no label for the target domain data, the statistics of the embeddings of their representations are forced to match with each other by MK-MMD.

- **Deep CORAL [144]**: The concept of Deep CORAL is very similar to MK-MMD [145]. The ConvNet is also fine-tuned by unlabelled target domain data by minimizing the features' statistics between different domains. Nevertheless, a new loss (*i.e.* CORAL) is utilized to minimize the *second-order* statistics of both domains' features at the last fully connected layer. It was argued

that this may facilitate the optimization and can be used in different network architectures more naturally [144].

- **Domain-Adversarial Neural Network (DANN) [104]**: The idea of DANN is to generate features that cannot be distinguished between source and target domains. During training, a ConvNet is trained to extract features that are discriminative for the task of interest, but indistinguishable between the two domains. To achieve this, the ConvNet is composed of three parts. The *feature extractor* will generate features from the input images. The features will then pass through the *label predictor* which is responsible for the prediction of the task of interest. It is just a normal feed-forward classifier (*i.e.* a few fully connected layers). During training, the features will also pass through the *domain classifier* via a newly proposed *gradient reversal layer*. The job of the domain classifier is to predict whether the input is from the source or target domains. Using back propagation to minimize the domain classification loss, the gradient reversal layer reverses the gradient and, hence, makes the distributions of the features of the two domains more similar. This makes them indistinguishable between the source and target domains, while the network is still able to achieve the task of interest through the feature extractor and label predictor. After the training, the domain classifier will be abandoned and only the feature extractor and label predictor are kept for prediction during inference.

4.2 Method

In Section 4.3.1, the problem setting considered in this chapter is first formulated, namely, adaptively localizing 2D ultrasound neuroimages in the 3D anatomical atlas. Next, the corresponding steps of the proposed framework, `AdLocUI`, namely training (Section 4.2.2), fine-tuning (Section 4.2.3) and inference (Section 4.2.4) are introduced. The whole pipeline is summarized in Fig. 4.1.

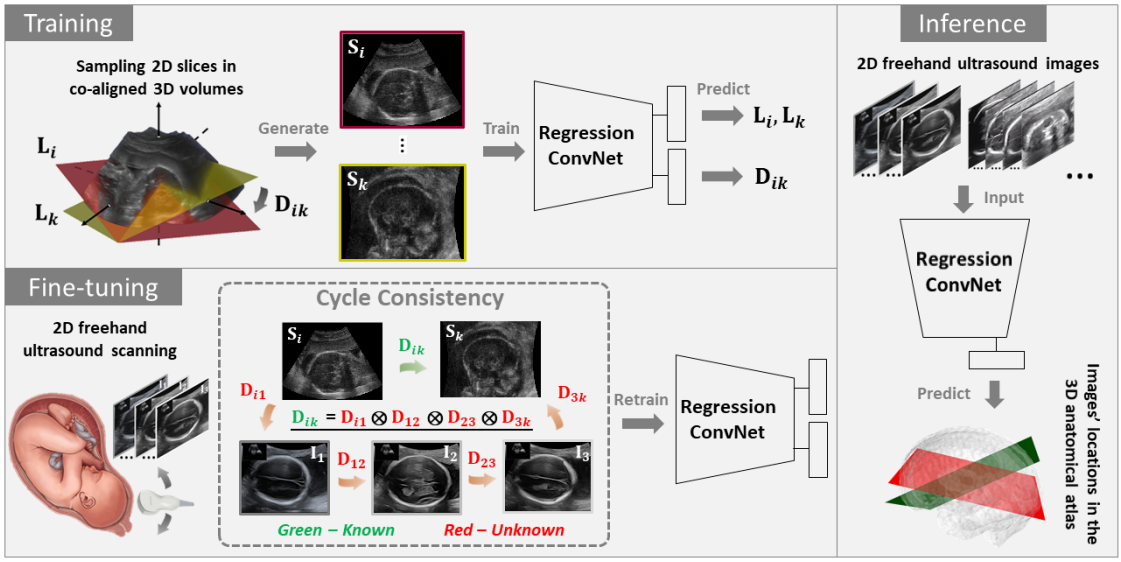


Figure 4.1: Pipeline of `AdLocUI`. During training, 2D slices, $\mathbf{S}_i$, sampled from co-aligned 3D volumes are used to train a regression ConvNet (*e.g.* `PlaneInVol`) to predict the locations, $\mathbf{L}_i$, and displacement $\mathbf{D}_{ik}$, of the 2D slices in the 3D anatomical atlas. The ConvNet is then fined-tuned in an **unsupervised** manner with 2D freehand ultrasound images, $\mathbf{I}_i$, based on the proposed cycle consistency. The fine-tuned ConvNet can then be used to localize $\mathbf{I}_i$ of the same domain (*i.e.* acquired with the same machines and protocols) in the predefined 3D anatomical atlas.

4.2.1 Problem Setup

Each ultrasound acquisition from different machines is considered as a different domain. In general, given a sequence or set of m 2D ultrasound images, $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m\}$, acquired from any domain, the goal is to predict their locations, $\mathcal{L}_{Img} = \{\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_m\}$, in a predefined 3D anatomical atlas, $\mathbb{R}^3_{atlas}$.

This problem is formulated in 3 stages (Fig. 4.1). In *training*, a regression ConvNet, $\psi(\cdot;\theta)$, parametrized by θ, is trained with n 2D slices, $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$, sampled from the corresponding plane locations, $\mathcal{L}_S = \{\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_n\}$, of a set of 3D ultrasound volumes co-aligned in $\mathbb{R}^3_{atlas}$. In this chapter, the regression ConvNet, $\psi(\cdot;\theta)$, is largely based on `PlaneInVol` (with some modifications that will be described in Section 4.2.2), but it can also be a ConvNet with any arbitrary architectures. Therefore, in the remaining of this chapter, it will just be referred as ConvNet for generalizability. After that, $\psi(\cdot;\theta)$ is retrained (*i.e. fine-tuned*) with $\mathcal{S}$ and $\mathcal{I}$, using cycle consistency in an **unsupervised** manner. $\psi(\cdot;\theta)$ can then

be used on $\mathcal{I}$ or images of the same domain as $\mathcal{I}$ during *inference*.

For clarification, in Chapter 3, both the 2D images and 2D slices are referred as **I**, where in this chapter, they have to be differentiated. Therefore, 2D slice sampled from the 3D training volumes is referred as **S** and the target domain 2D ultrasound image is referred as **I**.

4.2.2 Training with Sampled 2D Slices from 3D Volumes

Conventionally, training $\psi(\cdot;\theta)$, requires paired training data (*i.e.* $\{\mathbf{I}_i, \mathbf{L}_i\}$), where $\mathbf{L}_i$ (parameterization of **L** is detailed below) needs to be manually annotated, which is very challenging and time-consuming. In Chapter 3, `PlaneInVol` is proposed to use 2D slices, $\mathcal{S}$, sampled from aligned 3D ultrasound volumes, as the training data. Therefore, the corresponding plane locations, $\mathcal{L}_S$, of the 2D slices are automatically known, voiding the need for further manual annotation. The same strategy is adopted in this chapter.

Data preparation pipeline. Similar to Chapter 3, a set of 3D ultrasound volumes is affinely registered to a common predefined anatomical atlas, $\mathbb{R}^3_{atlas}$, either manually, or by alignment algorithms such as [138], followed by minor manual correction. This is the **only** manual annotation required by `AdLocUI`. 2D slices, $\mathcal{S}$, were then randomly sampled from the aligned volumes, using Fibonacci sphere sampling of polar coordinates [135], on the fly during training. The details were described in Chapter 3.2.1.

Training objectives. With a set of n paired training data, $\{\mathbf{S}_i, \mathbf{L}_i\}_{i=1}^n$, a regression ConvNet, $\psi(\cdot;\theta)$, is trained. $\psi(\cdot;\theta)$ is composed of 3 parts, namely the encoder $\psi_{enc}(\cdot;\theta_{enc})$, location prediction $\psi_{loc}(\cdot;\theta_{loc})$ and displacement prediction $\psi_{disp}(\cdot;\theta_{disp})$. When compared to `PlaneInVol` proposed in Chapter 3, the displacement prediction, $\psi_{disp}(\cdot;\theta_{disp})$, is a new component introduced in this chapter. First, $\mathcal{S}$ are randomly

augmented by scaling, in-plane translation, contrast adjustment and random noise. A feature vector, $\mathbf{v}_i$, is then generated by the encoder part, $\psi_{enc}(\cdot;\theta_{enc})$, for each $\mathbf{S}_i$:

$$[\mathbf{v}_1,\ \mathbf{v}_2, ...,\ \mathbf{v}_n] = [\psi_{enc}(\mathbf{S}_1;\theta_{enc}),\ \psi_{enc}(\mathbf{S}_2;\theta_{enc}),\ ...,\ \psi_{enc}(\mathbf{S}_n;\theta_{enc})] \tag{4.1}$$

Similar to `PlaneInVol` as proposed in Chapter 3, the feature vectors, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, are used to predict the plane locations, $\mathcal{L}_S$, by the location prediction part, $\psi_{loc}(\cdot;\theta_{loc})$:

$$[\hat{\mathbf{L}}_1,\ \hat{\mathbf{L}}_2, ...,\ \hat{\mathbf{L}}_n] = [\psi_{loc}(\mathbf{v}_1;\theta_{loc}),\ \psi_{loc}(\mathbf{v}_2;\theta_{loc}),\ ...,\ \psi_{loc}(\mathbf{v}_n;\theta_{loc})] \tag{4.2}$$

where $\hat{}$ indicates predicted values. Unlike `PlaneInVol`, $\mathbf{L}_i$ and the *displacement*, $\mathbf{D}_{ik}$, between each pair of slices, $\mathbf{S}_i$ and $\mathbf{S}_k$, in $\mathbb{R}^3_{atlas}$ are simultaneously predicted by the displacement prediction part, $\psi_{disp}(\cdot;\theta_{disp})$:

$$[...,\ \hat{\mathbf{D}}_{ik}, ...,\ \hat{\mathbf{D}}_{nn}] = [...,\ \psi_{disp}(\mathbf{v}_i, \mathbf{v}_k;\theta_{disp}),\ ...,\ \psi_{disp}(\mathbf{v}_n, \mathbf{v}_n;\theta_{disp})] \tag{4.3}$$

Parameterization of L and D. Following the practice as described in Chapter 3.2.2, the plane location, $\mathbf{L}_i \in \mathbb{R}^{3\times3}$, is parameterized by three anchor points (*i.e.* their *x, y* and *z* coordinates), namely the *top right*, *top left* and *bottom right* corners, of $\mathbf{S}_i$. The displacement, $\mathbf{D}_{ik}$, from $\mathbf{S}_i$ to $\mathbf{S}_k$, in $\mathbb{R}^3_{atlas}$ is therefore parameterized as $(\mathbf{L}_i - \mathbf{L}_k)$. There are other parameterization methods, such as Euler angles and quaternions [134, 135], which is not the focus of this chapter and may be investigated in the future work.

Training loss ($\mathfrak{L}_t$). Since $\hat{\mathbf{L}}$ and $\hat{\mathbf{D}}$ are simultaneously predicted, it can be represented as a multi-task learning problem and weighted mean least-squared error (MSE) is used as the loss function:

$$\mathfrak{L}_t = w_L \cdot \text{MSE}\left(\hat{\mathbf{L}}, \mathbf{L}\right) + w_D \cdot \text{MSE}\left(\hat{\mathbf{D}}, \mathbf{D}\right) \tag{4.4}$$

where w_L and w_D are the weights of the respective MSE loss.

4.2.3 Fine-tuning with 2D Ultrasound Images

The trained ConvNet, $\psi(\cdot;\theta)$, can then be fine-tuned with a new set of m 2D ultrasound images, $\mathcal{I}$, acquired from any domain. The retraining relies on *cycle consistency* and uses both the training data (*i.e.* $\{\mathcal{S}, \mathcal{L}_S\}$) and the new set of images, $\mathcal{I}$, *without* further manual annotation.

Cycle consistency. Although the plane locations, $\mathcal{L}_{Img}$ of the new $\mathcal{I}$ are unknown, by cycle consistency, it is known that the overall displacement, $\mathbf{D}$, of a sequence of images in $\mathbb{R}^3_{atlas}$ must be equal to $\mathbf{D}$ from the first image to the last of that sequence. For example, as illustrated in Fig. 4.1, the overall displacement (i) $\mathbf{S}_i \rightarrow \mathbf{I}_1$ (*i.e.* $\mathbf{D}_{i1}$), and $\mathbf{I}_1 \rightarrow \mathbf{I}_2$ (*i.e.* $\mathbf{D}_{12}$), and $\mathbf{I}_2 \rightarrow \mathbf{I}_3$ (*i.e.* $\mathbf{D}_{23}$), and $\mathbf{I}_3 \rightarrow \mathbf{S}_k$ (*i.e.* $\mathbf{D}_{3k}$) is equal to (ii) $\mathbf{S}_i \rightarrow \mathbf{S}_k$ (*i.e.* $\mathbf{D}_{ik}$). While every $\mathbf{D}$ in (i) is unknown, $\mathbf{D}_{ik}$ in (ii) is known from the original training data. Therefore, the cycle consistency loss ($\mathfrak{L}_c$) can be constructed with this equality to retrain $\psi(\cdot;\theta)$:

$$\mathfrak{L}_c = \text{MSE}\left(\hat{\mathbf{D}}_{i1} \otimes \hat{\mathbf{D}}_{12} \otimes \hat{\mathbf{D}}_{23} \otimes \hat{\mathbf{D}}_{3k},\ \mathbf{D}_{ik}\right) \tag{4.5}$$

where $\otimes$ depends on the choice of the parameterization of $\mathbf{D}$ and, hence, $\otimes$ is simply *subtraction* in this chapter (similar to the derivation of $\mathbf{D}$ from $\mathbf{L}$ described in Section 4.2.2). When predicting two consecutive displacements (*e.g.* $\mathbf{D}_{12}$ and $\mathbf{D}_{23}$), the common image involved (*i.e.* $\mathbf{I}_2$) is augmented differently, which coincides with the recent self-supervised and unsupervised learning studies [89, 91, 146] that emphasize the importance of data augmentation.

Fine-tuning loss ($\mathfrak{L}_f$). Since the goal of `AdLocUI` is to predict the corresponding plane location, $\mathcal{L}_{Img}$, of $\mathcal{I}$, relying solely on the cycle consistency loss, $\mathfrak{L}_c$, (*i.e.* supervise only on $\mathbf{D}$) may diverge the prediction or even fall into trivial solutions [97]. Therefore, the original training loss, $\mathfrak{L}_t$ (Eq. 4.4), is added to regularize the retraining. The overall fine-tuning loss, $\mathfrak{L}_f$ is:

$$\mathfrak{L}_f = w_c \cdot \mathfrak{L}_c + \mathfrak{L}_t \tag{4.6}$$

where w_c is the weight of the cycle consistency loss, $\mathfrak{L}_c$.

4.2.4 Inference

The fine-tuned ConvNet, $\psi(\cdot;\theta)$, can be used on the set of 2D ultrasound images, $\mathcal{I}$, or other 2D ultrasound images of the same domain (*i.e.* acquired from the same machine with the same acquisition protocols) to predict their corresponding locations, $\mathcal{L}_{Img}$, in the predefined 3D anatomical atlas, $\mathbb{R}^3_{atlas}$:

$$[\hat{\mathbf{L}}_1,\ \hat{\mathbf{L}}_2, ...,\ \hat{\mathbf{L}}_m] = [\psi(\mathbf{I}_1;\theta),\ \psi(\mathbf{I}_2;\theta),\ ...,\ \psi(\mathbf{I}_m;\theta)] \tag{4.7}$$

4.3 Experiment

4.3.1 Experimental Setup

`AdLocUI` and other baseline approaches were first trained with 2D slices, $\mathcal{S}$, sampled from 50 3D volumes acquired by Philips HD9 (*Training* in Fig. 4.1). The training dataset and details follow those described in Chapter 3.3.1 and 3.3.2. The trained networks were then fine-tuned and evaluated (*Fine-tuning* and *Inference* in Fig. 4.1) on both volume-sampled 2D images and native 2D freehand images. The training and testing images were acquired from different clinical sites and machines, simulating the cross-domain variance observed in reality. `AdLocUI` was compared with `PlaneInVol` (Chapter 3) and that fine-tuned by popular unsupervised deep domain adaptation methods, namely MK-MMD [145], DANN [104] and CORAL [144], as introduced in Section 4.1. The implementation details of different approaches are summarized in Table 4.1.

For MK-MMD [145], DANN [104] and CORAL [144], their respective loss, $\mathfrak{L}_r$, replaced $\mathfrak{L}_c$ of Eq. 4.6:

$$\mathfrak{L}_f = w_r \cdot \mathfrak{L}_r + \mathfrak{L}_t \tag{4.8}$$

<table>
<tr><td>Approaches</td><td>PlaneInVol</td><td>AdLocUI</td></tr>
<tr><td>Encoder
$\psi_{enc}(\cdot;\theta_{enc})$</td><td colspan="2" rowspan="2">Refer to Table 3.2</td></tr>
<tr><td>Location prediction
$\psi_{loc}(\cdot;\theta_{loc})$</td></tr>
<tr><td>Displacement prediction
$\psi_{disp}(\cdot;\theta_{disp})$</td><td>-</td><td>(FC - ReLU)×3

FC size from
512 to 256 to 9</td></tr>
<tr><td rowspan="2">Training hyperparameters</td><td>-</td><td>- $w_L = 1$
- $w_D = 0.5$</td></tr>
<tr><td colspan="2">- Batch size of 80
- Learning rate of 0.0001
- lr halved when errors plateaued
- Early stop when errors further plateaued
- ADAM optimization</td></tr>
<tr><td>Fine-tuning hyperparameters</td><td>- w_r for MK-MMD[145] loss = 10
- w_r for CORAL[144] loss = 1
- w_r for DANN[104] loss = 1</td><td>- $w_c = 1$</td></tr>
<tr><td>Other details</td><td colspan="2">- Python 3.7, pytorch 1.9
- Nvidia GTX 1080ti, 12GB memory</td></tr>
</table>

Table 4.1: Implementation details of different approaches.

where w_r is the weight of $\mathfrak{L}_r$. With some hyperparameters tuning, the w_r of each baseline methods are summarized in Table 4.1.

For volume-sampled 2D images, two different settings were considered, both corresponding to realistic scenarios. Firstly, it was the scenario where the *same* set of images was used for fine-tuning and then testing. This is relevant when **offline analysis** is performed, where there is sufficient time for fine-tuning with the test images before final analysis. Secondly, it was the scenario where *different* sets of images (from the same domain) were used for fine-tuning and testing. This was achieved by using half of the testing 3D volumes for fine-tuning and the other half for testing. This setting corresponds to **online prediction**, for example scanning guidance, where a set of example images were acquired in advance for fine-tuning.

4.3.2 Testing Dataset

Volume-sampled testing images. `AdLocUI` and other baseline approaches were tested on 2D slices sampled from 17 aligned 3D volumes (resize and crop to $160 \times 160 \times 160$ voxels at a resolution of around $0.6 \times 0.6 \times 0.6$ mm^3) acquired by GE Voluson E10, which were different from the training volumes (Philips HD9). The volumes were acquired at the Leiden University Medical Center, between 19 and 21 gestational weeks and aligned to the same common 3D atlas as the training data to ensure consistent evaluation. 3000 slices were sampled from each testing volume uniformly. Although they are not native 2D freehand ultrasound images, domain shift still exists and the availability of ground-truth enables more complete quantitative analysis.

Native 2D freehand images. Images from video sequences of 2D freehand ultrasound brain scans, acquired by GE Voluson E10 (4 sequences, 829 2D images in total) and Voluson E8 (3 sequences, 531 images in total) from two different clinical centers, were tested and analyzed. Each image was cropped to 160×160 pixels.

4.3.3 Evaluation metrics

Volume-sampled testing images. Following Chapter 3.3.3, the Euclidean distance (ED) between the coordinates of the predicted and ground-truth planes in the $\mathbb{R}^3_{atlas}$ and the dihedral plane angle (PA) between them were used as the evaluation metrics when testing on volume-sampled testing images.

Native 2D freehand images. As the ground-truth locations were not available for native 2D freehand images, it was not possible to achieve the same detailed quantitative analysis as the volume-sampled images. Therefore, another quantitative test was proposed. As the acquisition of the video sequences was smooth and

Table 4.2: Evaluation results (mean±standard deviation) on volume-sampled 2D images on two settings, 4.2a and 4.2b, evaluated by Euclidean distance (ED) and dihedral angle (DA). ↓ indicates lower values being more accurate. * indicates manual annotation being used.

	ED ↓ (voxel)	**DA ↓ (rad)**	**ED ↓ (voxel)**	**DA ↓ (rad)**
`PlaneInVol`				
without fine-tuning	71.1±29.9	0.264±0.177	70.6±25.3	0.265±0.137
with MK-MMD[145]	71.4±27.0	0.266±0.153	72.6±26.0	0.267±0.140
with CORAL[144]	79.3±29.9	0.276±0.159	80.8±28.1	0.278±0.149
with DANN[104]	72.8±30.5	0.265±0.160	72.4±27.9	0.266±0.143
*supervised fine-tuning	11.3±1.57	0.172±0.055	28.6±14.2	0.202±0.084
`AdLocUI` **(ours)**				
without fine-tuning	63.0±29.0	0.251±0.166	62.7±25.0	0.253±0.138
proposed fine-tuning	**23.7±9.01**	**0.198±0.092**	**33.0±15.1**	**0.211±0.097**

(a) Fine-tune and test on the *same* set of images **(b)** *different* set of images

continuous, the locations of consecutive images should not change abruptly, but show a gradual transition. Such a rate of change (Δc) was quantified as:

$$\Delta c = \frac{\mathrm{ED}(\hat{\mathbf{P}}_i, \hat{\mathbf{P}}_{i+1})}{1 - \mathrm{NCC}(\mathbf{I}_i, \mathbf{I}_{i+1})} \tag{4.9}$$

where $\hat{\mathbf{P}}_i$ is the coordinates of the predicted plane of $\mathbf{I}_i$ and NCC is the normalized cross-correlation. Normalized (*i.e.* by the mean of Δc) standard deviation (NSTD) was used to quantify the consistency of Δc throughout the whole video sequence, which should be low ideally.

4.4 Results

4.4.1 Volume-Sampled Images

`AdLocUI`, via ablation studies, was compared to different baseline approaches in two different settings as described in Section 4.3.1.

Offline analysis (Table 4.2a). The *same* set of images was used for fine-tuning and then testing, which corresponds to performing offline analysis, where there is sufficient time for fine-tuning with the test images before final analysis. From

Table 4.2a, the original `PlaneInVol` (*i.e.* without fine-tuning) achieved ED=71.1 and DA=0.264, which was slightly worse than `AdLocUI` without fine-tuning (ED=63.0 and DA=0.251). The multi-task learning (*i.e.* additional task of predicting $\mathbf{D}_{ij}$) contributed to such improvement. The proposed fine-tuning step, which does not require any additional manual annotation, contributed to a significant ($p<0.05$, student's t-test) improvement (ED=23.7 and DA=0.198). An *unlikely* situation was also analyzed, where it was assumed to have the ground-truth locations of the testing images for fine-tuning (*i.e.* retraining) `PlaneInVol` in a supervised manner. This can be viewed as the *oracle* of the accuracy of the prediction (ED=11.3 and DA=0.172).

Online prediction (Table 4.2b). *Different* sets of images (from the same domain) were used for fine-tuning and testing. This is relevant when *online* prediction is performed, for example scanning guidance, where a set of example images is acquired in advance for fine-tuning . From Table 4.2b, without fine-tuning, `PlaneInVol` (ED=70.6 and DA=0.265) and `AdLocUI` (ED=62.7 and DA=0.253) performed similarly as the first scenario. Compared to the first scenario, a pronounced drop in performance was seen for supervised fine-tuning of `PlaneInVol` (ED=28.6 and DA=0.202) when the fine-tuning and testing images were no longer the same. This had less severe impact to `AdLocUI` with the proposed fine-tuning (ED=33.0 and DA=0.211), which was still significantly ($p<0.05$) better than the baselines. Despite its slightly better performance, supervised fine-tuning requires manually annotated image locations to retrain the network for every new machine or protocol, which is not applicable in practice. On the contrary, `AdLocUI` just needs the raw 2D images for fine-tuning, which is much more achievable in neuroimaging studies.

Fine-tuned with existing DA methods. `AdLocUI` was also compared with `PlaneInVol` fine-tuned by popular unsupervised DA methods (*i.e.* MK-MMD [145], DANN [104] and CORAL [144]). Despite some trials of hyperparameters tuning, as shown in Table 4.2, their results were still comparable or worse than no fine-tuning.

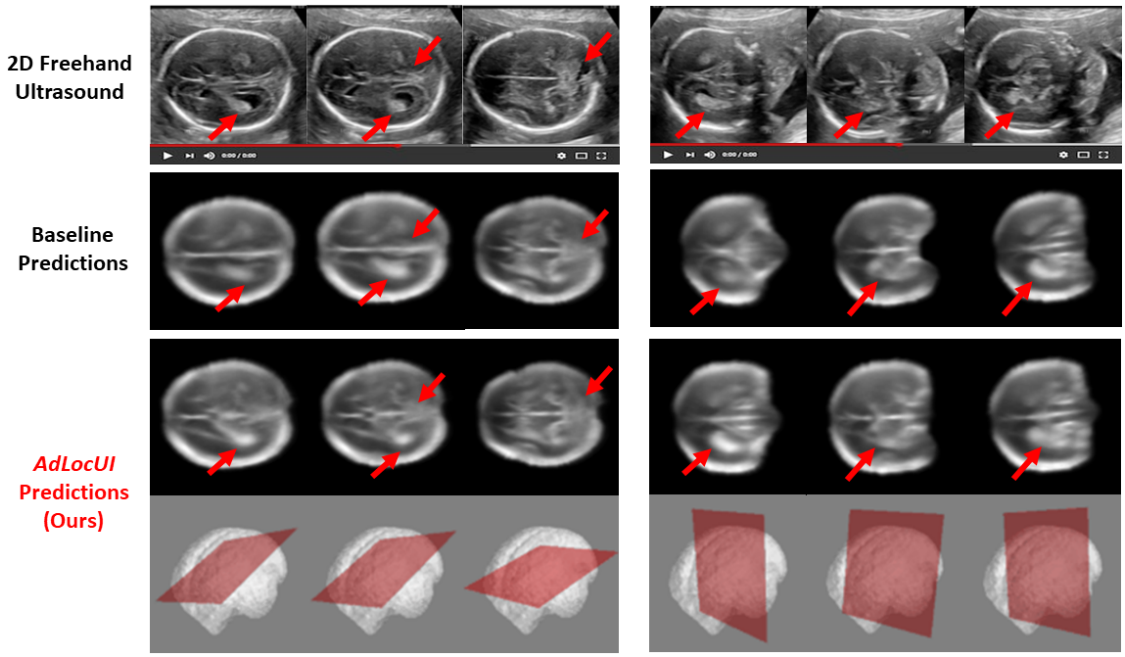


Figure 4.2: Localization of 2D freehand ultrasound images in the 3D brain atlas. 2D slices sampled from the 3D atlas using image locations predicted by the baseline (*i.e.* `PlaneInVol`) and `AdLocUI` are presented, where `AdLocUI`'s predictions show better correspondence (*i.e.* emphasized by the red arrows) with the ultrasound images, suggesting more accurate 3D localization prediction by `AdLocUI`.

This may be due to the fact that most DA approaches were designed for classification tasks, which may not be directly applicable to the regression task [147]. This further verifies the value of the work presented in this chapter.

4.4.2 Native Freehand Images

In the experiments on native 2D freehand ultrasound images, the predicted image locations were used to sample the corresponding slices from the 3D atlas, to which the 3D training volumes were co-aligned. The sampled slices should match with the corresponding input images for accurate predictions. As shown in Fig. 4.2, predictions from `AdLocUI` clearly demonstrated a much better match, in terms of similarity and anatomical structures present as indicated by the red arrows in the figure, with the corresponding input images at different orientations, when compared to `PlaneInVol`. By the proposed quantitative test (*i.e.* NSTD of Δc)

as described in Section 4.3, AdLocUI achieved a result of 0.553, which was lower than both PlaneInVol (0.706) and AdLocUI without fine-tuning (0.726), suggesting that the predicted localization of AdLocUI was more consistent throughout the ultrasound video sequence, which was indicative of the smooth frame-to-frame transitions expected in freehand scanning. Both the qualitative and quantitative results showed AdLocUI's superior performance when being applied on native 2D freehand ultrasound images in practice.

4.5 Conclusion

In this chapter, AdLocUI, a framework for localizing 2D ultrasound brain images in the 3D anatomy, is proposed. It extends PlaneInVol (Chapter 3) by using an intuitive cycle consistency loss to adapt the localization model to images acquired from different machines and protocols in an **unsupervised** manner.

AdLocUI is benchmarked on 3 different datasets of ultrasound images, acquired from diverse machines and acquisition protocols differing from those of the training data. It is shown, with ablation studies, that the introduction of the proposed unsupervised cycle consistency leads to a significant improvement on localization accuracy when compared to naïve PlaneInVol, and PlaneInVol fine-tuned by popular unsupervised domain adaptation algorithms, namely MK-MMD [145], DANN [104] and CORAL [144]. Furthermore, the improvement applies on two different realistic settings, namely *offline* analysis and *online* prediction, which may further verify AdLocUI's potential by the bedside under different circumstances.

All in all, Chapter 3 and this chapter propose a complete framework for localizing 2D fetal ultrasound images in the 3D brain, which may be further used for volumetric reconstruction (Chapter 5) and freehand guidance for training and facilitating more objective analysis and diagnosis. This framework satisfies the three fundamental principles described in Chapter 1.1. In specific, PlaneInVol (Chapter 3) is **sensorless**, which is trained with **minimal human annotation**

(*i.e.* co-alignment of a set of 3D volumes). `AdLocUI` proposed in this chapter further addresses the **generalization** issue to maximize the impact of the framework.

5

Volumetric Reconstruction from 2D Ultrasound Images

Building upon the ability to predict the position of 2D ultrasound fetal brain scans in the 3D brain atlas as proposed in previous chapters, this chapter presents `ImplicitVol` to reconstruct 3D volumes from non-sensor-tracked 2D ultrasound images, where their 3D locations are predicted by `PlaneInVol`. Specifically, using the predicted 3D location of individual 2D ultrasound images as an initialization, `ImplicitVol` will then jointly refine the image location and learn the volumetric reconstruction using *implicit representation*. The work presented in this chapter is submitted to (under review):

Yeung, P.H., Hesse, L., Aliasi, M., Haak, M., the INTERGROWTH-21st Consortium, Xie, W. and Namburete, A.I.,: Sensorless Volumetric Reconstruction of Fetal Brain Freehand Ultrasound Scans with Deep Implicit Representation., *submitted to Medical Image Analysis*.

Contents

5.1 Introduction

Two-dimensional freehand ultrasound is most commonly used at the bedside, and the operator (sonographer) freely controls the image acquisition path by navigating the probe during the examination. While methods presented in the previous chapters may facilitate more accessible and standardized scanning by localizing individual 2D scan in the 3D brain atlas, one of the major limitations of 2D ultrasound is still not answered, namely the fact that a given 2D ultrasound image frame is limited to representing a cross-sectional view of the 3D anatomy, which inherently fails to capture richer contextual volumetric information. 3D ultrasound, typically acquired using a 3D ultrasound transducer, may capture the structures in their entirety within a volumetric image which displays the spatial arrangement of structures within a region of interest. Despite all the advantages of 3D ultrasound over its 2D counterpart as reviewed in Chapter 2.2, for example richer diagnostic potential [39, 40, 44, 148] and easier offline and secondary examination [33, 34], due to its more sophisticated hardware requirements and larger footprint, a 3D ultrasound system may cost ten times more than a 2D system. This limits its use in practical scenarios, such as scanning by the bedside, or in point-of-care settings.

Three-dimensional transducers and the corresponding supporting hardware are not typical in obstetric and point-of-care settings. In the absence of such equipment, acquiring a set of adjacent 2D freehand images, along with the transducer's positional information, can still enable post-hoc reconstruction of a volumetric image. This can be achieved by using external positional sensor to track the ultrasound probe [149]. Different types of sensors have been proposed, such as optical positioners [150] or electromagnetic sensors [151]. While such a solution would be valid for reconstructing volumes of relatively static structures (*e.g.* liver [152]) or those (*e.g.* carotid artery [150]) affected by regular motion cycles (*e.g.* respiration), it would prove impractical in obstetric scanning where fetal motion is independent and uncorrelated with the freehand positioning of the probe. Furthermore, other sensor-related issues, such as electromagnetic interference from nearby metal objects [153] and attachment of obtrusive markers or sensors on the probe [154], may further hamper the acquisition of 3D volumes.

In this chapter, `ImplicitVol` is proposed to reconstruct 3D volumes from a set of freehand 2D ultrasound scans, which are available in routine clinical screening, without using any external sensors. This approach presents the diagnostic advantages of 3D ultrasonography while simultaneously offering the accessibility and ease-of-use of 2D ultrasonography.

The standard approach to slice-to-volume ultrasound reconstruction is to register the 2D scans into the 3D volumes, and explicitly perform interpolations in the resulting volumetric representation. Despite its promise, multi-step reconstruction often suffers from different challenges, such as, error accumulation due to the incorrect estimation from 2D scans to their corresponding 3D locations, and limited resolution from the low tessellated grid.

In this chapter, the goal is to directly address these challenges by proposing a novel slice-to-volume reconstruction pipeline based on *implicit representation*. Specifically, the proposed approach parameterises the 3D volume as a deep neural network, which jointly refines the slice-to-volume registrations and learns a full 3D reconstruction based on only a set of 2D scans. To the best of my knowledge,

the framework, `ImplicitVol`, is the first study to propose a *genuinely sensorless* (*i.e.* in both training and inference) 3D reconstruction pipeline based on deep implicit representation. As freehand ultrasound is the mainstay in prenatal health monitoring, and is the most widely used and rapidly evolving technique for fetal imaging, the proposed technique is demonstrated for task of reconstructing 3D volumes of the fetal brain.

5.1.1 Conventional 3D Ultrasound Reconstruction

Existing methods to construct a slice-to-volume ultrasound reconstruction pipeline were surveyed in [155], and can be summarized by some or all of the following steps:

3D Localization: First, the 3D location of each tomographic (2D) ultrasound image is estimated, where an external sensor tracking is required at either the *training* [156, 157] or *inference* [150, 151] stage. These methods are subject to errors caused by the subjects' internal motion (*e.g.* fetal movement).

Interpolation: A 3D volume, represented *discretely* as a tensor of intensities, is reconstructed by 'registering' the localized 2D scans back to the 3D space, with holes being interpolated. However, such slice-to-volume back-projections are often prone to errors, leading to artifacts and thus require*post-hoc* corrections.

Correction. Finally, approaches, such as [158, 159], have been proposed to correct the aforementioned reconstruction artifacts, based on kernel smoothing and denoising. However, the effect may be limited, as the source of inaccuracy from the localization of the 2D images is unsolved.

Conventional 3D reconstruction pipelines have been extensively built on *explicit* representations: representing a volume in terms of a 3D array. Instead, in this chapter, a deep neural network is parameterised to *implicitly* represent the 3D

	Explicit	**Implicit**
continuity	discrete voxel grid	continuous function
memory-efficiency	lower	higher
resolution	defined by grid	arbitrary
gradient & derivatives	limited by discretization	continuous & well-defined

Table 5.1: Comparison between the explicit and implicit representations for 3D volumes.

volume. Such a representation is continuous, and enables the querying of intensities at arbitrary spatial coordinates. With only a set of 2D scans available, it can produce jointly optimal 3D structures (voxel intensities) and 3D location estimations for these scans. The difference and comparison between *explicit* and *implicit* representations are reviewed in the next section (Section 5.1.2).

5.1.2 Construction of 3D Representations

By its very nature, a 3D volume is a one-to-one mapping from a set of 3D spatial positions (*i.e.* 3D coordinates) to their corresponding intensity values in the real world. In general, there are two different ways for representing a 3D volume, either *explicitly* or *implicitly*, as described and compared in Table 5.1 and as follows:

Explicit Representation. Conventionally, a 3D volume, $\mathbf{V} \in \mathbb{R}^{H \times W \times D \times C}$, is represented *discretely* and *explicitly* as a tensor with height (H), width (W), depth (D), and intensity channels (C). Besides the slice-to-volume ultrasound reconstruction described in Section 5.1.1, most other medical applications involving 3D volumes, for example volume segmentation [160–162], and registration [138, 163], also rely on using such representation.

Implicit Representation. As an alternative, a 3D volume can also be represented as a zero level set of a *continuous* function parameterized by Θ. Such *implicit* representation compresses the volumetric information and encodes it as parameters of a model, for example a deep neural network, that maps the 3D coordinates, $\mathbf{x} = (x, y, z)$, to intensities, *i.e.* $F_\Theta : \mathbf{x} \rightarrow c$. Implicit representation has been proposed

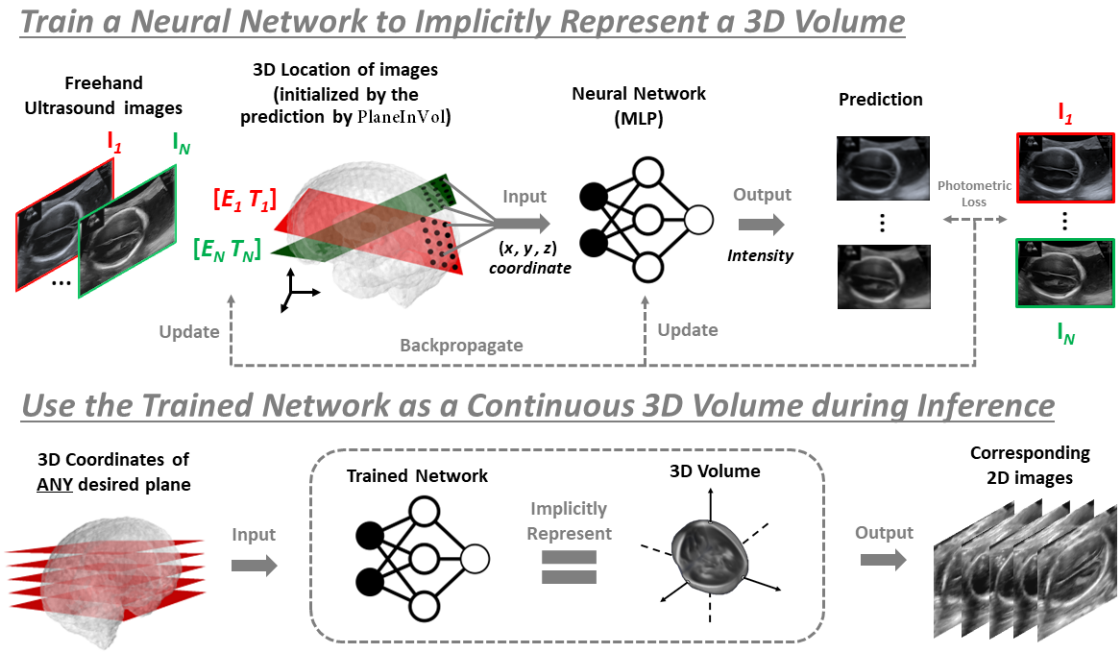


Figure 5.1: Pipeline of `ImplicitVol`. During training, a set of 2D freehand ultrasound images, $\{\mathbf{I}_i\}_{i=1}^N$, and their estimated 3D location, $\{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, are used to train a deep neural network to implicitly represent the *continuous* 3D volume from which $\{\mathbf{I}_i\}_{i=1}^N$ are acquired. During inference, images at arbitrarily oriented planes can be obtained as output, by feeding the grid coordinates of the corresponding query plane to the network.

as an alternative approach for novel view synthesis for natural scenes [164–167]. However, it is less frequently studied and applied to medical imaging tasks [168, 169]. This work is the first to apply implicit representation techniques to ultrasound imaging. Due to speckle and strong view-dependency, ultrasound is likely to reap the most benefit from compounding information from multiple views, and thus vastly improve the structural representation.

5.2 Methods

In this section, the problem setting of this chapter, namely reconstructing a 3D volume from only a sparse set of 2D fetal brain scans with implicit representation, is first formulated. Next, in accordance with the three conventional steps of 3D reconstruction summarized in Section 5.1.1, the corresponding components of `ImplicitVol` and joint optimization are introduced. The pipeline of `ImplicitVol`

is summarized in Fig. 5.1

5.2.1 Problem Setup

Consider a set of 2D ultrasound images, $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$, capturing N different cross-sectional views of a region of interest (e.g. fetal brain) at the corresponding 3D locations, parameterized by $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, with $\mathbf{E} = \{\theta_x, \theta_y, \theta_z\}$ being the 3D Euler angles and $\mathbf{T} = \{t_x, t_y, t_z\}$ the translation. The goal is to reconstruct the volume, such that any 2D cross-sectional view of arbitrary resolution can be generated by inputting the corresponding 3D coordinates, $\mathbf{x} = (x, y, z)$.

Inspired by [164, 165], the volume is represented as a continuous function, parameterized by a multi-layer perceptron (MLP) representation network $F_\Theta : \mathbf{x} \rightarrow c$. The weights, Θ, are learned by minimizing the discrepancy between the actual and network-predicted intensities of $\{\mathbf{I}_i\}_{i=1}^N$, when the 3D coordinates, $\mathbf{x}$, computed from the corresponding $\{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, are passed as input to the network.

5.2.2 Sensorless 3D Localization of 2D Scans

`PlaneInVol` is used to estimate the 3D locations, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, of the set of 2D ultrasound images, $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$. Not requiring any external tracking, `PlaneInVol` is trained with a set of 2D slices, sampled from a set of co-aligned 3D brain volumes, and their locations in the 3D aligned space. The details of the training and inference are included in Chapter 3.

5.2.3 3D Reconstruction with Implicit Representation

Conceptually, the idea is to *store* the 3D volume in a MLP, the weights of which are learned through a set of training data, namely the 2D ultrasound images, $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$ and their pre-computed 3D locations, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, detailed in Section 5.2.2.

During training, the 3D coordinate, $\mathbf{x}_i^p \in \mathbb{R}^{3\times1}$, for pixel p of the 2D ultrasound image, $\mathbf{I}_i \in \mathbb{R}^{H\times W}$, is first derived from the estimated $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$:

$$\mathbf{x}_i = \varepsilon(\mathbf{E}_i, \mathbf{T}_i) \tag{5.1}$$

Specifically, ε is achieved by first rotating the 3D coordinate, $\mathbf{r}^p \in \mathbb{R}^{3\times1}$, of pixel p of the reference xy plane by $\mathbf{E}_i = \{\theta_x, \theta_y, \theta_z\}$, and then translating it by $\mathbf{T}_i$. Computationally, a rotation matrix, $\mathbf{R}_i \in \mathbb{R}^{3\times3}$ is first generated by:

$$\mathbf{R} = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} \tag{5.2}$$

followed by transforming the 3D coordinate, $\mathbf{r}^p \in \mathbb{R}^{3\times1}$, of pixel p of a reference plane ($\mathbb{R}^{3\times H\times W}$), by the rotation matrix, $\mathbf{R}_i \in \mathbb{R}^{3\times3}$, and the translation $\mathbf{T}_i \in \mathbb{R}^{3\times1}$:

$$\mathbf{x}_i^p = \mathbf{R}_i \cdot \mathbf{r}^p + \mathbf{T}_i \tag{5.3}$$

Positional Encoding. Mapping the 3D coordinate, $\mathbf{x}$, to a higher dimensional space better represents the high frequency variation in the object's intensity and geometry [165, 170]. Therefore, $\mathbf{x}$ is encoded by the function $E : \mathbb{R} \rightarrow \mathbb{R}^{2L}$ [165], where L is the encoding dimension which is a hyperparameter:

$$E(n) = (\sin(2^0\pi n), \cos(2^0\pi n), ..., \sin(2^{L-1}\pi n), \cos(2^{L-1}\pi n)) \tag{5.4}$$

where n are the normalized values (*i.e.* from -1 to 1) of each x, y and z. In the following sections, 3D coordinate refers to the encoded coordinate, $\mathbf{x} \in \mathbb{R}^{3\times2L}$.

Network Training. With the training set, $\{\mathbf{x}_i^p, \mathbf{I}_i^p\}\ \forall i, p$, the weights, Θ, of the representation network, F_Θ, can be learned through normal back propagation:

$$\Theta^* = \arg\min_\Theta \sum_{i,p} \mathfrak{L}(F_\Theta(\mathbf{x}_i^p), \mathbf{I}_i^p) \tag{5.5}$$

where Θ^* are the optimized weights of the representation network and $\mathfrak{L}$ is the photometric loss, *i.e.* structural similarity (SSIM) loss [171]:

5.2.4 Joint Optimization for Location Refinement

In practice, the 3D locations, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, predicted by `PlaneInVol`, are imperfect due to prediction error. Inspired by [166], during training the network, F_Θ, the pre-computed 3D locations, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, were updated (*i.e. refined*) simultaneously through joint optimization which can be summarized as:

$$\Theta^*, \Lambda^* = \arg\min_{\Theta,\Lambda} \mathfrak{L}(F_\Theta(E(\varepsilon(\Lambda))), \mathcal{I}) \tag{5.6}$$

where $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$ and $\varepsilon(\cdot)$ and $E(\cdot)$ are from Eq. 5.1 and 5.4, respectively.

Iterative Reinitialization. As the 3D locations, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, and the network, F_Θ, are optimized and trained *simultaneously*, the network may overfit to a set of sub-optimal 3D locations. Therefore, *iterative refinement* is proposed, where the network weights, Θ, are iteratively reinitialized every e epochs for t times while the 3D locations, Λ, are continuously and jointly optimized. Despite the simplicity of the proposed iterative reinitialization, significant improvement is demonstrated.

5.2.5 Inference

The trained representation network, F_Θ, represents a *continuous* 3D fetal brain captured by the set of 2D images. The 3D volume or any 2D cross-sectional view at any resolution can be easily obtained as the output, by feeding the corresponding grid coordinates for the desired slice to the network, as illustrated in the bottom half of Fig. 5.1.

5.3 Experiment

5.3.1 Overview of study design

The aim of this chapter is to propose a framework, `ImplicitVol`, for reconstructing a 3D volume from a set of 2D ultrasound images. The proposed `ImplicitVol`, with

ablation studies, was tested on several experimental settings. It was compared against two baseline approaches which are based on *explicit* representation. Baseline 1 relies on `PlaneInVol` for *3D localization* and *interpolation* as described in Section 5.1.1. Baseline 2 further implements the step of *correction* by SVRTK [172]. Therefore, the two baseline methods encompass the general pipeline of most conventional slice-to-volume ultrasound reconstruction methods. The experimental settings can be summarized as follows:

1. **Reconstruction from Volume-Sampled Images:** 2D cross-sectional images were sampled from native 3D ultrasound volumes (*i.e.* ground-truth volumes). The 3D volumes were then reconstructed back from the 2D images using `ImplicitVol` and the baseline approaches. The reconstructed volumes were then compared with the ground-truth volumes to quantify how similar they were. The estimated 3D locations refined by `ImplicitVol` and those predicted by the baseline approaches were also compared to the ground-truth locations (*i.e.* the 2D images' locations in the native volumes).

2. **Segmentation on Reconstructed Volumes:** Extended from the first experiment, a well-trained deep learning-based segmentation network [162] was employed to segment different anatomical structures from the native 3D volumes (*i.e.* ground-truth) as well as from the volumes reconstructed by `ImplicitVol` and the baseline approaches. The segmentations from the reconstructed volumes were then compared with those from the corresponding native volumes to semantically evaluate the volumetric reconstruction performance.

3. **Reconstruction from Native 2D Ultrasound Images:** To demonstrate the transferability of `ImplicitVol` to the bedside, volumes were reconstructed from images acquired as native 2D freehand video sequences. These videos were acquired following the routine protocol for the fetal brain anomaly scan, typically performed in the second trimester of pregnancy [4]. The 3D volumes were then reconstructed from the 2D video frames using `ImplicitVol`, and

compared the reconstructed volumes with those generated by the baseline approaches.

5.3.2 Implementation Details

The proposed representation network, F_Θ, was a 5-layer MLP, with the hidden layer dimension of 128 and SIREN [55] as the activation function. Encoding dimension, L, from Eq. 5.4, was set to 10 and the set of 3D locations, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, was initialized by the estimated locations predicted by `PlaneInVol`. The learning and decay rates followed those adopted in [166]. Specifically, the initial learning rate was 0.001. It decayed by 0.9954 every 10 epochs when updating the network parameters, Θ, and by 0.9 every 100 epochs when updating the 3D location parameters, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$. A representation network, F_Θ, was trained for one set of images for 10000 epochs to represent one 3D volume. The hyperparameters of the iterative reinitialization, namely e and t, were empirically set to 500 and 10, respectively.

5.3.3 Comparison Baselines

Baseline 1. With the set of 2D ultrasound images, $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$, and the corresponding 3D locations, $\Lambda = \{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^N$, predicted by `PlaneInVol`, $\mathbf{V} \in \mathcal{R}^{H \times W \times D}$ was explicitly reconstructed by interpolating the intensity at each voxel by an inverse distance-weighted average from the 20 nearest pixels of $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$.

Baseline 2. Slice-to-volume registration has been well investigated for super-resolution reconstruction of motion-corrupted MRI. SVRTK [172], designed for fetal brain MRI motion correction, was implemented to the `PlaneInVol`-interpolated volume to verify if using a technique developed for a similar task but in a different modality (*i.e.* MRI) may help in the problem setting in this chapter.

5.3.4 Evaluation metrics

To demonstrate the experimental results, three types of evaluation were employed to quantify different accuracy, which is described as follows:

Volumetric reconstruction accuracy. Structural similarity index measure (SSIM) [171] is a common metric to quantify the similarity between two images. In this study, the 3D version of SSIM was employed to compare the similarity between two 3D volumes, for example the reconstructed and native volumes. SSIM ranges from 0 to 1, where higher value represents higher similarity. Suggested by [173], visual information fidelity (VIF) [174] was also computed to measure the similarity between pair of 2D slices, for example 2D slices sampled from the 3D reconstructed volumes and the corresponding slices sampled from the native volumes. A higher VIF value represents higher similarity.

Positional estimation accuracy. Since `ImplicitVol` refines the estimated 3D locations of the 2D images during volumetric reconstruction, absolute difference between rotation angles (θ_{diff}) and absolute distance between translations (T_{diff}) were used to compare two sets of locations, for example the predicted/refined locations and the ground-truth locations. For both metrics, lower values represent smaller localization differences.

Structural segmentation accuracy. Following the practice of [162], dice similarity coefficient (DSC), 95^{th} percentile Hausdorff distance (H_{95}) and unsigned relative volume differences ($|\Delta V_{rel}|$) were used to compare the segmentations from the reconstructed volumes with those from the native volumes. DSC ranges from 0 to 1, where a higher value corresponds to a better match between the two sets of segmentations, thus suggesting better structural fidelity. For H_{95} and $|\Delta V_{rel}|$, a lower value suggests better performance.

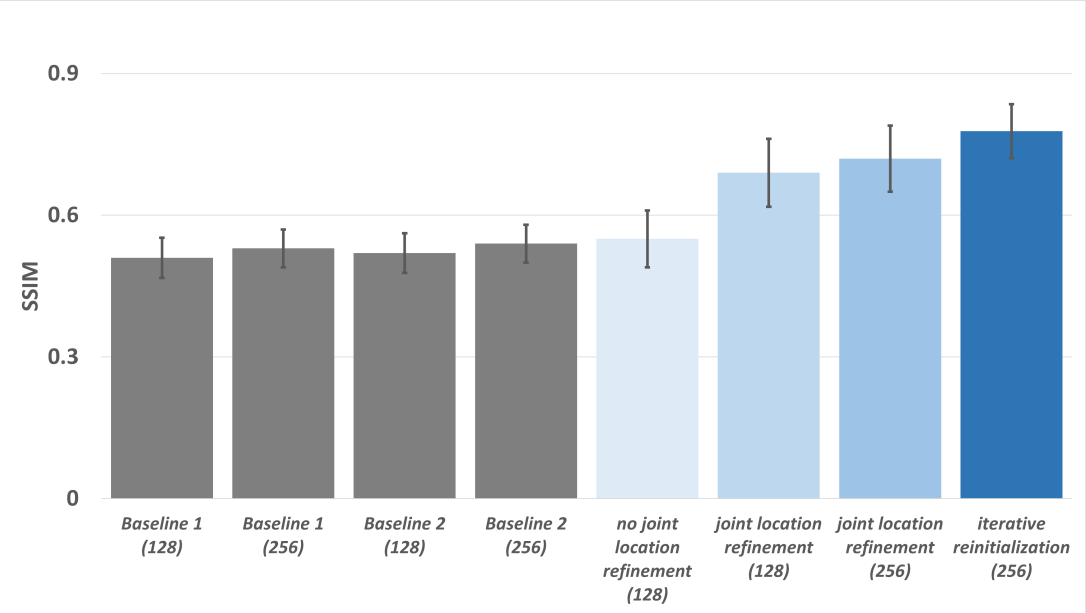
5.3.5 Dataset

Part of the data involved in this chapter overlapped with those described in Chapter 3 and 4. For clarification, they were described again as follows:

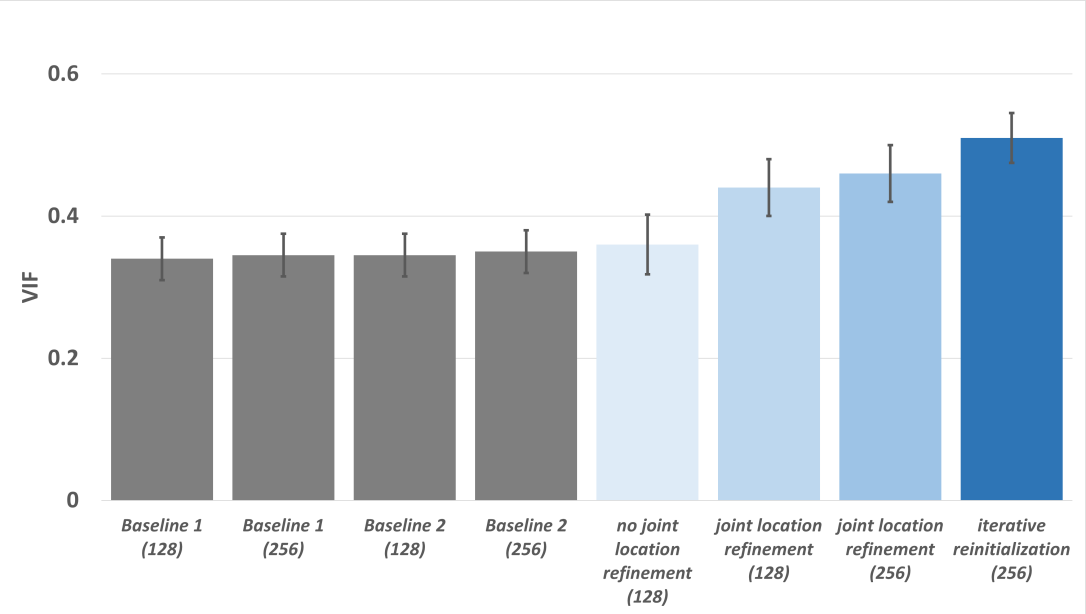
Dataset A. A set of fifteen 3D ultrasound fetal brain volumes ($160 \times 160 \times 160$ voxels at a resolution of $0.6 \times 0.6 \times 0.6$ mm^3), which were obtained as part of the INTERGROWTH-21st study [140]. The volumes were acquired between 19 and 21 gestational weeks, within which the routine fetal anomaly ultrasound scan is recommended by [3]. The 3D ultrasound fetal brain volumes were collected using a Philips HD9 curvilinear probe at a 2–5 MHz wave frequency. Each volume was aligned to a common 3D atlas using [138].

Dataset B. Seventeen 3D ultrasound fetal brain volumes (resize and crop to $160 \times 160 \times 160$ voxels at a resolution of around $0.6 \times 0.6 \times 0.6$ mm^3) were collected using a GE Voluson E10 at the Department of Obstetrics at the Leiden University Medical Center. The volumes were also acquired between 19 and 21 gestational weeks and aligned to the same common 3D atlas of Dataset A using [138] to ensure consistent evaluation.

Native 2D freehand video sequences. Four videos of native freehand 2D brain scans with around 250 frames each were collected at 20 weeks' gestational age at the Leiden University Medical Center using a GE Voluson E10 ultrasound scanner. Each frame was cropped and resized to 160×160 pixels, with a resolution of approximately 0.6×0.6 mm^2.

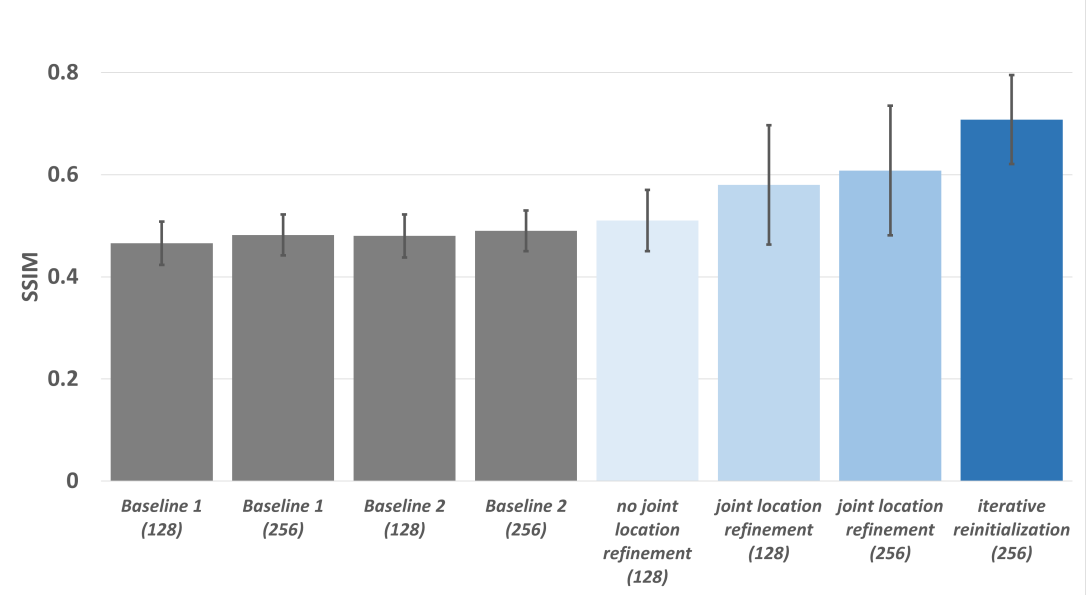


(a) SSIM of Dataset A (higher better)

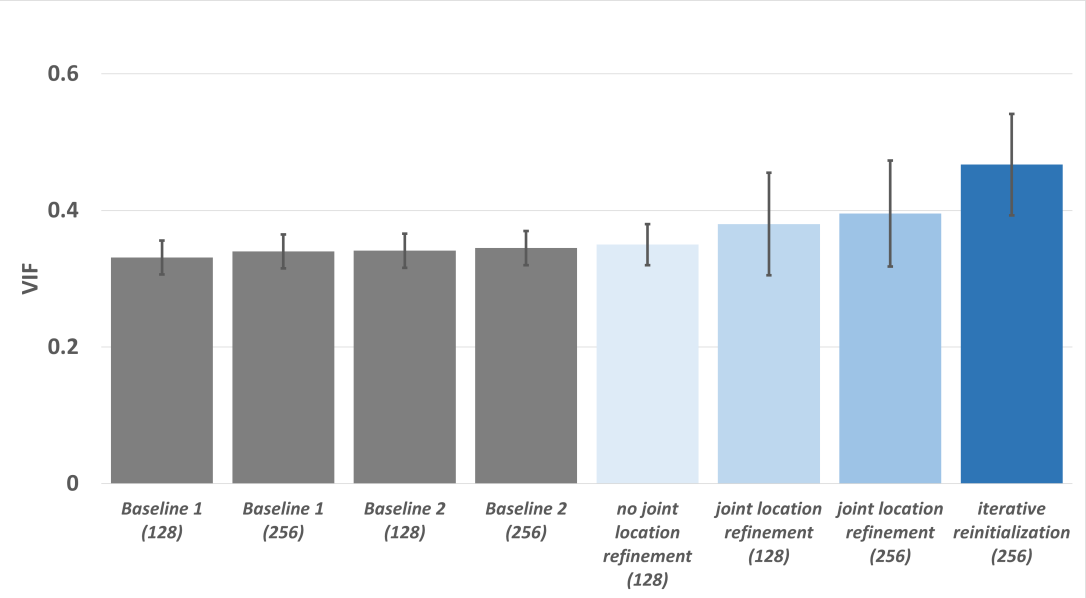


(b) VIF of Dataset A (higher better)

Figure 5.2: Volume reconstruction results of Dataset A. The gray bars represent the baseline methods and the blue bars represent `ImplicitVol` with ablation studies. The number in the bracket is N, the number of 2D slices used for volumetric reconstruction.



(a) SSIM of Dataset B (higher better)



(b) VIF of Dataset B (higher better)

Figure 5.3: Volume reconstruction results of Dataset B. The gray bars represent the baseline methods and the blue bars represent `ImplicitVol` with ablation studies. The number in the bracket is N, the number of 2D slices used for volumetric reconstruction.

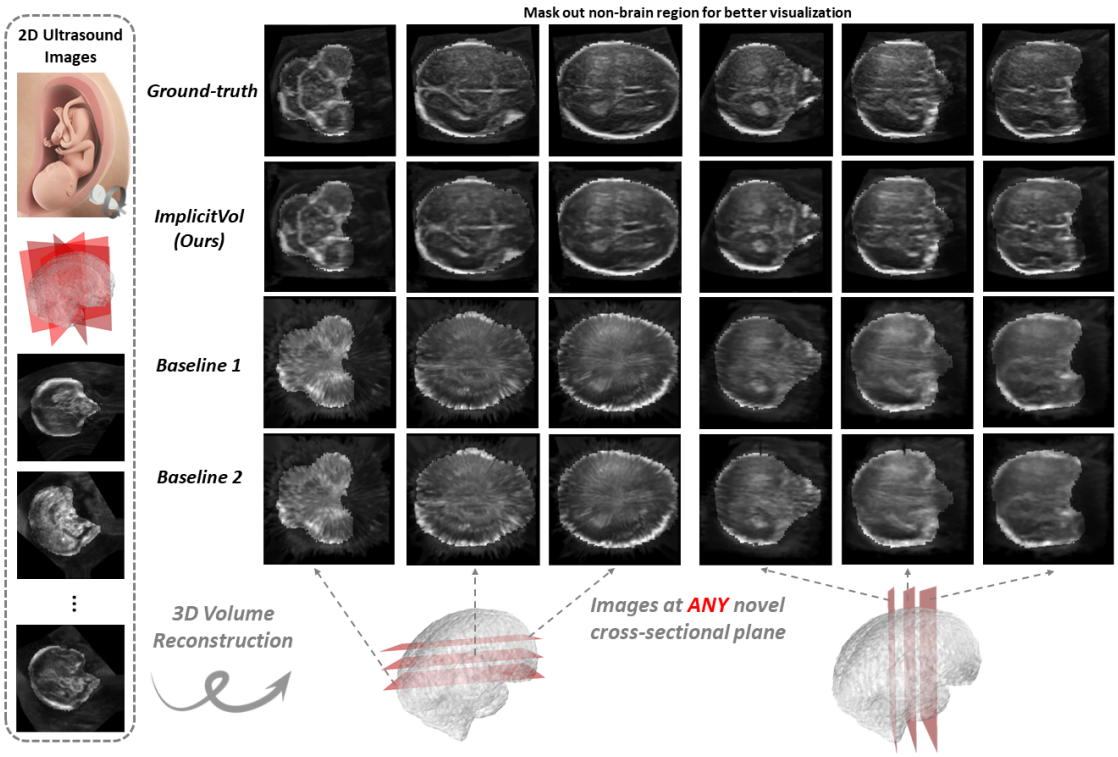


Figure 5.4: Visualization of 3D reconstruction from volume-sampled testing images by different approaches. Images sampled at different novel cross-sectional views, masked by [175] for better visualization, are presented and compared with the ground-truth.

5.4 Results

5.4.1 Reconstruction from Volume-Sampled Images

Two-dimensional cross-sectional images were sampled from native 3D ultrasound volumes from two independently-acquired datasets: Dataset A and Dataset B (details in Section 5.3.5). For each volume, a set of 2D images ($N \in \{128, 256\}$ for each dataset, respectively) were non-uniformly sampled around the central axis of the brain, to simulate actual freehand acquisition by rotating the probe (Fig. 5.4).

Three-dimensional volumes and 2D slices sampled at new cross-sectional views along the *coronal*, *sagittal* and *axial* directions from both the native (*i.e.* ground-truth) and reconstructed volumes, were analyzed. The quantitative results are presented in Figs. 5.2a and 5.2b for Dataset A and Figs. 5.3a and 5.3b for Dataset B, while the qualitative results are presented in Fig. 5.4. The reconstructed volumes were rigidly aligned to the ground-truth volume for fairer comparison, as global

rigid shifts may be introduced to the volumes during the location refinement of the reconstruction.

Overall, from Figs. 5.2 and 5.3, the 3D volumes reconstructed from the proposed approach, `ImplicitVol` (blue bars), showed a better match with the corresponding ground-truth as suggested by both the higher SSIM (Figs. 5.2a and 5.3a) and VIF values (Figs. 5.2b and 5.3b), outperforming the baseline approaches (gray bars). For both datasets, there was a boost in performance by over 50% (SSIM) and 40% (VIF) when comparing `ImplicitVol` with the baselines. This can be qualitatively verified in Fig. 5.4, where the 2D slices sampled from volumes reconstructed by the baseline approaches displayed visible artifacts when compared to those sampled from the ground-truth volumes.

When joint location refinement and the proposed iterative reinitialization were excluded (leftmost blue bars in Figs. 5.2 and 5.3), `ImplicitVol` achieved comparable reconstruction performance as the baseline. Adding the joint location refinement (second left blue bars in Figs. 5.2 and 5.3) resulted in a performance gain of 24.5% (Dataset A, SSIM), 15.6% (Dataset B, SSIM), 22.2% (Dataset A, VIF) and 9% (Dataset B, VIF). While increasing the training set from 128 to 256 slices (third left blue bars in Figs. 5.2 and 5.3) led to a slight increase in performance, with the full version of `ImplicitVol` (rightmost blue bars in Figs. 5.2 and 5.3) achieving the best performance, and leading to a further improvement of 8.1% (Dataset A, SSIM), 16.7% (Dataset B, SSIM), 10.9% (Dataset A, VIF) and 17.9% (Dataset B, VIF).

`ImplicitVol` was compared with two baseline approaches, which encompass the general pipeline (Section 5.1.1) of most conventional approaches that rely on *explicit* representation. The core part of Baseline 1 is to localize each 2D images in the 3D space by `PlaneInVol` (Chapter 3), such that registration and interpolation can be performed post-hoc. As mentioned in Section 5.1.1, other existing approaches [150, 151, 156, 157] may also achieve a similar task (*i.e.* localization in 3D space). However, most of these require external sensor tracking in either the training or inference stages, whereas `ImplicitVol` is a fully sensor-free framework, which may integrate seamlessly into the clinical workflow.

`ImplicitVol` refines the 3D locations of the 2D images while *simultaneously* learning the volumetric reconstruction. This approach yielded superior performance over methods that implemented *3D localization*, *interpolation* and *correction* separately (*i.e.* Baseline 2). The paucity of reconstruction studies for ultrasound imaging data, and lack of open-source software, made it difficult to reproduce the results of existing works. Nevertheless, the task for which SVRTK [172] was proposed (*i.e.* super-resolution reconstruction of motion-corrupted MRI) is relevant to the presented work, and the hyperparameters have been optimally fine-tuned in the experiments (*i.e.* Baseline 2).

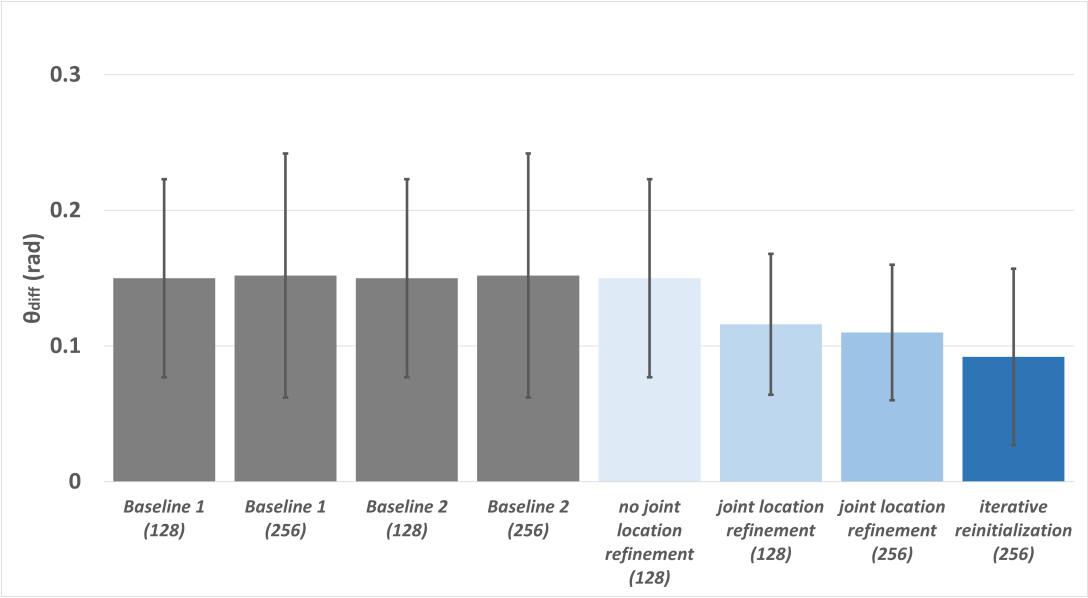
5.4.2 Location Refinement from Volume-Sampled Images

Besides the reconstruction performance, the 3D localization accuracy of the different approaches was also assessed by comparing the estimated 3D locations refined by `ImplicitVol` to the ground-truth locations, and those predicted by other baseline approaches. The results are presented in Figs. 5.5a and 5.5b for Dataset A and Figs. 5.6a and 5.6b for Dataset B.

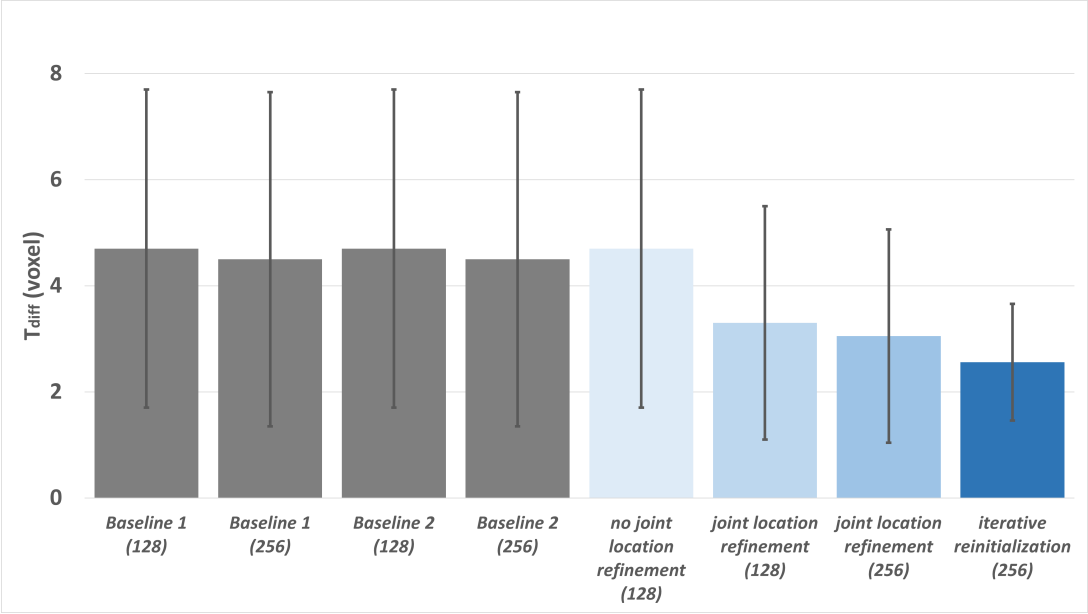
The estimated 3D locations refined by `ImplicitVol` showed a significant improvement over those predicted by the baseline approaches by over 35% (θ_{diff}) and 40% (T_{diff}) for both datasets. Similar to the experimental results of the volumetric reconstruction, ablation studies suggested that the joint location refinement and iterative re-initialization of `ImplicitVol` have led to the respective performance gains. Note that, such refinement requires no extra supervision cost, which manifests `ImplicitVol`'s additional potential in slice-to-volume registration of ultrasound, for example strengthening `PlaneInVol` (Chapter 3).

5.4.3 Structural segmentation on reconstructed volumes

Semantic-level evaluation was conducted by using a 3D segmentation network [162], which was trained (to convergence) with a dataset of volumes acquired with a

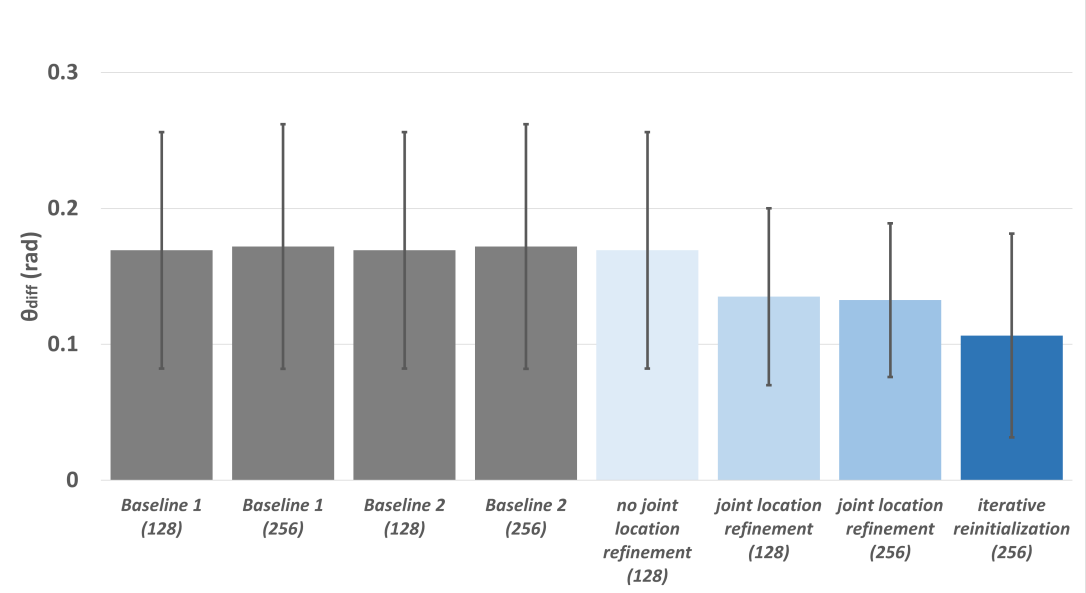


(a) θ_{diff} of Dataset A (lower better)

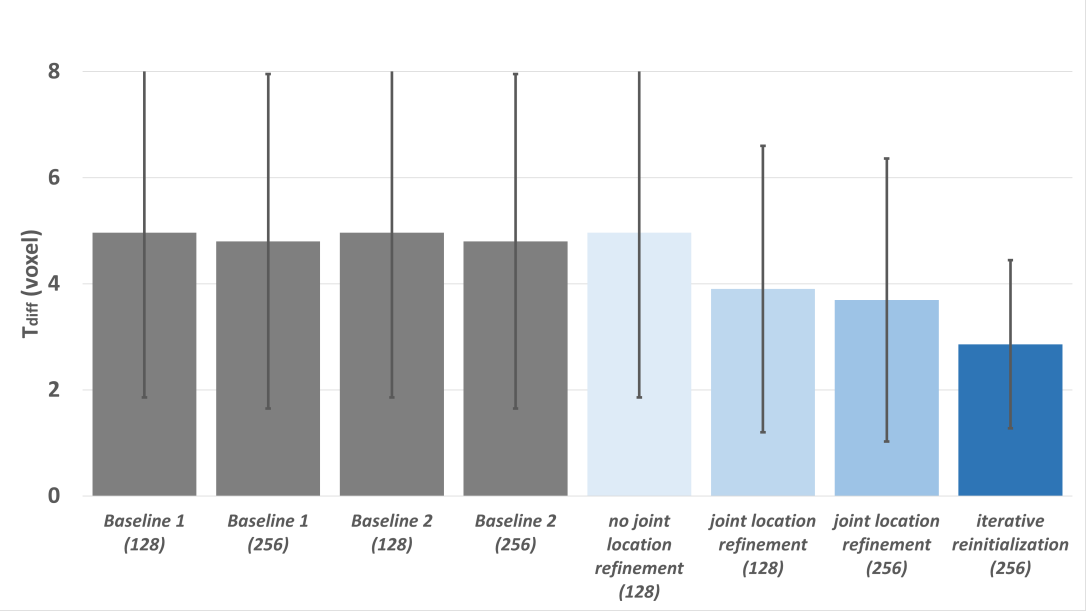


(b) T_{diff} of Dataset A (lower better)

Figure 5.5: Location refinement results of Dataset A. The gray bars represent the baseline methods and the blue bars represent `ImplicitVol` with ablation studies. The number in the bracket is N, the number of 2D slices used for volumetric reconstruction.

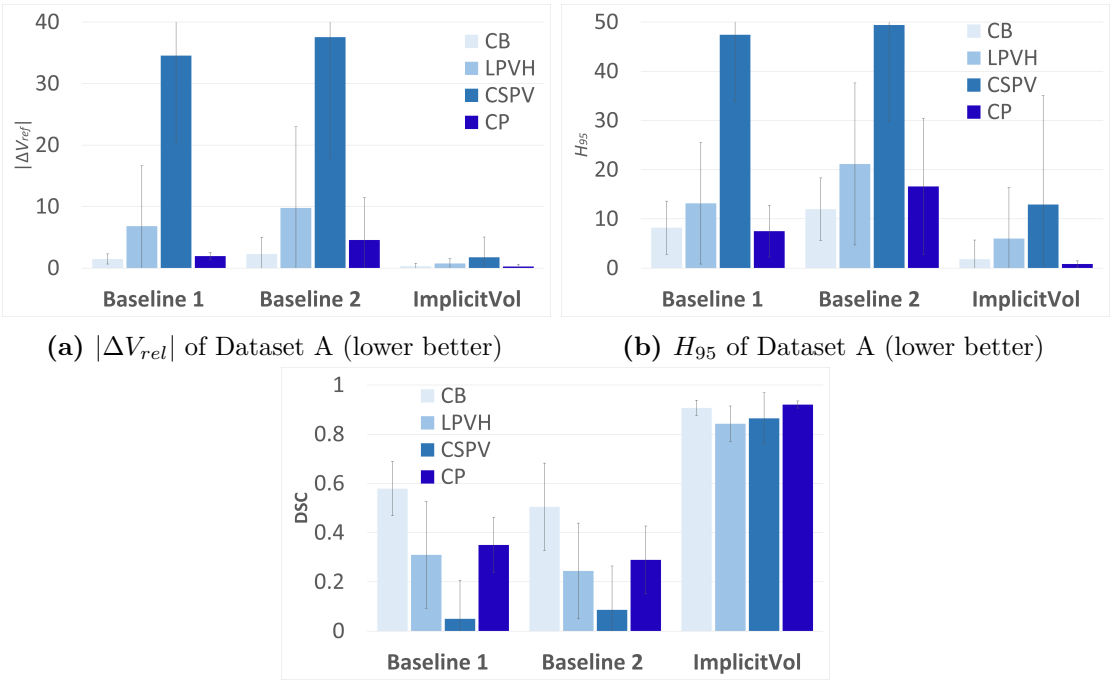


(a) θ_{diff} of Dataset B (lower better)



(b) T_{diff} of Dataset B (lower better)

Figure 5.6: Location refinement results of Dataset B. The gray bars represent the baseline methods and the blue bars represent `ImplicitVol` with ablation studies. The number in the bracket is N, the number of 2D slices used for volumetric reconstruction.



(a) $|\Delta V_{rel}|$ of Dataset A (lower better)

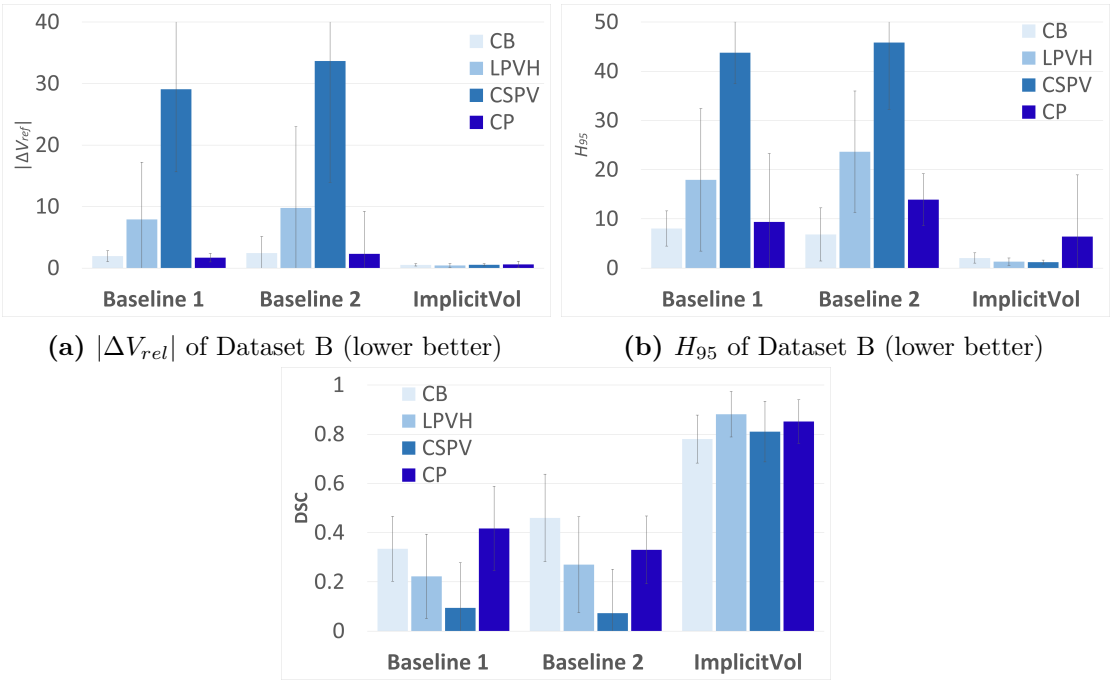
(b) H_{95} of Dataset A (lower better)

(c) DSC of Dataset A (higher better)

Figure 5.7: Quantitative results of segmentation from 3D volumes of Dataset A reconstructed by different approaches. Four subcortical structures: the choroid plexus (CP), lateral posterior ventricle horns (LPVH), cavum septum pellucidum et vergae (CSPV), and cerebellum (CB) were segmented and analyzed.

3D transducer. The network was applied to segment four subcortical structures: the choroid plexus (CP), lateral posterior ventricle horns (LPVH), cavum septum pellucidum et vergae (CSPV), and cerebellum (CB) in the native and reconstructed 3D volumes. Quantitative results in Figs. 5.7a to 5.7c for Dataset A and Figs. 5.8a to 5.8c for Dataset B, showed a better semantic similarity to `ImplicitVol`, when compared to the baseline approaches. This was verified by all the three evaluation metrics for all four segmented subcortical structures. This is further supported by the qualitative comparison in Fig. 5.9, where segmentations of Baselines 1 and 2 showed an obvious mismatch with those of the ground-truth (*i.e.* native volumes), for one representative example.

This novel evaluation method is conceptually similar to perceptual loss [176] and different CNN-based image quality metrics [177, 178], which utilize another trained neural network to quantify the high-level image quality. In practice, segmentation is usually required in many downstream tasks on the ultrasound volumes, for



(a) $|\Delta V_{rel}|$ of Dataset B (lower better)

(b) H_{95} of Dataset B (lower better)

(c) DSC of Dataset B (higher better)

Figure 5.8: Quantitative results of segmentation from 3D volumes of Dataset B reconstructed by different approaches. Four subcortical structures: the choroid plexus (CP), lateral posterior ventricle horns (LPVH), cavum septum pellucidum et vergae (CSPV), and cerebellum (CB) were segmented and analyzed.

example structural analysis and volumetric evaluation. The superior performance of `ImplicitVol` over other baseline approaches may also suggest its potential in different clinical applications.

5.4.4 Volumetric reconstruction on native freehand sweeps

In order to assess the sensitivity of `ImplicitVol` to motion artifacts (e.g. transducer jitter, fetal movement), volumes were reconstructed from video sequences (sweeps) with noise added to the positional information (Fig. 5.10). As shown in Fig. 5.11, images sampled from `ImplicitVol` showed better visual quality in motion-corrupted regions (red boxes), thanks to the localization refinement achieved by the joint optimization. `ImplicitVol` also performed better in under-sampled regions (yellow boxes), where the freehand scanning yielded limited coverage of the region of interest.

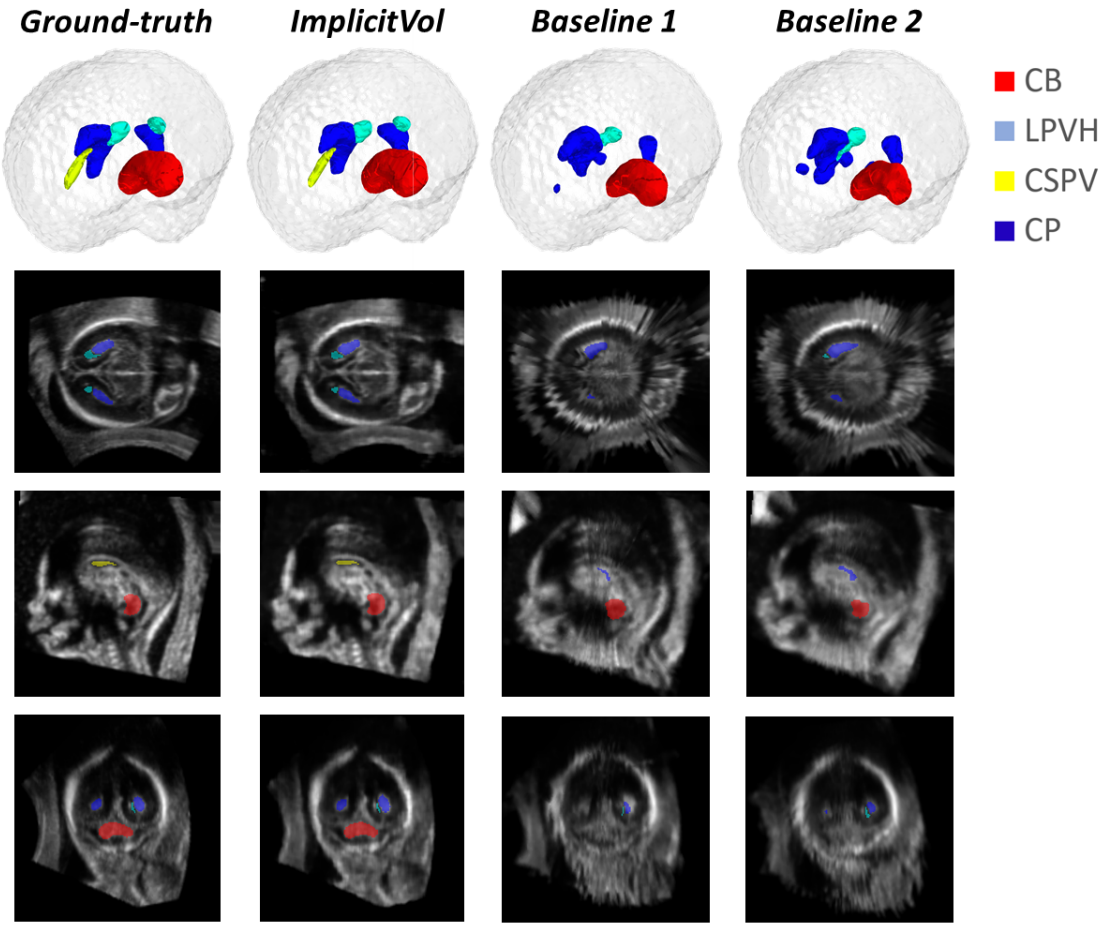
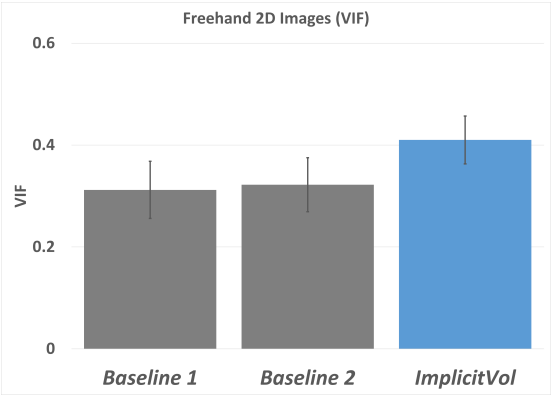


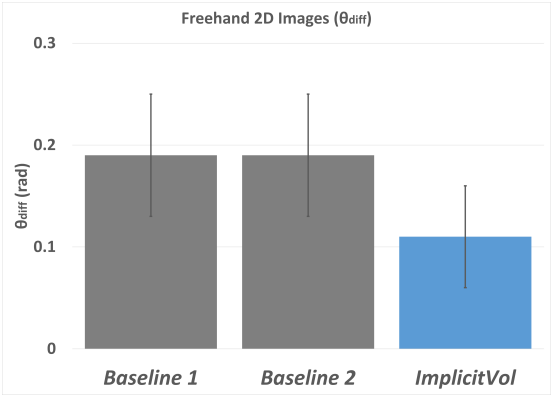
Figure 5.9: Visualization of quantitative segmentation results. Segmentation of CB, LPVH, CSPV and CP by different approaches are presented in the axial, coronal and sagittal planes. Results from the two baselines showed an obvious mismatch with those of the ground-truth, suggesting the difference in the quality of volumetric reconstruction by different approaches.

In such areas, the baseline approaches reconstructed misleading results due to the inaccuracy caused by extrapolation from spatially distant neighbours.

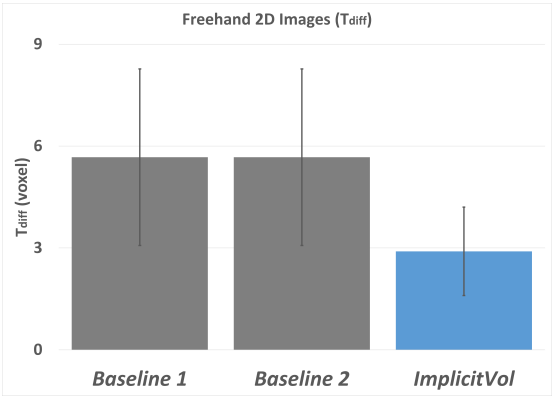
Since there were no native (*i.e.* ground-truth) volumes associated with these native freehand 2D ultrasound sequences, quantitative analysis was performed by using 80% of each 2D ultrasound sequence for volumetric reconstruction and the remaining 20% for testing. These testing images were compared (*i.e.* with VIF) to the corresponding slices sampled from the reconstructed volumes at the estimated 3D locations. Over 30% of improvement was achieved by `ImplicitVol` comparing to the baseline approaches (Fig. 5.10a). The refined 3D position estimations of the testing images were also compared with the original predicted 3D locations



(a) VIF of native freehand ultrasound images (higher better)



(b) θ_{diff} of native freehand ultrasound images (lower better)



(c) T_{diff} of native freehand ultrasound images (lower better)

Figure 5.10: Quantitative results of volumetric reconstruction from native freehand 2D ultrasound images.

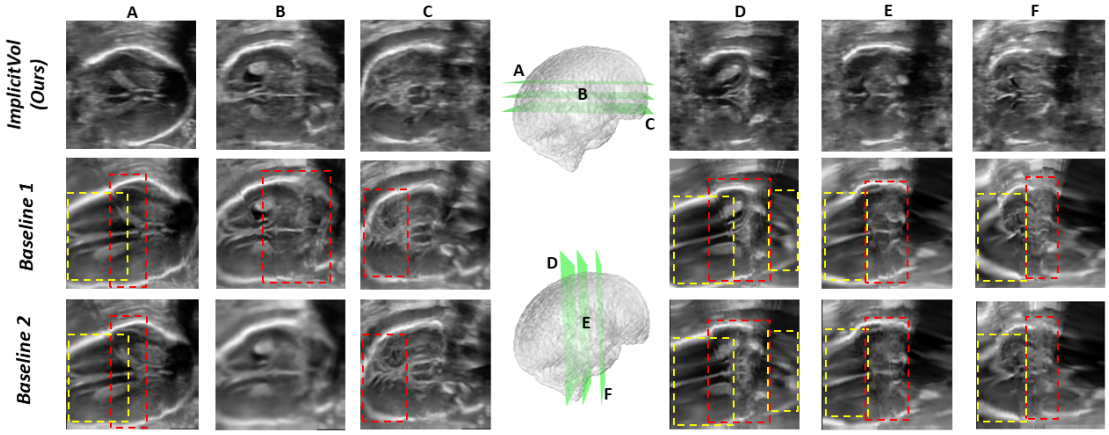


Figure 5.11: Qualitative results of 3D reconstruction from native freehand 2D ultrasound. Novel view images sampled from different planes from volumes reconstructed by different approaches are presented. `ImplicitVol`, shows better visual quality in under-sampled region (yellow) and is more robust against inaccurate position estimation (red).

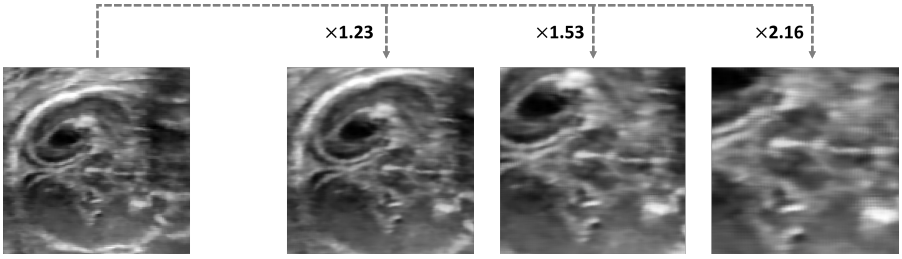


Figure 5.12: `ImplicitVol` is able to reconstruct the volumes at arbitrarily high resolutions. Image on the left is sampled at the original resolution, while the three images on the right are sampled at arbitrarily higher resolutions.

to quantify the 3D localization accuracy. From Figs. 5.10b and 5.10c, there is an evident performance boost of around 50% when comparing `ImplicitVol` with the baselines. These results verified that `ImplicitVol` not only outperforms the baseline approaches on volume-sampled ultrasound images, but also shows superior volumetric reconstruction performance on native freehand 2D ultrasound images. `ImplicitVol`'s ability to reconstruct 3D volumes at arbitrarily high resolutions (Fig. 5.12) was also demonstrated. This is enabled by the continuous property of the *implicit* representation and may facilitate zooming in for closer inspection of anatomies of interest.

5.5 Conclusion

In this chapter, `ImplicitVol`, a sensor-free approach to reconstruct 3D ultrasound volumes from a sparse set of 2D images with deep implicit representation, is proposed. Instead of using standard slice-to-volume ultrasound reconstruction by registering the 2D scans into the 3D volumes, and explicitly performing interpolations in the resulting volumetric representation, a novel slice-to-volume reconstruction pipeline based on implicit representation by parameterising the 3D volume as a deep neural network is proposed, which jointly refines the slice-to-volume registrations and learns a full 3D reconstruction based on a set of 2D scans.

The proposed framework outperforms conventional reconstruction approaches, in terms of the quality of the reconstruction, both directly and semantically, as well

as the refinement of the 3D localization, which highlights `ImplicitVol`'s additional potential in slice-to-volume registration. It is flexible, and capable of accurately reconstructing volumes from native freehand 2D ultrasound images, without a tracking sensor, causing minimal interruption to the routine scanning protocols while providing richer information for diagnosis and evaluation of the 3D anatomies. This strengthens the potential of the framework proposed in this thesis, from 2D (Chapter 3 and 4) to 3D (this chapter).

6

Volumetric Segmentation from a Single Slice Annotation

Following from the previous chapter that considered volumetric reconstruction, a more general problem regarding 3D medical volumes, namely semi-automatic segmentation, is investigated in this chapter. Specifically, `Sli2Vol`, a self-supervised learning framework trained with just raw 3D volumes, which can be used to propagate a single-slice annotation to the whole 3D volume, for any structure across different modalities, is presented. The work presented in this chapter has been published in:

Yeung, P.H., Namburete, A.I., and Xie, W,: Sli2Vol: Annotate a 3D Volume from a Single Slice with Self-Supervised Learning., *International conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

Contents

6.1 Introduction

In previous chapters, the limitation of 2D image, namely the failure to capture rich contextual volumetric information, has been discussed. Such limitation can be overcome by using 3D volumes. Nevertheless, 3D medical volume, on its own, may not be sufficient for the diagnosis and analysis. Image segmentation is commonly required, making it arguably one of the most important tasks in medical image analysis, as it identifies the structure of interest (SOI) with arbitrary shape (*i.e.* pixel level predictions), encompassing rich information, such as the position and size. In recent years, the development and application of different ConvNets, for example U-Net [67], have significantly boosted the accuracy of computer-aided medical image segmentation.

Nevertheless, training fully automatic segmentation models comes with several limitations: *firstly*, acquiring annotations for the training volumes are usually costly and time consuming; *secondly*, once domain shift appears, (*i.e.* from differences in scanner, acquisition protocol or the SOI varies during inference), the model may suffer a catastrophic drop in performance, requiring new annotations and additional fine-tuning. These factors have limited the use of the automatic segmentation approaches to applications with inter-vendor and inter-operator variance. As an alternative, semi-automatic approaches are able to operate interactively with the end users: this is the scenario considered in this chapter. Specifically, the goal is to segment any *arbitrary* SOIs in 3D volumes while only requiring manual annotation of a *single* slice within the volume, which may facilitate more flexible analysis of

arbitrary SOIs with the desired generalizability (*e.g.* inter-scanner variability), and significantly reduce the annotating cost for fully supervised learning.

Similar tools have been developed with level set or random forest methods, which show excellent performance as reported in [179–183]. However, implementation of specific regularization and heavy parameter-tuning are usually required for different SOIs, limiting its use in practice. On the other hand, related work in medical image registration explores the use of pixelwise correspondence from optical flow [184, 185] or unsupervised approaches [163, 186, 187], which in principle could be harnessed for the propagation of a 2D mask between slices within a volume. However, they are prone to error drift, *i.e.* error accumulation, introduced by inter-slice propagation of registration errors.

In this chapter, the approach taken is to propagate a labelled 2D slice of segmentation through the entire 3D volume by matching correspondences between consecutive slices. This chapter makes the following contributions: *first*, mask propagation approaches based on unsupervised/self-supervised registration of slices, namely, naïve optical flow [188] and VoxelMorph [163], and the proposed self-supervised `Sli2Vol`, which is based on learning to match slices' correspondences [189, 190], are explored. *Second*, to alleviate the error accumulation in mask propagation, a simple verification module is proposed and exploited for refining the mask during inference time. *Third*, `Sli2Vol` is benchmarked on 8 public CT and MRI datasets [191–194], spanning 9 anatomical structures. Without any parameter-tuning, a *single* `Sli2Vol` model achieves Dice scores (0-100 scale) above 80 for most of the benchmarks, which outperforms other supervised and unsupervised approaches for all datasets in cross-domain evaluation. To the best of my knowledge, this is the first study to undertake cross-domain evaluation on such large-scale and diverse benchmarks for semi-automatic segmentation approaches, which shifts the focus to *generalizability* across different devices, clinical sites and anatomical SOIs.

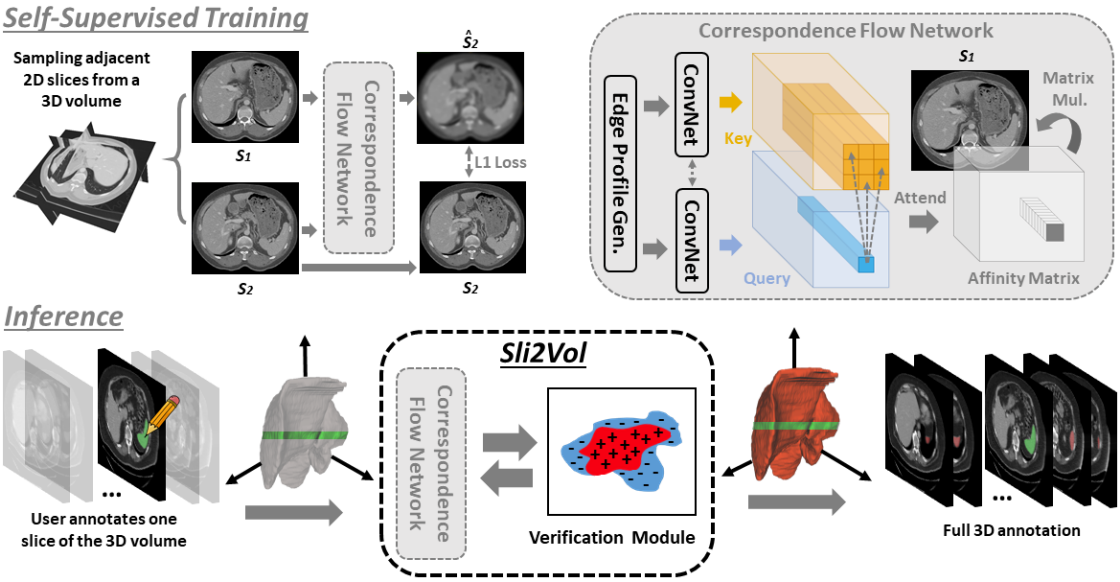


Figure 6.1: Pipeline of `Sli2Vol`. During *self-supervised training*, pair of adjacent slices sampled from 3D volumes are used to train a correspondence flow network. Provided with the 2D mask of a single slice of a volume, the trained network with the verification module can be used to propagate the initial annotation to the whole volume during *inference*.

6.2 Methods

In Section 6.2.1, the problem setting in this chapter, namely semi-automatic segmentation for 3D volume with *single* slice annotation, is first formulated. Next, the training stage of the proposed approach, `Sli2Vol`, is introduced in Section 6.2.2 and the proposed edge profile generator in Section 6.2.3. This is followed by the computations for inference (6.2.4), including the proposed verification module (6.2.5).

6.2.1 Problem Setup

In general, given a 3D volume, denoted by $\mathbf{V} \in \mathbb{R}^{H \times W \times D}$, where H, W and D are the height, width and depth of the volume, respectively, the goal is to segment the SOI in the volume based on a user-provided 2D segmentation mask for the *single* slice, *i.e.* $\mathbf{M}_i \in \mathbb{R}^{H \times W \times 1}$ with $1's$ indicating the SOI, and $0's$ as background. The outputs will be a set of masks for an individual slice, *i.e.* $\{\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_D\}$.

Inspired by [189, 190], this problem is formulated as learning feature representations that establish robust pixelwise correspondences between adjacent slices in a 3D volume, which results in a set of affinity matrices, $\mathbf{A}_{i \to i+1}$, for propagating the 2D mask between consecutive slices by *weighting and copying*. Model training follows a self-supervised learning scheme, where only raw volume data is used, and only one slice annotation is required during inference time.

6.2.2 Self-Supervised Training of Sli2Vol

This section includes the details of the self-supervised approach for learning the dense correspondences. Conceptually, the idea is to task a deep network for slice reconstruction by *weighting and copying* pixels from its neighboring slice. The affinity matrices used for weighting are acquired as a by-product, and can be directly used for mask propagation during inference.

During training, a pair of adjacent slices, $\{\mathbf{S}_1, \mathbf{S}_2\}, \mathbf{S}_i \in \mathbb{R}^{H \times W \times 1}$, are sampled from a training volume, and then fed to a ConvNet, parametrized by $\psi(\cdot; \theta)$ (as shown in the upper part of Fig. 6.1):

$$[\mathbf{k}_1,\ \mathbf{q}_2] = [\psi(g(\mathbf{S}_1);\ \theta),\ \psi(g(\mathbf{S}_2);\ \theta)] \tag{6.1}$$

where $g(\cdot)$ denotes an *edge profile generator* (details in Section 6.2.3) and $\mathbf{k}_1, \mathbf{q}_2 \in \mathbb{R}^{H \times W \times c}$ refer to the feature representation (c channels) computed from corresponding slices, termed as *key* and *query* respectively (Fig. 6.1). The difference in notation (*i.e.* $\mathbf{q}$ and $\mathbf{k}$) is just for emphasizing their functional difference.

Reshaping $\mathbf{k}_1$ and $\mathbf{q}_2$ to $\mathbb{R}^{HW \times c}$, an affinity matrix, $\mathbf{A}_{1 \to 2} \in \mathbb{R}^{HW \times \delta}$, is computed to represent the feature similarity between the two slices (Fig. 6.1):

$$\mathbf{A}_{1 \to 2}(u, v) = \frac{exp\langle \mathbf{q}_2(u, :), \mathbf{k}_1(v, :) \rangle}{\sum_{\lambda \in \Omega} exp\langle \mathbf{q}_2(u, :), \mathbf{k}_1(\lambda, :) \rangle} \tag{6.2}$$

where $\langle \cdot, \cdot \rangle$ is the dot product between two vectors and Ω is the window surrounding pixel v (*i.e.* in $\mathbb{R}^{H \times W}$ space) for computing local attention, with $n(\Omega) = \delta$.

Loss Function. During training, $\mathbf{A}_{1\to2}$ is used to *weight and copy* pixels from $\mathbf{S}_1$ (*i.e.* reshape to $\mathbb{R}^{HW\times1}$) to reconstruct $\mathbf{S}_2$, denoted as $\hat{\mathbf{S}}_2$, by:

$$\hat{\mathbf{S}}_2(u,1) = \sum_{v}^{\Omega} \mathbf{A}_{1\to2}(u,v)\mathbf{S}_1(v,1). \tag{6.3}$$

Mean absolute error (MAE) between $\mathbf{S}_2$ and $\hat{\mathbf{S}}_2$ is applied as the training loss.

6.2.3 Edge Profile Generator

Essentially, the basic assumption of the above-mentioned idea is that, to better reconstruct $\mathbf{S}_2$ via copying pixel from $\mathbf{S}_1$, the model must learn to establish reliable correspondences between the two slices. However, naïvely training the model may actually incur trivial solutions, for example, the model can perfectly solve the reconstruction task by simply matching the *pixel intensity* of $\mathbf{S}_1$ and $\mathbf{S}_2$.

In Lai *et al.* [189, 190], the authors showed that input color channel (*i.e. RGB* or *Lab*) dropout is an effective information bottleneck, which breaks the correlation between the color channels and forces the model to learn more robust correspondences. However, this is usually not feasible in medical images, as only single input channel is available in most of the modalities.

A *profile of edges* is proposed to be used as an *information bottleneck* to avoid trivial solution. Specifically, for each pixel, its intensity value is converted to a normalized edge histogram, by computing the derivatives along d different directions at s different scales, *i.e.* $g(\mathbf{S}_i) \in \mathbb{R}^{H\times W\times(d\times s)}$, followed by a *softmax* normalization through all the derivatives. Intuitively, $g(\cdot)$ explicitly represents the edge distributions centered each pixel of the slice $\mathbf{S}_i$, and force the model to pay more attentions to the edges during reconstruction. Experimental results in Section 6.4.3 verify the essence of this design in improving the model performance.

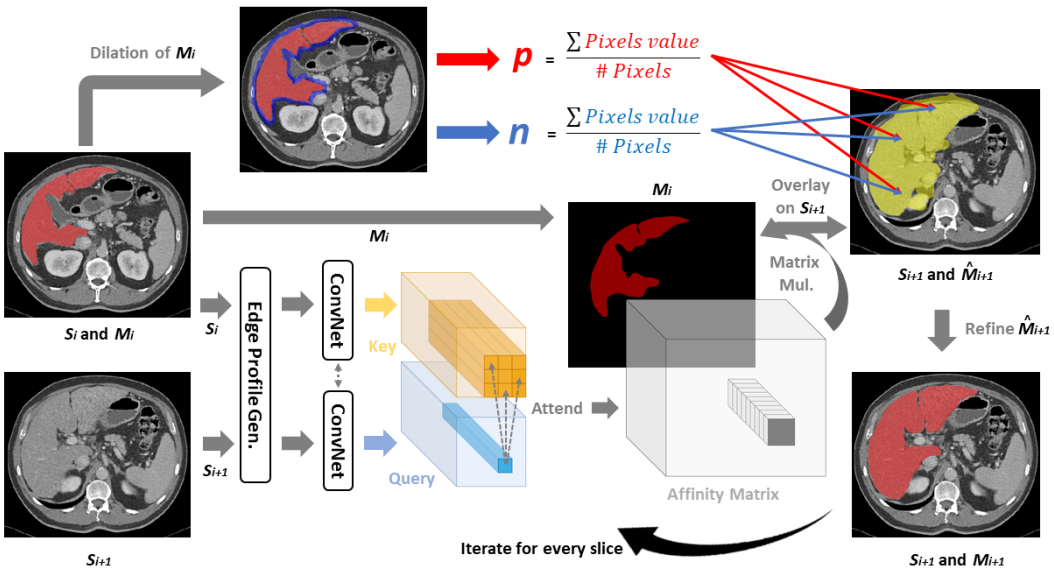


Figure 6.2: Computation of each iteration of `Sli2Vol` during *inference*. $\{\mathbf{S}_i, \mathbf{S}_{i+1}\}$, sampled from $\mathbf{V}$ are fed into the trained correspondence flow network to obtain the affinity matrix to propagate $\mathbf{M}_i$ to $\hat{\mathbf{M}}_{i+1}$. $\hat{\mathbf{M}}_{i+1}$ is then refined by $\mathbf{p}$ and $\mathbf{n}$, obtained by $\mathbf{M}_i$ and $\mathbf{S}_i$, to get the final mask, $\mathbf{M}_{i+1}$.

6.2.4 Inference

Given a volume, $\mathbf{V}$ and an initial mask at the i-th slice, $\mathbf{M}_i$, the affinity matrix, $\mathbf{A}_{i\to i+1}$, output from $\psi(\cdot;\theta)$ is used to propagate $\mathbf{M}_i$ iteratively to the whole $\mathbf{V}$.

Specifically, two consecutive slices, $\{\mathbf{S}_i, \mathbf{S}_{i+1}\}$, are sampled from the volume $\mathbf{V}$ and fed into $\psi(g(\cdot);\theta)$ to get $\mathbf{A}_{i\to i+1}$, which is then used to propagate $\mathbf{M}_i$, using Eq. 6.3, ending up with $\hat{\mathbf{M}}_{i+1}$. This set of computations (Fig. 6.2) is then repeated for the next two consecutive slices, $\{\mathbf{S}_{i+1}, \mathbf{S}_{i+2}\}$, in either direction, until the whole volume is covered.

6.2.5 Verification Module

In practice, it is found that directly using $\hat{\mathbf{M}}_{i+1}$ for further propagation will potentially accumulate the prediction error after each iteration. To alleviate this drifting issue, and further boost the performance, a simple verification module is proposed to correct the mask after each iteration of mask propagation.

Specifically, two regions, namely positive ($\mathbf{P} \in \mathbb{R}^{H \times W}$) and negative ($\mathbf{N} \in \mathbb{R}^{H \times W}$) regions, are constructed. $\mathbf{P}$ refers to the delineated SOI in $\mathbf{M}_i$, and $\mathbf{N}$ is identified by subtracting $\mathbf{P}$ from its own morphologically dilated version. Intuitively, the negative region denotes the thin and non-overlapping region surrounding $\mathbf{P}$ (Fig. 6.2). The *mean intensity value* is maintained within each region:

$$p = \frac{1}{|P_i|}\langle P_i, S_i \rangle \qquad n = \frac{1}{|N_i|}\langle N_i, S_i \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes Frobenius inner product, p and n refer to the positive and negative query values respectively.

During inference time, assuming $\hat{\mathbf{M}}_{i+1}$ is the predicted mask from the propagation, each of the proposed foreground pixels u in $\mathbf{S}_{i+1}$, is then compared to p and n and being re-classified according to its distance to the two values by:

$$\mathbf{M}^u_{i+1} = \begin{cases} 1, & \text{if } \hat{\mathbf{M}}^u_{i+1} = 1 \text{ and } \sqrt{(\mathbf{S}^u_{i+1} - p)^2} < \sqrt{(\mathbf{S}^u_{i+1} - n)^2} \\ 0, & \text{otherwise} \end{cases} \tag{6.4}$$

This set of computations is then repeated for the next round of propagation, where p and n are updated using the corrected mask, $\mathbf{M}_{i+1}$, and $\mathbf{S}_{i+1}$.

6.3 Experiment

The framework, `Sli2Vol`, was benchmarked on 8 different public datasets, spanning 9 different SOIs, and compare with a variety of fully supervised and semi-automatic approaches. In Section 6.3.1, the datasets used in this chapter are introduced. In Section 6.3.2, the experiments conducted for this study are summarized.

6.3.1 Dataset

Four training and eight testing datasets were involved. For **chest and abdominal CT**, a *single* model was trained on 3 unannotated dataset (*i.e.* C4KC-KiTS [195], CT-LN [196] and CT-Pancreas [197]) and tested on 7 other datasets (*i.e.* Sliver07 [191], CHAOS [192], 3Dircadb-01, 02 [193], and Decath-Spleen, Liver and Pancreas [194]).

For **cardiac MRI**, models were trained on the 2D video dataset from Kaggle [198], and tested on a 3D volume dataset (*i.e.* Decath-Heart [194]), which manifested large domain shift. The details of the datasets are summarized as follows and in Table 6.1.

Chest and Abdominal CT (Training)

- **C4KC-KiTS [195]**: The CT volumes were acquired from patients undergoing either partial or radical nephrectomy for at least one kidney tumor [199]. All the patients received treatments at the University of Minnesota Medical Center but the scans were collected by different scanners with different acquisition protocols and hence heterogeneity existed. The thickness of the CT slice ranged from 1mm to 5mm. 300 patients participated in the CT scanning, resulting in 300 volumes being collected. 210 of them were released as the training data for the KiTS19 Challenge and they were used as part of the training data in this chapter. For the KiTS19 Challenge, manual segmentation labels of the kidney and tumor were included but they were not used in this study.

- **CT-LN [196]**: 86 CT scans, with a total of 595 manually annotated abdominal lymph nodes, were collected from the National Institutes of Health, Clinical Center. For the purpose of this study, only the CT volumes were used for training.

- **CT-Pancreas [197]**: The CT volumes were acquired from 53 male and 27 female subjects (ages ranged from 18 to 76 years old, with a mean of 46.8 years old) at the National Institutes of Health Clinical Center. All of them did not have major abdominal nor pancreatic diseases. The CT scans were collected by different scanners (*i.e.* Philips and Siemens MDCT scanners) and the thickness of the CT slice ranged from 1.5mm to 2.5mm. Manual segmentation labels of the pancreas were included but not used in this study.

Chest and Abdominal CT (Testing)

- **Sliver07 [191]**: The CT volumes were collected, in the central venous phase, by different manufacturer scanners, with either 4, 16 or 64 detector rows. The in-plane resolution ranged from 0.55mm to 0.80mm, while the inter-slice distance ranged from 1mm to 3mm. Most of the patients involved in the CT collection were diagnosed of different diseases, including cancers and cysts. 20 volumes, with the manual segmentation of the liver, were used for testing.

- **CHAOS [192]**: The CT volumes were acquired from healthy potential liver donors at portal venous phase. The scans were acquired by three different manufacturer scanners (*i.e.* Philips SecuraCT, 16 detectors; Philips Mx8000 CT, 64 detectors and Toshiba AquilionOne, 320 detectors) The in-plane resolution ranged from 0.7mm to 0.8mm, while the inter-slice distance ranged from 3mm to 3.2mm. 20 volumes, with the manual segmentation of the liver, were used for testing.

- **Decath-Spleen [194]**: The CT volumes were acquired from patients receiving chemotherapy treatment for liver cancers at Memorial Sloan Kettering Cancer Center. They were scanned at the portal venous phase, with the slice thickness ranged from 2.5mm to 5mm. 41 volumes, with the manually-adjusted segmentation of the spleen, were used for testing.

- **Decath-Liver [194]**: The CT volumes were acquired from patients diagnosed of liver-related cancers, including hepatocellular carcinoma and metastatic liver disease, at various clinical centers, including Ludwig Maximilian University of Munich, Radboud University Medical Center of Nijmegen, Polytechnique and CHUM Research Center Montreal, Tel Aviv University, Sheba Medical Center, IRCAD Institute Strasbourg, and Hebrew University of Jerusalem, for the LiTS challenge [200]. The scans were collected either before or after therapy and may contain metal artifacts. The in-plane resolution ranged from 0.5mm to 1mm, while the slice thickness ranged from 0.45mm to 6mm. 131 volumes, with the manual segmentation of the liver, were used for testing.

- **Decath-Pancreas** [**194**]: The CT volumes were acquired from patients receiving resection of pancreatic masses at the Memorial Sloan Kettering Cancer Center. The slice thickness ranged from 2.5mm to 5mm. 281 volumes, with the manual segmentation of the pancreas, were used for testing.

- **3Dircadb-01** [**193**]: The CT volumes were acquired from 10 male and 10 female subjects with liver cancer. The in-plane resolution ranged from 0.56mm to 0.87mm, while the inter-slice distance ranged from 1mm to 4mm. Multiple manual annotations, including spleen, heart, gall-bladder, kidney, surrenal-gland, liver, lung and pancreas, were used for testing.

- **3Dircadb-02** [**193**]: 2 CT volumes were acquired from a patient with hepatic focal nodular hyperplasia. One was acquired at the arterial phase during inhalation, while the other was acquired at the portal phase during exhalation. The in-plane resolution is 0.961mm, while the inter-slice distance ranged from 1.8mm to 2.4mm. Together with 3Dircadb-01, manual annotations of 7 spleens, 3 hearts, 8 gall-bladders, 17 kidneys, 11 surrenal-glands, 22 liver, 12 lungs and 4 pancreases, were used for testing.

Cardiac MRI (Training)

500 video sequence, each contained approximately 30 2D images across the cardiac cycle, were collected for **Kaggle** [**198**] challenge. 14370 images, in total, were used for the training. Only images, but not any manual annotations, were used in this study.

Cardiac MRI (Testing)

The MRI volumes (**Decath-Heart** [**194**]) were collected by King's College London for the LASC Challenge [201]. The scans were acquired by a a 1.5T Philips Healthcare Achieva scanner, covering the entire heart. The in-plane resolution was 1.25mm, while the inter-slice distance was 2.7mm. 30 volumes, with the manual

segmentation of the left atrium, were used for testing.

6.3.2 Experimental Design

`Sli2Vol` and a set of baseline approaches were tested under different settings, as described as follows:

Semi-automatic (Propagation). `Sli2Vol` propagates a single-slice annotation across the whole volume iteratively. Other approaches, such as VoxelMorph [163] and optical flow, may also achieve similar goal by registering consecutive slices and they need the same amount of manual annotation as `Sli2Vol`. In order to test whether the proposed `Sli2Vol` may achieve better performance than those related approaches, it was first compared with Optical Flow [188, 202] and a 2D variation of VoxelMorph [163], which will be referred as VoxelMorph2D in this chapter. The implementation details of different approaches were given in Section 6.3.3. For all these approaches, one of the ± 3 slices around the slice with the largest ground-truth annotation was randomly picked as the initial mask. This simulates the process of a user sliding through the whole volume and roughly identifying the slice with the largest SOI to annotate, which is achievable in reality.

Semi-automatic (Supervised). Using single-slice annotation, one can train a 2D segmentation model (*e.g.* 2D-UNet) on the manually annotated slices and use the trained model to segment every slice in the volume to get a complete 3D segmentation. Therefore, `Sli2Vol` was compared with a 2D segmentation model, which was trained by a single slice annotation in each testing volume. For example, in *Sliver07*, the model, trained on 20 slice annotations (single slice from each volume), was tested on the same set of 20 volumes. This approach utilized the same amount of manual annotation as `Sli2Vol`, so as to investigate if a model trained on single slice annotations is sufficient to generalize to the whole volume. The same mechanism (*i.e.* randomly picking one of the ± 3 slices around the slice with the

largest ground-truth annotation) was used to obtain the single-slice annotations as mentioned in the last setting.

Fully Automatic (Same Domain). In this setting, `Sli2Vol` was compared with 3D segmentation models (*e.g.* 3D-UNet) that were trained on fully annotated 3D data. Here, it is referring to the scenario where the training and testing data come from the *same* benchmark dataset. The aim of this experiment was to test the upper-bound of segmentation prediction, where the ideal conditions were satisfied, namely complete 3D annotated training data and training and testing data were collected from the same domain.

Fully Automatic (Different Domain). Different from the last setting, `Sli2Vol` was compared with 3D segmentation models that were trained and tested on fully annotated 3D data from *different* domains. For example, the model was trained on *Sliver07* and tested on *CHAOS*. The aim was to evaluate the generalizability of fully supervised approaches and compare their performance with `Sli2Vol`.

6.3.3 Implementation Details

The details of the implementation of different approaches are summarized as follows and in Table 6.2.

Semi-automatic (Propagation)

- `Sli2Vol`: ResNet18 [66], without max pooling, was used as the backbone encoder. The stride at every layer was set to be equal to 1 to ensure that the dimensional size did not change. There were 16 channels at the first level. Other hyperparameters were set as follows: $d = 8$, $s = 3$, $\Omega = 15 \times 15$. Optimization was achieved using the ADAM algorithm [68] with mini-batches of size 10. The initial learning rate was set to 10^{-4}, which was decreased by half when errors plateaued. The input image dimension is 256×256.

Modality	Abdominal and Chest CT										Cardiac MRI	
Task	**Training**			**Testing**							**Training**	**Testing**
Name	C4KC-KiTS	CT-LN	CT-Pancreas	Sliver07	CHAOS	Decath-Spleen	Decath-Liver	Decath-Pancreas	3Dircadb-01	3Dircadb-02	Kaggle	Decath-Heart
Type	3D Volumes										2D Video Sequence	3D Volume
SOI	-	-	-	Liver	Liver	Spleen	Liver	Pancreas	Multiple	Multiple	-	Left atrium
Number	210	86	82	20	20	41	131	281	20	2	14370	20
Scanner	Multiple	Multiple	Philips & Siemens MDCT	Multiple	Philips Secura Philips Mx8000 Toshiba AquilionOne	NA	Multiple	NA	NA	NA	NA	Philips 1.5T Achieva
Resolution (xy) (mm)	Varying	Varying	Varying	0.55-0.8	0.7-0.8	Varying	0.5-1.0	Varying	Varying	Varying	Varying	1.25
Resolution (z) (mm)	Varying	Varying	1.5-2.5	1.0-3.0	3.0-3.2	2.5-5.0	0.45-6.0	2.5	Varying	Varying	-	2.7
Details	[195]	[196]	[197]	[191]	[192]	[194]	[194]	[194]	[193]	[193]	[198]	[194]

Table 6.1: Summarization of different datasets used for the experiments.

Approaches	**Fully Supervised - Same Domain**		**Fully Supervised - Different Domain**	**Fully Supervised - Single Slice**	**Optical Flow**	**VoxelMorph2D - UNet**	**VoxelMorph2D - ResNet18Stride1**	`Sli2Vol`
Backbone Architecture	Varying	- 3D Unet - 16 filters at first level		- 2D Unet - 64 filters at first level	conventional off-the-shelf optical flow algorithm [188]	- 2D Unet - 64 filters at first level	- ResNet18 without max pooling and stride at every layer equals 1 - 16 filters at first level	
Training hyper-parameter	Varying	- Batch size of 1 - Learning rate (lr) of 0.0001 - lr halved when errors plateaued - ADAM optimization		- Batch size of 10	-	- Batch size of 10 - Learning rate (lr) of 0.0001 - lr halved every epoch - ADAM optimization		
Input dim.	Varying	(128, 128, 128)		(256, 256)				
Remarks	Results from [192, 203–205]	Results from model trained by ourselves	-	-	Hyperparameters from OpenCV [202]: pyr_scale = 0.5 level = 3 winsize = 7	-	-	Other hyper-parameters: d = 8 s = 3 Ω = 15x15

Table 6.2: Implementation details of `Sli2Vol` and other baseline approaches.

- **Optical Flow**: Off-the-shelf optical algorithm [188] was implemented using OpenCV [202]. 3 levels of pyramid layers were used, where the height and width of each subsequent layer is halved. The window size is 7. The input image dimension is 256×256.
- **VoxelMorph2D - UNet**: A UNet backbone as proposed originally in [163] was used. There were 64 channels at the first level. Optimization was achieved using the ADAM algorithm [68] with mini-batches of size 10. The initial learning rate was set to 10^{-4}, which was decreased by half when errors plateaued. The input image dimension is 256×256.
- **VoxelMorph2D - ResNet18Stride1**: The same backbone as `Sli2Vol` was used for the VoxelMorph2D. The aim was to ensure that any potential difference of performance between `Sli2Vol` and VoxelMorph2D is not due to the difference of backbone architecture. Other settings were the same as those of **VoxelMorph2D - UNet**.

Semi-automatic (Supervised). 2D UNets, with 64 channels at the first level, were used for this setting. The models were referred as **Fully Supervised - Single Slice**. Optimization was achieved using the ADAM algorithm [68] with mini-batches of size 10. The initial learning rate was set to 10^{-4}, which was decreased by half when errors plateaued. The input image dimension is 256×256.

Fully Automatic (Same Domain). Results from both state-of-the-art methods [192, 203–205] and 3D UNets, with 16 channels at the first level, trained in this study were reported, such that their difference can be compared. The models in this setting were referred as **Fully Supervised - Same Domain**. For the trained UNets, optimization was achieved using the ADAM algorithm [68] with mini-batches of size 10. The initial learning rate was set to 10^{-4}, which was decreased by half when errors plateaued. The input volume dimension is $128 \times 128 \times 128$.

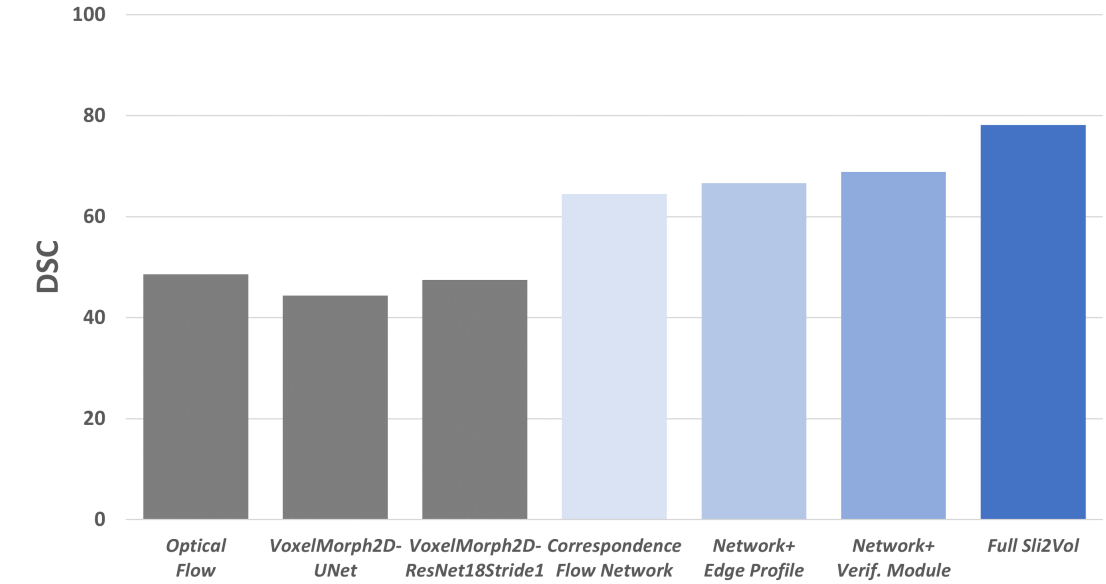
Fully Automatic (Different Domain). 3D UNets, with 16 channels at the first level, were used for this setting. The models were referred as **Fully Supervised - Different Domain**. Optimization was achieved using the ADAM algorithm [68] with mini-batches of size 10. The initial learning rate was set to 10^{-4}, which was decreased by half when errors plateaued. The input volume dimension is $128 \times 128 \times 128$.

6.4 Results

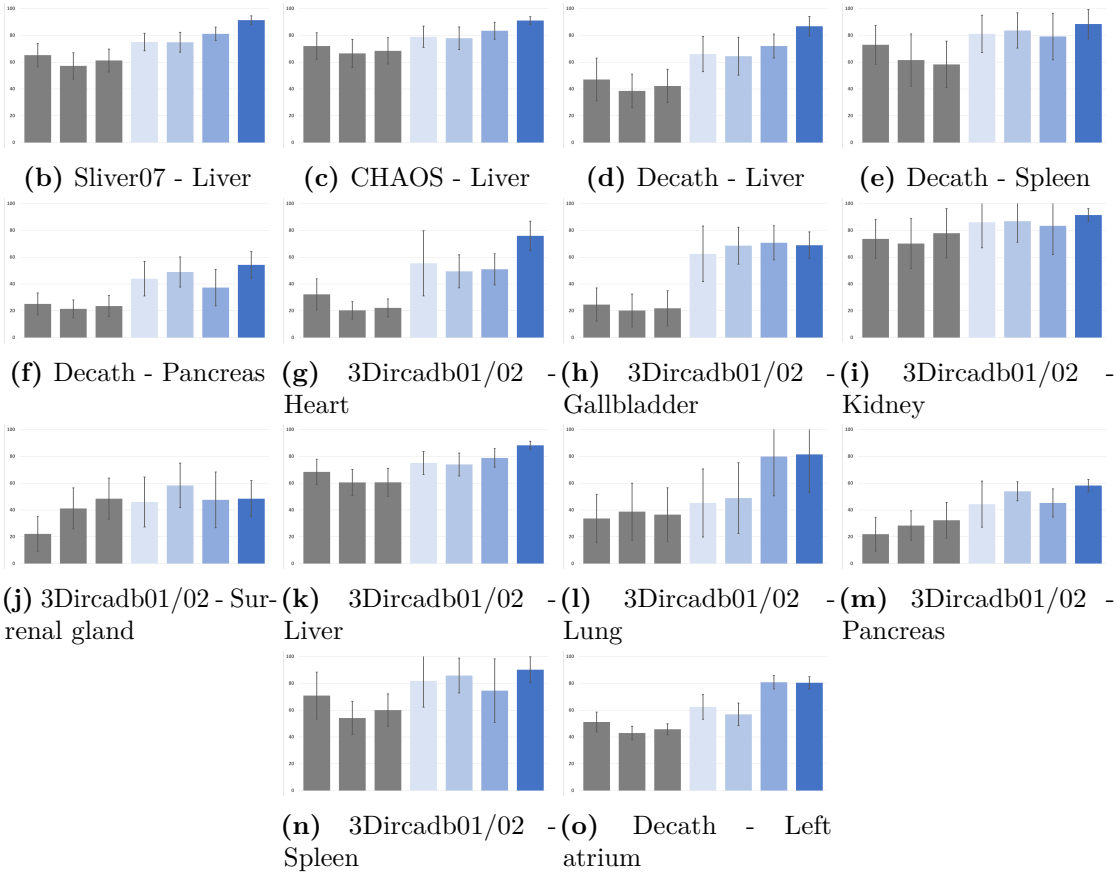
In this section, the results of the semi-automatic settings are first presented in Section 6.4.1 and Figs. 6.3 and 6.4. After that, the results of the automatic settings are presented in Section 6.4.2 and Figs. 6.5 and 6.6. Some examples of the qualitative segmentation results generated by `Sli2Vol` are displayed in Fig. 6.7. Finally, an analysis on `Sli2Vol` is presented in Section 6.4.3.

6.4.1 Semi-Automatic Approaches

`Sli2Vol` was first compared with all other semi-automatic approaches that rely on single-slice annotation. The results of propagation-based approaches, namely **Optical Flow** and **VoxelMorph2D**, are presented in Fig. 6.3. As shown in Fig. 6.3a, which shows the overall results over all the datasets, higher DSC of **Correspondence Flow Network** (leftmost blue bars) over all the gray bars suggested that solely self-supervised correspondence matching may incur less severe error drift and, hence, be more suitable than all the baseline methods for mask propagation within a volume. Comparison of results of **VoxelMorph2D - UNet**, **VoxelMorph2D - ResNet18Stride1** and **Correspondence Flow Network** further verified that the backbone architecture was not the determining factor for the superior performance achieved by `Sli2Vol`. The addition of the proposed edge profile and verification module contribute to the **Full `Sli2Vol`** (rightmost blue bars), which showed even better performance. Fig. 6.3b to Fig. 6.3o, which shows



(a) Overall



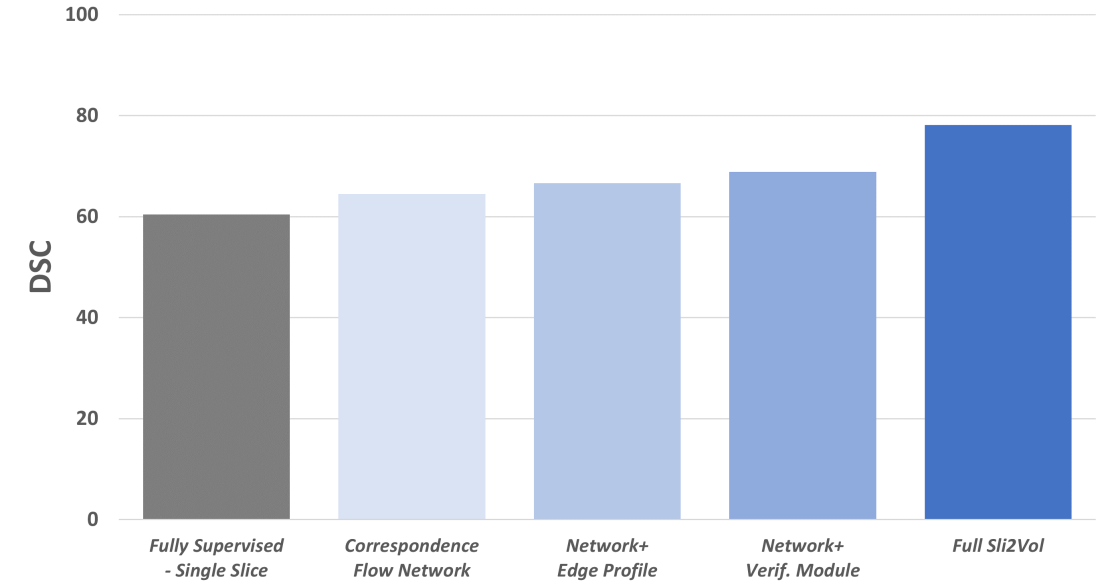
(b) Sliver07 - Liver **(c)** CHAOS - Liver **(d)** Decath - Liver **(e)** Decath - Spleen

(f) Decath - Pancreas **(g)** 3Dircadb01/02 - Heart **(h)** 3Dircadb01/02 - Gallbladder **(i)** 3Dircadb01/02 - Kidney

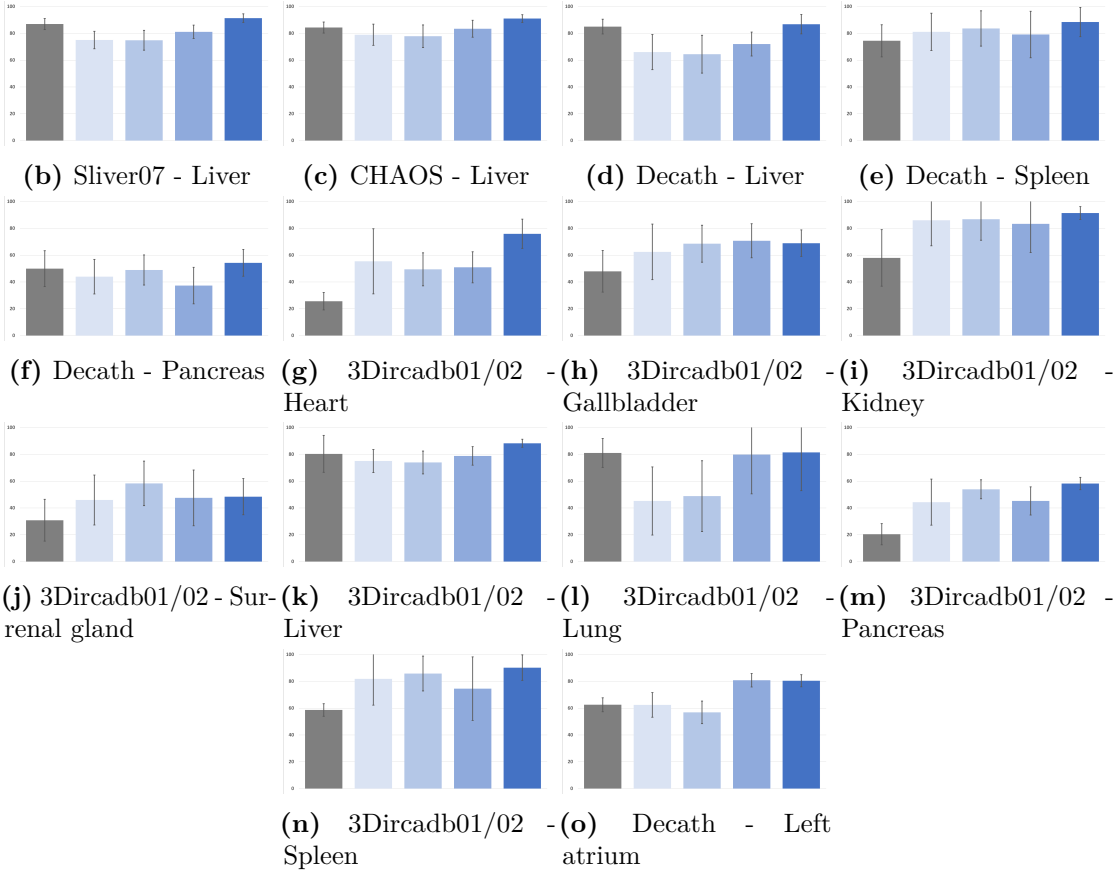
(j) 3Dircadb01/02 - Sur-renal gland **(k)** 3Dircadb01/02 - Liver **(l)** 3Dircadb01/02 - Lung **(m)** 3Dircadb01/02 - Pancreas

(n) 3Dircadb01/02 - Spleen **(o)** Decath - Left atrium

Figure 6.3: Quantitative segmentation results of semi-automatic propagation-based methods. The gray bars represent the baseline methods and the blue bars represent `Sli2Vol` with ablation studies. 6.3a is the overall results of all the datasets, while 6.3b to 6.3o correspond to each individual dataset.



(a) Overall



(b) Sliver07 - Liver **(c)** CHAOS - Liver **(d)** Decath - Liver **(e)** Decath - Spleen

(f) Decath - Pancreas **(g)** 3Dircadb01/02 - Heart **(h)** 3Dircadb01/02 - Gallbladder **(i)** 3Dircadb01/02 - Kidney

(j) 3Dircadb01/02 - Surrenal gland **(k)** 3Dircadb01/02 - Liver **(l)** 3Dircadb01/02 - Lung **(m)** 3Dircadb01/02 - Pancreas

(n) 3Dircadb01/02 - Spleen **(o)** Decath - Left atrium

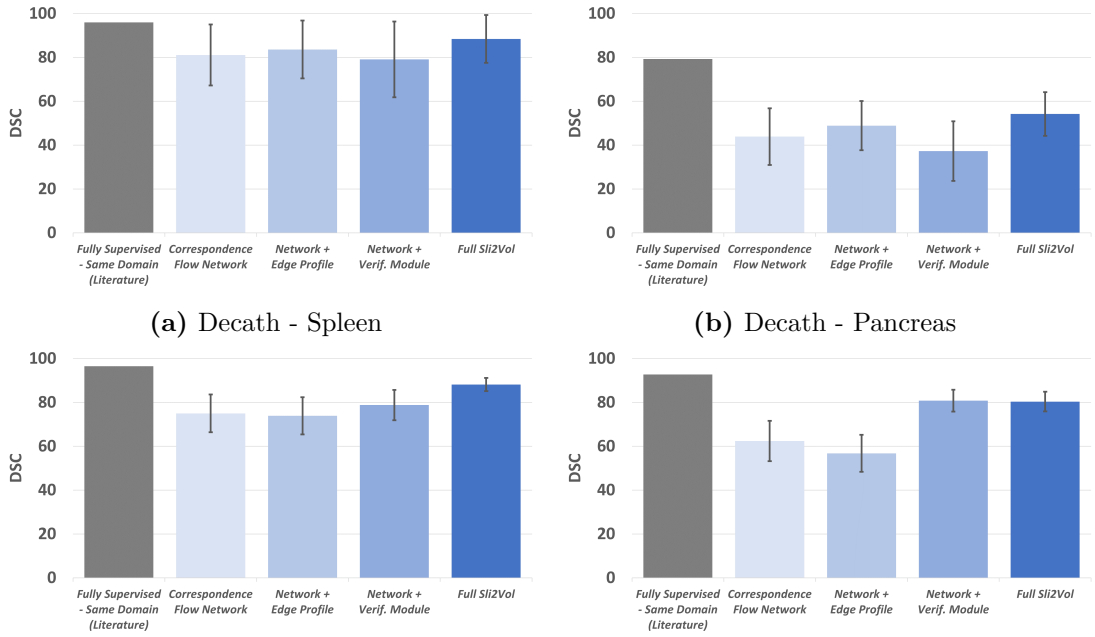
Figure 6.4: Quantitative segmentation results of `Sli2Vol` and semi-automatic supervised methods. The gray bars represent the baseline methods and the blue bars represent `Sli2Vol` with ablation studies. 6.4a is the overall results of all the datasets, while 6.4b to 6.4o correspond to each individual dataset.

the performance of different approaches on individual dataset, manifested similar trend of performance. This may further verify the robustness of `Sli2Vol` over **Optical Flow** and **VoxelMorph2D** for mask propagation within a volume.

With the same amount of annotation, `Sli2Vol` was also compared with supervised semi-automatic approach, which is shown in Fig. 6.4. A 2D-UNet, trained by a single-slice annotation in each testing volume, was tested on the same volumes. As shown in Fig. 6.4a, which shows the overall results over all the datasets, `Sli2Vol` with the ablation studies (all the blue bars) clearly outperformed the supervised semi-automatic approach (*i.e.* **Fully Supervised - Different Domain** represented by the gray bars) on all benchmarks (Fig. 6.4b to Fig. 6.4o), with an average DSC margin of over 18. Furthermore, for supervised semi-automatic approach, a new network needs to be trained for every new dataset, which is time and resources demanding. On the other hand, only a single `Sli2Vol` model is needed for all different datasets, which is much more efficient and practical.

6.4.2 Automatic Approaches

`Sli2Vol` was also compared with 3D segmentation models that were trained on fully annotated 3D data. As shown in Fig. 6.5, when the training and testing data belonged to the same domain, the best performance was achieved by the fully supervised approaches (gray bars). This was not surprising as fully annotated 3D data were used for training the models. However, as shown in Fig. 6.6, a significant performance drop (*i.e.* over 20 DSC) can be observed for cross-domain (*i.e.* same SOI, different datasets) evaluation, for example **Fully Supervised - Different Domain** was trained on *Sliver07* and tested on *CHAOS*. This was manifested by comparing the two gray bars on the left and in the middle, to the rightmost lightly gray bars in in Fig. 6.6. The leftmost gray bars report the results from the literature [192, 203–205], while the middle ones report the results obtained by the 3D UNet trained in this study. Both results were reported for a fair comparison with **Fully Supervised - Different Domain** (rightmost lightly gray bars).



(a) Decath - Spleen **(b)** Decath - Pancreas

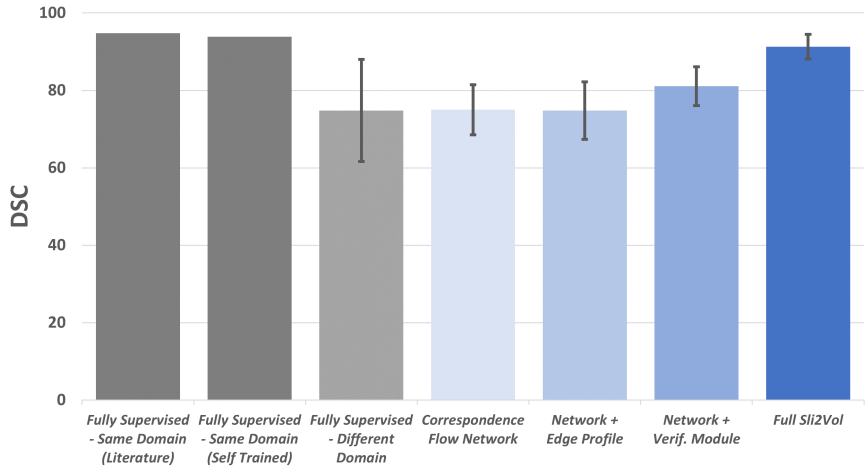
(c) 3Dircadb01/02 - Liver **(d)** Decath - Left Atrium

Figure 6.5: Quantitative segmentation results of `Sli2Vol` and fully automatic methods tested on the same-domain data. The gray bars represent the baseline methods and the blue bars represent `Sli2Vol` with ablation studies.

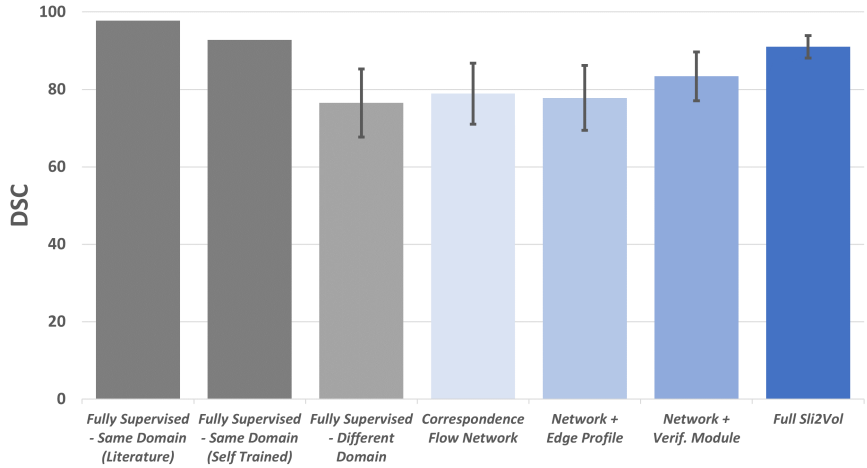
Such decline of performance encountered by the fully automatic approaches may be partially minimized by increasing the amount and diversity of training data, better design of training augmentation, and application of domain adaptation techniques. However, these may not always be practical in real-world scenarios, due to the high cost of data annotation and frequent domain shifts, for example variation of scanners and acquisition protocols in different clinical sites. Under this scenario, `Sli2Vol` outperformed the fully supervised approaches significantly ($p < 0.05$, t-test), by more than 20 DSC, as shown in Fig. 6.6, and the annotation efforts are much lower, namely only a single slice per volume.

6.4.3 Analysis on `Sli2Vol`

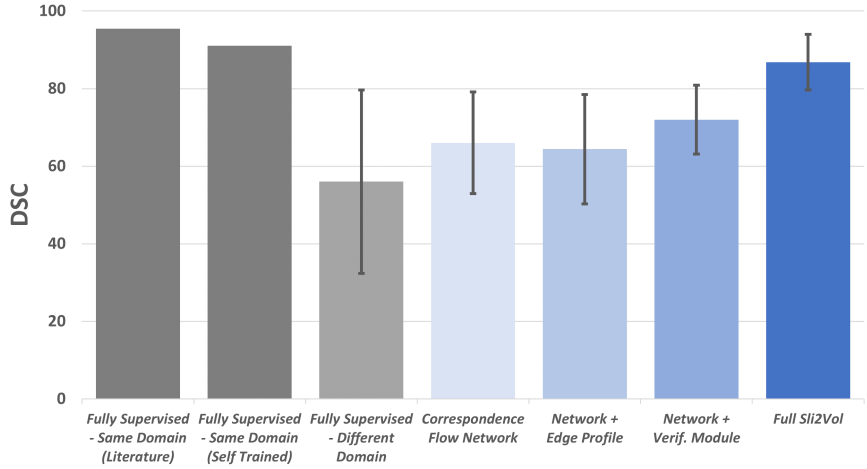
As shown in Fig. 6.3, `Sli2Vol` trained with self-supervised learning is agnostic to SOIs and domains. As for abdominal and chest CT (Fig. 6.3b to Fig. 6.3n), a *single* `Sli2Vol` model without any fine-tuning achieves a mean DSC of 78.0



(a) Sliver07 - Liver



(b) CHAOS - Liver



(c) Decath - Liver

Figure 6.6: Quantitative segmentation results of `Sli2Vol` and fully automatic methods tested on the different-domain data. The gray bars represent the baseline methods and the blue bars represent `Sli2Vol` with ablation studies.

when testing on 7 datasets spanning 8 anatomical structures. As for the cardiac MRI experiments with large training-testing domain shift (Fig. 6.3o), `Sli2Vol` still performs reasonably well with a DSC of 80.4. For the ablation studies, the proposed edge profile (**Network + Edge Profile**) was shown to be a more effective bottleneck than using the original slice as input and it further boosted the marginal benefit of the verification module, which was manifested by the last rightmost blue bars (**Full `Sli2Vol`**). Some examples of segmentation result generated by `Sli2Vol` are shown in Fig. 6.7

6.5 Conclusion

This chapter investigates on semi-automatic 3D segmentations, where any *arbitrary* SOIs in 3D medical volumes are segmented while only requiring manual annotation of a single slice. The proposed architecture, `Sli2Vol`, is trained with self-supervised learning to output affinity matrices between consecutive slices through correspondence matching, which are then used to propagate the segmentation through the volume during inference. The proposed edge profile generator and verification module further improve `Sli2Vol`'s performance and robustness.

Benchmarking on 8 public CT and MRI datasets with 9 different SOIs, `Sli2Vol` shows superior generalizability and accuracy as compared to other baseline approaches, agnostic to the SOI. Although only being tested on CT and MRI volumes, with its generalizability demonstrated, `Sli2Vol` shows great potential to be utilized on ultrasound volumes in the future works, which may ultimately couple with `ImplicitVol` (Chapter 5) to extract structural volumetric biometrics from fetal ultrasound brain volumes reconstructed by 2D freehand images. Furthermore, the goal is to provide end users with more flexibility to segment and analyze different SOIs with `Sli2Vol`, to facilitate the community to study various anatomical structures, and minimize the cost of annotating large dataset.

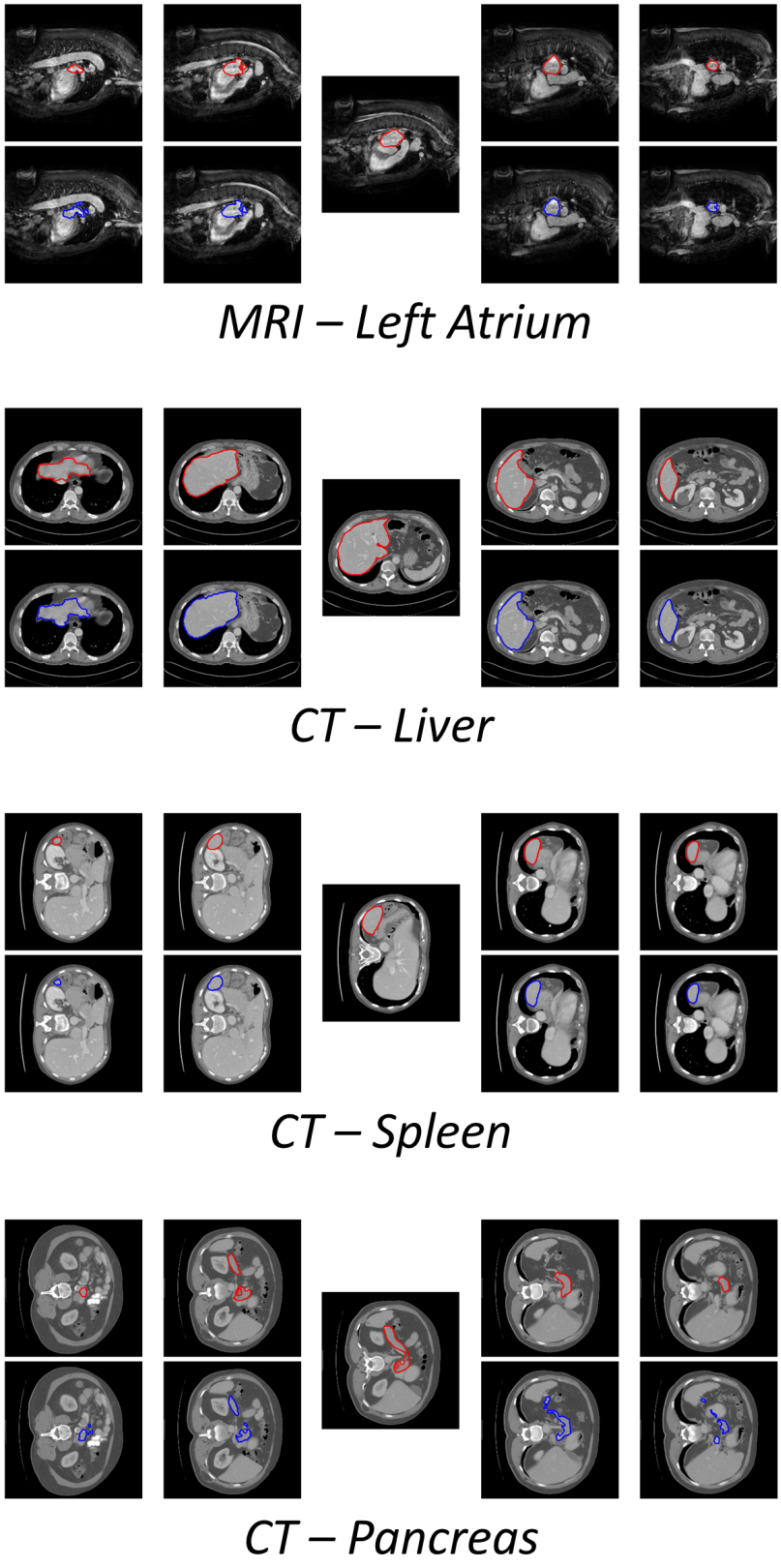


Figure 6.7: Examples of segmentation results generated by `Sli2Vol`. The middle slice is the initial annotation. Red contours represent ground-truth segmentation while blue contours represent segmentation generated by `Sli2Vol`.

7 Conclusion

This thesis has presented a framework for optimizing the utilization and diagnostic power of 2D freehand ultrasound in fetal brain imaging. Four respective components, namely plane localization (Chapter 3), unsupervised domain adaptation (Chapter 4), slice-to-volume reconstruction (Chapter 5) and semi-automatic segmentation (Chapter 6), have been proposed. In this final chapter, the major contributions of this thesis are first summarized in Section 7.1. The limitations of the proposed framework and the potential future works are then discussed in Section 7.2.

Contents

7.1 Contributions

This thesis presents an end-to-end framework to assist the acquisition of 2D ultrasound fetal brain images and the extraction of volumetric and structural information from them. Several contributions have been made accordingly, where three fundamental principles are followed when proposing the framework, namely minimal human annotation, generalizability and sensorless operation, to maximize its impact and utilization in practice.

7.1.1 Localizing 2D Ultrasound Images in a 3D Brain Atlas

The first contribution of this thesis is to predict the location of 2D ultrasound fetal brain scans in a 3D brain atlas. The task is formulated as a self-supervised learning problem, where the proposed `PlaneInVol` is trained by sampling 2D slices from aligned 3D fetal brain ultrasound volumes. In order to optimize generalizability, `AdLocUI` is further proposed to adapt a trained localization model (*e.g.* `PlaneInVol`) to freehand 2D ultrasound images acquired from arbitrary machines and with different acquisition protocols. The proposed domain adaptation mechanism is *unsupervised*, ensuring that no extra manual annotation is needed. In overall, the framework presented is sensor-free, which is trained with minimal human annotations and can be adapted to images from different domains. This may be useful for different potential clinical applications: (i) training novices because the model may help them visualize the correspondence between 2D scans and 3D atlas and structures, (ii) mapping 2D ultrasound images to 3D space may facilitate a variety of tasks, such as quality control and guiding the scanning by human-computer interaction.

7.1.2 Reconstructing Brain Volumes from 2D Images

The other contribution of this thesis is to reconstruct 3D fetal brain volumes from non-sensor-tracked 2D ultrasound images. As the first study that relies on *implicit representations* for the ultrasound volumetric reconstruction task, the proposed `ImplicitVol` can jointly refine the images' locations and learn the volumetric reconstruction, which demonstrates superior performance than conventional 3D reconstruction pipelines that have been extensively built on *explicit* representations. The proposed method could be useful particularly in settings where only 2D transducers are available but 3D assessment would be beneficial for detailed assessment or diagnostic confirmation. Without requiring extra equipment or substantial changes to the routine scanning procedures, `ImplicitVol` may enhance the diagnostic power of 2D obstetric ultrasound, by extracting volumetric information from standard 2D videos. This may facilitate the transformation of ultrasound from a screening to a powerful diagnostic tool, eventually offering personalised and advanced monitoring to the most vulnerable members of society, while capitalizing on the affordability and ubiquity of 2D ultrasound imaging at the bedside.

7.1.3 Semi-Automatic Segmentation from a Single Slice

The last contribution of this thesis is to segment any *arbitrary* structure of interest (SOI) in 3D volumes while only requiring manual annotation of a *single* slice. The proposed `Sli2Vol` requires only raw volumes, but not any manual annotation, for training. It has been demonstrated that a *single* trained model can work on wide variety of datasets and anatomical structures, without any parameter-tuning. It provides end users with more flexibility to segment and analyze different SOIs, facilitates the community to study various anatomical structures, and minimizes the cost of annotating large dataset.

7.2 Limitations and Potential Future Works

The major limitations of this thesis are discussed in this section, followed by the corresponding potential future works for further improvement.

7.2.1 Fetal Brain Imaging

In this thesis, most of the components of the proposed framework were only demonstrated on fetal brain imaging. In practice, there are other applications that may be benefited by the proposed framework. For example, for 2D echocardiography, measuring different biometrics, such as ventricle wall thickness and volume, is not straightforward [206]. Some standard planes of views may need to be identified first, similar to 2D fetal brain imaging as reviewed in Chapter 2.1. 3D localization of 2D ultrasound images (*i.e.* `PlaneInVol`) may be useful. Furthermore, 3D echocardiography is becoming more popular, which may facilitate more direct and objective measurement of volumetric and geometric biometrics [207]. Volumetric reconstruction from 2D ultrasound images (*i.e.* `ImplicitVol`) can be an alternative way for obtaining the 3D volumes.

Future works. The proposed framework may not be directly applied to other applications and may need further investigation due to the unique properties of fetal brain imaging:

- Fetal brain images, when compared to ultrasound images of other anatomical structures (*e.g.* liver), present richer contextual and structural information (*e.g.* choroid plexus and cerebellum). This may help the network learn to localize the images.
- The structure of fetal brain is relatively static, when compared to scanning of the heart, which involves cardiac motions. Therefore, temporal change of the anatomical structures and artifacts caused by motion were not the focus when designing the framework.

Therefore, when investigating other applications of the proposed framework in the future works, extra design needs to be considered:

- For anatomical structures without too many distinguishable features, other information, such as temporal continuity and other regularizations, may be needed to guide the network to learn.
- If the anatomical structures are not static, multiple atlases that correspond to different deformed states or a deformable atlas may be needed. Furthermore, if the motion is regular or uniform (*e.g.* cardiac cycles), the temporal conditions can also be incorporated into the change of atlas.
- For images corrupted by different artifacts, probability-based approaches [208] may be used to quantify the uncertainty associated with the prediction and reconstruction. Generative models [209, 210] can then be used to refine the reconstructed regions with high uncertainty.

7.2.2 Computational Cost for Volumetric Reconstruction

One major limitation of `ImplicitVol` (Chapter 5) is the computational cost required. For every volume reconstructed, a MLP needs to be trained from scratch on GPUs for tens of minutes. This poses several limitations:

- Since training of neural network is required, use of at least one GPU is recommended. This is not a common hardware that most hospitals and clinics may have, let alone the healthcare facilities in LMIC. This may limit the use of the proposed framework in those settings.
- The reconstruction (*i.e.* training) may take tens of minutes, which is not optimal. Firstly, routine scanning session normally spans 20 to 30 minutes. The reconstruction time is longer than that, which may disrupt the normal practice. This contradicts to one of the goals of the proposed framework, which is to integrate seamlessly into the clinical workflow. Secondly, the long

reconstruction time also prevents sonographers or clinicians from repeated acquisition. Specifically, those medical professionals should be able to acquire new set of images if the reconstructed volumes are sub-optimal. Such feedback and correction mechanism may not be practical if the reconstruction (*i.e.* waiting) time is too long.

Although the long computational time can be partially addressed by having more and better hardware, this is not practical in many resources-limited settings and the fundamental inefficiency of the reconstruction pipeline is still not addressed properly.

Future works. One of the potential future research directions for speeding up the volumetric reconstruction is from the perspective of implicit representation. Specifically, implicit representation is widely used for novel views synthesis of natural scenes. Recent studies in that field of research demonstrated the potential of speeding up the process through different techniques, such as generalizable pre-training [211] and explicit representation modeling [212]. Examining those related techniques and designing approaches that fit to the problem setting proposed in this thesis can be a potential future research direction.

If the reconstruction can be significantly sped up, a follow-up question would be if a user-in-the-loop technique can be developed such that the reconstruction performance can be iteratively improved by acquiring more 2D images recommended by a software system. Specifically, if volumetric reconstruction can be achieved within a short period of time (*e.g.* in several minutes), a feedback system can be developed to assess the quality of the current set of 2D images and recommend to the sonographer an extra set of images to be acquired. This process is performed iteratively until the system determines that there are sufficient images to achieve a structurally-sound volumetric reconstruction.

7.2.3 Segmentation of Ultrasound Volumes

In Chapter 6, although `Sli2Vol` was tested on different datasets and modalities (*i.e.* MRI and CT) to demonstrate its generalizability on semi-automatic medical image segmentation, the testing on ultrasound volumes was missing. Due to the unique characteristics of ultrasound images and volumes, for example the difference in resolution and artifacts presented, `Sli2Vol` may have different performance when tested on ultrasound dataset. Semi-automatic segmentation of arbitrary structures is important as many useful biometrics can be derived from them, for example the surface area of cerebellar vermis can potentially be used for early diagnosis of vermian anomalies [44]. Therefore, this may be essential for strengthening the proposed framework.

Future works. One of the potential future works will be testing `Sli2Vol` on ultrasound volumes. Due to the difference in image characteristics, it may need to be further fine-tuned to achieve optimal performance. Since a complete single-slice segmentation is required for `Sli2Vol`, which is still relatively tedious, potential future research direction is to further simplify this requirement. Specifically, simpler modes of annotation, for example using several points and lines, can be used to replace full segmentation when the users annotate the SOIs. This may further facilitate the use of `Sli2Vol` at the bedside, where carefully drawing the boundary of a structure may not be practical enough.

Bibliography

[1] Dario Paladini, Gustavo Malinger, Ana Monteagudo, Gianluigi Pilu, Ilan Timor-Tritsch, and Ants Toi. "Sonographic examination of the fetal central nervous system: guidelines for performing the 'basic examination' and the 'fetal neurosonogram'". In: *Ultrasound in Obstetrics and Gynecology* 29.1 (2007), pp. 109–116.

[2] Jordana N Peake, Rachel L Knowles, Jill Shawe, Judith Rankin, and Andrew J Copp. "Maternal ethnicity and the prevalence of British pregnancies affected by neural tube defects". In: *Birth Defects Research* 113.12 (2021), pp. 968–980.

[3] Public Health England. *NHS Fetal Anomaly Screening Programme Handbook*. Guidance. Public Health England, 2018.

[4] L. J. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. L. Johnsen, K. Kalache, K.-Y. Leung, G. Malinger, H. Munoz, F. Prefumo, A. Toi, W. Lee, and on behalf of the ISUOG Clinical Standards Committee. "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan". In: *Ultrasound in Obstetrics & Gynecology* 37.1 (2011), pp. 116–126.

[5] Joseph D Seffah and Richard MK Adanu. "Obstetric ultrasonography in low-income countries". In: *Clinical Obstetrics and Gynecology* 52.2 (2009), pp. 250–255.

[6] Beryl R Benacerraf. "Three-dimensional Fetal Sonography: Use and Misuse". In: *Journal of Ultrasound in Medicine* 21.10 (2002), pp. 1063–1067.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[8] Longlong Jing and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2020), pp. 4037–4058.

[9] Stuart Campbell. "A short history of sonography in obstetrics and gynaecology". In: *Facts, Views & Vision in ObGyn* 5.3 (2013), p. 213.

[10] Robbie Kerr and Rachel Liebling. "The fetal anomaly scan". In: *Obstetrics, Gynaecology & Reproductive Medicine* 31.3 (2021), pp. 72–76.

[11] Michael Bethune, Ekaterina Alibrahim, Braidy Davies, and Eric Yong. "A pictorial guide for the second trimester ultrasound". In: *Australasian Journal of Ultrasound in Medicine* 16.3 (2013), pp. 98–113.

[12] Pam Loughna, Lyn Chitty, Tony Evans, and Trish Chudleigh. "Fetal Size and Dating: Charts Recommended for Clinical Obstetric Practice". In: *Ultrasound* 17.3 (2009), pp. 160–166.

[13] Ramireddy Harikiran Reddy, Kumar Prashanth, and Mahale Ajit. "Significance of foetal transcerebellar diameter in foetal biometry: a pilot study". In: *Journal of Clinical and Diagnostic Research: JCDR* 11.6 (2017), TC01.

[14] Olufemi Adebari Oloyede, Tessie O Shorunmu, Peter O Adefuye, and Mkpe Abbey. "Foetal Transcerebellar Diameter (TCD) measurement between 18 and 23 weeks of pregnancy". In: *Annals of Health Research* 3.1 (2017), pp. 60–65.

[15] ASM Vinkesteijn, CLR Jansen, FJ Los, PGH Mulder, and JW Wladimiroff. "Fetal transcerebellar diameter and chromosomal abnormalities". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 17.6 (2001), pp. 502–505.

[16] PL Hilpert, BE Hall, and AB Kurtz. "The atria of the fetal lateral ventricles: a sonographic study of normal atrial size and choroid plexus volume." In: *AJR. American Journal of Roentgenology* 164.3 (1995), pp. 731–734.

[17] John H Gilmore, Lauren C Smith, Honor M Wolfe, Barbara S Hertzberg, J Keith Smith, Nancy C Chescheir, Dianne D Evans, Chaeryon Kang, Robert M Hamer, Weili Lin, and Guido Gerig. "Prenatal mild ventriculomegaly predicts abnormal development of the neonatal brain". In: *Biological Psychiatry* 64.12 (2008), pp. 1069–1076.

[18] G Malinger, D Lev, D Kidron, F Heredia, R Hershkovitz, and T Lerman-Sagie. "Differential diagnosis in fetuses with absent septum pellucidum". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 25.1 (2005), pp. 42–49.

[19] P Falco, S Gabrielli, A Visentin, A Perolo, G Pilu, and L Bovicelli. "Transabdominal sonography of the cavum septum pellucidum in normal fetuses in the second and third trimesters of pregnancy". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 16.6 (2000), pp. 549–553.

[20] Thomas C Winter, Anne M Kennedy, Jan Byrne, and Paula J Woodward. "The cavum septi pellucidi: why is it important?" In: *Journal of Ultrasound in Medicine* 29.3 (2010), pp. 427–444.

[21] Roy A Filly, Jimmy D Cardoza, Ruth B Goldstein, and Anthony J Barkovich. "Detection of Fetal Central Nervous System Anomalies: A Practical Level of Effort for a Routine Sonogram". In: *Radiology* 172.2 (1989), pp. 403–408.

[22] Kelsey A Stewart, Sergio M Navarro, Sriharsha Kambala, Gail Tan, Revanth Poondla, Sara Lederman, Kelli Barbour, and Chris Lavy. "Trends in ultrasound use in low and middle income countries: a systematic review". In: *International Journal of Maternal and Child Health and AIDS* 9.1 (2020), p. 103.

[23] Ruzica Maksimovic and Adriana Velazquez Berumen. "Innovative technology in addressing global health issues: the who perspective". In: *Geneva, Switzerland: World Health Organization (WHO/HQ/HSS/EHT/DIM)* (2011).

[24] Traci B Fox. "Multiple pregnancies: Determining chorionicity and amnionicity". In: *Journal of Diagnostic Medical Sonography* 22.1 (2006), pp. 59–65.

[25] Blair J Wylie, Linda Kalilani-Phiri, Mwayi Madanitsa, Gladys Membe, Osward Nyirenda, Patricia Mawindo, Redson Kuyenda, Albert Malenga, Abbey Masonbrink, Bonus Makanani, Phillip Thesing, and Miriam K Laufer. "Gestational age assessment in malaria pregnancy cohorts: a prospective ultrasound demonstration project in Malawi". In: *Malaria Journal* 12.1 (2013), pp. 1–11.

[26] Eunsoo Timothy Kim, Kavita Singh, Allisyn Moran, Deborah Armbruster, and Naoko Kozuki. "Obstetric ultrasound use in low and middle income countries: a narrative review". In: *Reproductive Health* 15.1 (2018), pp. 1–26.

[27] Babette Müller-Rockstroh. "Appropriate and appropriated technology: Lessons learned from ultrasound in Tanzania". In: *Medical Anthropology* 31.3 (2012), pp. 196–212.

[28] Heidi Harbison Kimberly, Alice Murray, Maria Mennicke, Andrew Liteplo, Jason Lew, J Stephen Bohan, Lynda Tyer-Viola, Roy Ahn, Thomas Burke, and Vicki E Noble. "Focused maternal ultrasound by midwives in rural Zambia". In: *Ultrasound in Medicine & Biology* 36.8 (2010), pp. 1267–1272.

[29] Stephanie Sippel, Krithika Muruganandan, Adam Levine, and Sachita Shah. "Use of ultrasound in the developing world". In: *International Journal of Emergency Medicine* 4.1 (2011), pp. 1–11.

[30] Patricia C Henwood, David C Mackenzie, Joshua S Rempell, Alice F Murray, Megan M Leo, Anthony J Dean, Andrew S Liteplo, and Vicki E Noble. "A practical guide to self-sustaining point-of-care ultrasound education programs in resource-limited settings". In: *Annals of Emergency Medicine* 64.3 (2014), pp. 277–285.

[31] Cheick Oumar Bagayoko, Diakaridia Traoré, Laurence Thevoz, Soumahila Diabaté, David Pecoul, Mahamoudane Niang, Georges Bediang, Seydou Tidiane Traoré, Abdrahamane Anne, and Antoine Geissbuhler. "Medical and economic benefits of telehealth in low-and middle-income countries: results of a study in four district hospitals in Mali". In: *BMC Health Services Research* 14.1 (2014), pp. 1–6.

[32] Nittaya Chamadol, Vallop Laopaiboon, Jiraporn Srinakarin, Watcharin Loilome, Puangrat Yongvanit, Bandit Thinkhamrop, and Narong Khuntikeo. "Teleconsultation ultrasonography: a new weapon to combat cholangiocarcinoma". In: *Esmo Open* 2.3 (2017), e000231.

[33] Luis F Gonçalves. "Three-dimensional ultrasound of the fetus: how does it help?" In: *Pediatric Radiology* 46.2 (2016), pp. 177–189.

[34] Tho Quynh Nguyen and Melanie Flores. "Accuracy of ultrasound measurements by novices: Pixels or voxels". In: *Donald School Journal of Ultrasound in Obstetrics and Gynecology* 5.3 (2011), pp. 303–309.

[35] E-K Ji, Dolores H Pretorius, Ruth Newton, K Uyan, AD Hull, K Hollenbach, and TR Nelson. "Effects of ultrasound on maternal-fetal bonding: a comparison of two-and three-dimensional imaging". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 25.5 (2005), pp. 473–477.

[36] Min-Long Chen, Chiung-Hsin Chang, Chen-Hsiang Yu, Yueh-Chin Cheng, and Fong-Ming Chang. "Prenatal diagnosis of cleft palate by three-dimensional ultrasound". In: *Ultrasound in Medicine & Biology* 27.8 (2001), pp. 1017–1023.

[37] Ramen Chmait, Dolores Pretorius, Marilyn Jones, Andrew Hull, Gina James, Tom Nelson, and Thomas Moore. "Prenatal evaluation of facial clefts with two-dimensional and adjunctive three-dimensional ultrasonography: a prospective trial". In: *American Journal of Obstetrics and Gynecology* 187.4 (2002), pp. 946–949.

[38] C Mittermayer, W Blaicher, PC Brugger, G Bernaschek, and A Lee. "Foetal facial clefts: prenatal evaluation of lip and primary palate by 2D and 3D ultrasound". In: *Ultraschall in der Medizin-European Journal of Ultrasound* 25.02 (2004), pp. 120–125.

[39] E Merz and C Welter. "2D and 3D Ultrasound in the evaluation of normal and abnormal fetal anatomy in the second and third trimesters in a level III center". In: *Ultraschall in der Medizin-European Journal of Ultrasound* 26.01 (2005), pp. 9–16.

[40] H-GK Blaas, P Taipale, H Torp, and SH Eik-Nes. "Three-dimensional ultrasound volume calculations of human embryos and young fetuses: a study on the volumetry of compound structures and its reproducibility". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 27.6 (2006), pp. 640–646.

[41] LR Pistorius, Ph Stoutenbeek, F Groenendaal, L De Vries, G Manten, E Mulder, and G Visser. "Grade and symmetry of normal fetal cortical development: a longitudinal two-and three-dimensional ultrasound study". In: *Ultrasound in Obstetrics & Gynecology* 36.6 (2010), pp. 700–708.

[42] Luís F Gonçalves, Wesley Lee, Jimmy Espinoza, and Roberto Romero. "Three- and 4-Dimensional Ultrasound in Obstetric Practice: Does It Help?" In: *Journal of Ultrasound in Medicine* 24.12 (2005), pp. 1599–1624.

[43] Anna M Dückelmann and Karim D Kalache. "Three-dimensional ultrasound in evaluating the fetus". In: *Prenatal Diagnosis* 30.7 (2010), pp. 631–638.

[44] F Vinals, M Munoz, R Naveas, J Shalper, and A Giuliano. "The fetal cerebellar vermis: anatomy and biometric assessment using volume contrast imaging in the C-plane (VCI-C)". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 26.6 (2005), pp. 622–627.

[45] Ana IL Namburete, Richard V Stebbing, Bryn Kemp, Mohammad Yaqub, Aris T Papageorghiou, and Alison Noble. "Learning-based prediction of gestational age from ultrasound images of the fetal brain". In: *Medical Image Analysis* 21.1 (2015), pp. 72–86.

[46] Ana IL Namburete, Weidi Xie, and Alison Noble. "Robust regression of brain maturation from 3d fetal neurosonography using CRNs". In: *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer, 2017, pp. 73–80.

[47] Yuanwei Li, Juan J Cerrolaza, Matthew Sinclair, Benjamin Hou, Amir Alansary, Bishesh Khanal, Jacqueline Matthew, Bernhard Kainz, and Daniel Rueckert. "Standard Plane Localisation in 3D Fetal Ultrasound Using Network with Geometric and Image Loss". In: *Medical Imaging with Deep Learning (MIDL)*. 2018.

[48] Lbnmedical. "How Much Does an Ultrasound Machine Cost?" In: *https://lbnmedical.com/* (2022). URL: `https://lbnmedical.com/ultrasound-price-guide/`.

[49] Ruth Reader. "This affordable, portable ultrasound device works with a smartphone". In: *FastCompany* (2019). URL: `https://www.fastcompany.com/90327979/this-affordable-ultrasound-device-works-with-a-smartphone`.

[50] Kunihiko Fukushima and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and Cooperation in Neural Nets*. Springer, 1982, pp. 267–285.

[51] Leonid Datta. "A survey on activation functions and their relation with xavier and he normal initialization". In: *arXiv preprint arXiv:2004.06632* (2020).

[52] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[53] Abien Fred Agarap. "Deep learning using rectified linear units (ReLu)". In: *arXiv preprint arXiv:1803.08375* (2018).

[54] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (ELUs)". In: *International Conference on Learning Representations (ICLR)*. 2016.

[55] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. "Implicit neural representations with periodic activation functions". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020).

[56] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. "Micro-batch training with batch-channel normalization and weight standardization". In: *arXiv preprint arXiv:1903.10520* (2019).

[57] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International Conference on Machine Learning (ICML)*. PMLR. 2015, pp. 448–456.

[58] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. "How does batch normalization help optimization?" In: *Advances in Neural Information Processing Systems (NeurIPS)* 31 (2018).

[59] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. "Rethinking normalization and elimination singularity in neural networks". In: *arXiv preprint arXiv:1911.09738* (2019).

[60] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[61] 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". In: *International Conference on Learning Representations (ICLR)* (2020).

[62] Yuxin Wu and Kaiming He. "Group normalization". In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.

[63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.

[64] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems (NeurIPS)* 25 (2012).

[65] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations (ICLR)*. 2015.

[66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[67] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2015, pp. 234–241.

[68] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[69] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *International Conference on Learning Representations (ICLR)*. 2019.

[70] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. "Albumentations: Fast and Flexible Image Augmentations". In: *Information* 11.2 (2020). URL: https://www.mdpi.com/2078-2489/11/2/125.

[71] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *International Conference on Learning Representations (ICLR)* (2015).

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[73] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014).

[74] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM". In: *Neural Computation* 12.10 (2000), pp. 2451–2471.

[75] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. URL: https://aclanthology.org/D14-1179.

[76] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7794–7803.

[77] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. "Learn to pay attention". In: *International Conference on Learning Representations (ICLR)* (2018).

[78] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention-Gated Networks for Improving Ultrasound Scan Plane Detection". In: *Medical Imaging with Deep Learning (MIDL)*. 2018.

[79] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention gated networks: Learning to leverage salient regions in medical images". In: *Medical Image Analysis* 53 (2019), pp. 197–207.

[80] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context". In: *European Conference on Computer Vision (ECCV)*. Springer. 2014, pp. 740–755.

[81] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786 (2006), pp. 504–507.

[82] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and composing robust features with denoising autoencoders". In: *International Conference on Machine Learning (ICML)*. 2008, pp. 1096–1103.

[83] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised visual representation learning by context prediction". In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 1422–1430.

[84] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. "Learning image representations by completing damaged jigsaw puzzles". In: *Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 793–802.

[85] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations". In: *International Conference on Learning Representations (ICLR)* (2018).

[86] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *International Conference on Learning Representations (ICLR)* (2013).

[87] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 649–666.

[88] Richard Zhang, Phillip Isola, and Alexei A Efros. "Split-brain autoencoders: Unsupervised learning by cross-channel prediction". In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1058–1067.

[89] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[90] Olivier Henaff. "Data-efficient image recognition with contrastive predictive coding". In: *International Conference on Machine Learning (ICML)*. PMLR. 2020, pp. 4182–4192.

[91] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International Conference on Machine Learning (ICML)*. PMLR. 2020, pp. 1597–1607.

[92] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning". In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9729–9738.

[93] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6707–6717.

[94] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification". In: *European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 527–544.

[95] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. "Unsupervised representation learning by sorting sequences". In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 667–676.

[96] Xiaolong Wang and Abhinav Gupta. "Unsupervised learning of visual representations using videos". In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 2794–2802.

[97] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. "Learning dense correspondence via 3d-guided cycle consistency". In: *Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 117–126.

[98] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. "AutoCorrect: Deep Inductive Alignment of Noisy Geometric Annotations". In: *British Machine Vision Conference (BMVC)*. 2019.

[99] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. "Image to image translation for domain adaptation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4500–4509.

[100] Hidetoshi Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of Statistical Planning and Inference* 90.2 (2000), pp. 227–244.

[101] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation". In: *AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 865–872.

[102] Mei Wang and Weihong Deng. "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312 (2018), pp. 135–153.

[103] Kowshik Thopalli, Jayaraman J Thiagarajan, Rushil Anirudh, and Pavan Turaga. "Salt: Subspace alignment as an auxiliary learning task for domain adaptation". In: *arXiv preprint arXiv:1906.04338* (2019).

[104] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. "Domain-adversarial training of neural networks". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2096–2030.

[105] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. "Adversarial discriminative domain adaptation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7167–7176.

[106] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014).

[107] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. "Unsupervised pixel-level domain adaptation with generative adversarial networks". In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3722–3731.

[108] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. "Cycada: Cycle-consistent adversarial domain adaptation". In: *International Conference on Machine Learning (ICML)*. Pmlr. 2018, pp. 1989–1998.

[109] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 2223–2232.

[110] Fengmao Lv, Jun Zhu, Guowu Yang, and Lixin Duan. "TarGAN: Generating target data with class labels for unsupervised domain adaptation". In: *Knowledge-Based Systems* 172 (2019), pp. 123–129.

[111] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. "Conditional adversarial domain adaptation". In: *Advances in Neural Information Processing Systems (NeurIPS)* 31 (2018).

[112] Ling Zhang, Siping Chen, Chien Ting Chin, Tianfu Wang, and Shengli Li. "Intelligent scanning: Automated standard plane selection and biometric measurement of early gestational sac in routine ultrasound examination". In: *Medical Physics* 39.8 (2012), pp. 5015–5027.

[113] Dong Ni, Tianmei Li, Xin Yang, Jing Qin, Shengli Li, Chien-Ting Chin, Shuyuan Ouyang, Tianfu Wang, and Siping Chen. "Selective Search and Sequential Detection for Standard Plane Localization in Ultrasound". In: *MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*. Springer, 2013, pp. 203–211.

[114] Xin Yang, Dong Ni, Jing Qin, Shengli Li, Tianfu Wang, Siping Chen, and Pheng Ann Heng. "Standard plane localization in ultrasound by radial component". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014, pp. 1180–1183.

[115] Hao Chen, Dong Ni, Jing Qin, Shengli Li, Xin Yang, Tianfu Wang, and Pheng Ann Heng. "Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks". In: *IEEE Journal of Biomedical and Health Informatics* 19.5 (2015), pp. 1627–1636.

[116] Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. "SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound". In: *IEEE Transactions on Medical Imaging* 36.11 (2017), pp. 2204–2215.

[117] Keyu Li, Jian Wang, Yangxin Xu, Hao Qin, Dongsheng Liu, Li Liu, and Max Q-H Meng. "Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning". In: *International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 8302–8308.

[118] Haoran Dou, Xin Yang, Jikuan Qian, Wufeng Xue, Hao Qin, Xu Wang, Lequan Yu, Shujun Wang, Yi Xiong, Pheng-Ann Heng, and Dong Ni. "Agent with warm start and active termination for plane localization in 3d ultrasound". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2019, pp. 290–298.

[119] Richard Droste, Lior Drukker, Aris T Papageorghiou, and Alison Noble. "Automatic probe movement guidance for freehand obstetric ultrasound". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2020, pp. 583–592.

[120] Hao Chen, Qi Dou, Dong Ni, Jie-Zhi Cheng, Jing Qin, Shengli Li, and Pheng-Ann Heng. "Automatic Fetal Ultrasound Standard Plane Detection Using Knowledge Transferred Recurrent Neural Networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 507–514.

[121] Weilin Huang, Christopher P Bridge, Alison Noble, and Andrew Zisserman. "Temporal HeartNet: Towards Human-Level Automatic Analysis of Fetal Cardiac Screening Video". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2017, pp. 341–349.

[122] Yuan Gao and Alison Noble. "Detection and Characterization of the Fetal Heartbeat in Free-hand Ultrasound Sweeps with Weakly-supervised Two-streams Convolutional Networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2017, pp. 305–313.

[123] Hosuk Ryou, Mohammad Yaqub, Angelo Cavallaro, Fenella Roseman, Aris Papageorghiou, and Alison Noble. "Automated 3D Ultrasound Biometry Planes Extraction for First Trimester Fetal Assessment". In: *Machine Learning in Medical Imaging (MLMI)*. Springer, 2016, pp. 196–204.

[124] Amir Alansary, Loic Le Folgoc, Ghislain Vaillant, Ozan Oktay, Yuanwei Li, Wenjia Bai, Jonathan Passerat-Palmbach, Ricardo Guerrero, Konstantinos Kamnitsas, and Benjamin Hou. "Automatic View Planning with Multi-scale Deep Reinforcement Learning Agents". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018, pp. 277–285.

[125] Enzo Ferrante and Nikos Paragios. "Slice-to-volume medical image registration: A survey". In: *Medical Image Analysis* 39 (2017), pp. 101–123.

[126] Amir Alansary, Martin Rajchl, Steven G McDonagh, Maria Murgasova, Mellisa Damodaram, David FA Lloyd, Alice Davidson, Mary Rutherford, Joseph V Hajnal, and Daniel Rueckert. "PVR: Patch-to-Volume Reconstruction for Large Area Motion Correction of Fetal MRI". In: *IEEE Transactions on Medical Imaging* 36.10 (2017), p. 2031.

[127] Francois Rousseau, Orit A Glenn, Bistra Iordanova, Claudia Rodriguez-Carranza, Daniel B Vigneron, James A Barkovich, and Colin Studholme. "Registration-based approach for reconstruction of high-resolution in utero fetal MR brain images". In: *Academic Radiology* 13.9 (2006), pp. 1072–1081.

[128] Shuzhou Jiang, Hui Xue, Alan Glover, Mary Rutherford, Daniel Rueckert, and Joseph V Hajnal. "MRI of Moving Subjects Using Multislice Snapshot Images With Volume Reconstruction (SVR): Application to Fetal, Neonatal, and Adult Brain Studies". In: *IEEE Transactions on Medical Imaging* 26.7 (2007), pp. 967–980.

[129] Maria Kuklisova-Murgasova, Gerardine Quaghebeur, Mary A Rutherford, Joseph V Hajnal, and Julia A Schnabel. "Reconstruction of fetal brain MRI with intensity matching and complete outlier removal". In: *Medical Image Analysis* 16.8 (2012), pp. 1550–1564.

[130] Ali Gholipour, Judy A Estroff, and Simon K Warfield. "Robust Super-Resolution Volume Reconstruction From Slice Acquisitions: Application to Fetal Brain MRI". In: *IEEE Transactions on Medical Imaging* 29.10 (2010), pp. 1739–1758.

[131] Bernhard Kainz, Markus Steinberger, Wolfgang Wein, Maria Kuklisova-Murgasova, Christina Malamateniou, Kevin Keraudren, Thomas Torsney-Weir, Mary Rutherford, Paul Aljabar, Joseph V Hajnal, and Daniel Rueckert. "Fast volume reconstruction from motion corrupted stacks of 2D slices". In: *IEEE Transactions on Medical Imaging* 34.9 (2015), pp. 1901–1913.

[132] Tiexiang Wen, Qingsong Zhu, Wenjian Qin, Ling Li, Fan Yang, Yaoqin Xie, and Jia Gu. "An accurate and effective FMM-based approach for freehand 3D ultrasound reconstruction". In: *Biomedical Signal Processing and Control* 8.6 (2013), pp. 645–656.

[133] Athanasios Karamalis, Wolfgang Wein, Oliver Kutter, and Nassir Navab. "Fast hybrid freehand ultrasound volume reconstruction". In: *Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*. Vol. 7261. SPIE. 2009, pp. 366–373.

[134] Benjamin Hou, Amir Alansary, Steven McDonagh, Alice Davidson, Mary Rutherford, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. "Predicting slice-to-volume transformation in presence of arbitrary subject motion". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2017, pp. 296–304.

[135] Benjamin Hou, Bishesh Khanal, Amir Alansary, Steven McDonagh, Alice Davidson, Mary Rutherford, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. "3D Reconstruction in Canonical Co-ordinate Space from Arbitrarily Oriented 2D Images". In: *IEEE Transactions on Medical Imaging* 37.8 (2018), pp. 1737–1750.

[136] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. "A simple neural network module for relational reasoning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 4967–4976.

[137] Weidi Xie, Li Shen, and Andrew Zisserman. "Comparator networks". In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 782–797.

[138] Ana IL Namburete, Weidi Xie, Mohammad Yaqub, Andrew Zisserman, and Alison Noble. "Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning". In: *Medical Image Analysis* 46 (2018), pp. 1–14.

[139] Felipe Moser, Ruobing Huang, Aris T Papageorghiou, Bartłomiej W Papież, and Ana IL Namburete. "Automated fetal brain extraction from clinical Ultrasound volumes using 3D Convolutional Neural Networks". In: *Medical Image Understanding and Analysis Conference (MIUA)*. Springer. 2019, pp. 151–163.

[140] Aris T Papageorghiou, Eric O Ohuma, Douglas G Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A Jaffer, Enrico Bertino, Michael G Gravett, and Manorama Purwar. "International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project". In: *The Lancet* 384.9946 (2014), pp. 869–879.

[141] The INTERBIO-21st Consortium. *INTERBIO-21st Study Protocol*. Protocol. Oxford, 2012.

[142] Jae-Chern Yoo and Tae Hee Han. "Fast normalized cross-correlation". In: *Circuits, Systems and Signal Processing* 28.6 (2009), p. 819.

[143] Elisabetta Sassaroli, Calum Crake, Andrea Scorza, Don-Soo Kim, and Mi-Ae Park. "Image quality evaluation of ultrasound imaging systems: advanced B-modes". In: *Journal of Applied Clinical Medical Physics* 20.3 (2019), pp. 115–124.

[144] Baochen Sun and Kate Saenko. "Deep coral: Correlation alignment for deep domain adaptation". In: *European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 443–450.

[145] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. "Learning transferable features with deep adaptation networks". In: *International Conference on Machine Learning (ICML)*. PMLR. 2015, pp. 97–105.

[146] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 21271–21284.

[147] Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. "Representation Subspace Distance for Domain Adaptation Regression". In: *International Conference on Machine Learning (ICML)*. PMLR. 2021, pp. 1749–1759.

[148] Ruobing Huang, Ana Namburete, and Alison Noble. "Learning to segment key clinical anatomical structures in fetal neurosonography informed by a region-based descriptor". In: *Journal of Medical Imaging* 5.1 (2018), p. 014007.

[149] Qinghua Huang and Zhaozheng Zeng. "A review on real-time 3D ultrasound imaging technology". In: *BioMed Research International* 2017 (2017).

[150] Shao-Wen Chung, Cho-Chiang Shih, and Chih-Chung Huang. "Freehand three-dimensional ultrasound imaging of carotid artery using motion tracking technology". In: *Ultrasonics* 74 (2017), pp. 11–20.

[151] Mohammad I Daoud, Abdel-Latif Alshalalfah, Falah Awwad, and Mahasen Al-Najar. "Freehand 3D ultrasound imaging system using electromagnetic tracking". In: *International Conference on Open Source Software Computing (OSSCOM)*. IEEE. 2015, pp. 1–5.

[152] Houssem-Eddine Gueziri, Sebastien Tremblay, Catherine Laporte, and Rupert Brooks. "Graph-Based 3D-Ultrasound Reconstruction of the Liver in the Presence of Respiratory Motion". In: *Reconstruction, Segmentation, and Analysis of Medical Images*. Ed. by Maria A. Zuluaga, Kanwal Bhatia, Bernhard Kainz, Mehdi H. Moghari, and Danielle F. Pace. Cham: Springer International Publishing, 2017, pp. 48–57.

[153] Aaron Fenster, Grace Parraga, and Jeff Bax. "Three-dimensional ultrasound scanning". In: *Interface Focus* 1.4 (2011), pp. 503–519.

[154] AM Goldsmith, PC Pedersen, and TL Szabo. "An inertial-optical tracking system for portable, quantitative, 3D ultrasound". In: *International Ultrasonics SymposiumV (IUS)*. IEEE. 2008, pp. 45–49.

[155] Farhan Mohamed and C Vei Siang. "A survey on 3D ultrasound reconstruction techniques". In: *Artificial Intelligence: Applications in Medicine and Biology* (2019).

[156] Richard W Prager, Andrew H Gee, Graham M Treece, Charlotte JC Cash, and Laurence H Berman. "Sensorless freehand 3-D ultrasound using regression of the echo intensity". In: *Ultrasound in Medicine & Biology* 29.3 (2003), pp. 437–446.

[157] Raphael Prevost, Mehrdad Salehi, Julian Sprung, Alexander Ladikos, Robert Bauer, and Wolfgang Wein. "Deep learning for sensorless 3D freehand ultrasound imaging". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2017, pp. 628–636.

[158] Xiankang Chen, Tiexiang Wen, Xingmin Li, Wenjian Qin, Donglai Lan, Weizhou Pan, and Jia Gu. "Reconstruction of freehand 3D ultrasound based on kernel regression". In: *Biomedical Engineering Online* 13.1 (2014), pp. 1–15.

[159] Hyungil Moon, Geonhwan Ju, Seyoun Park, and Hayong Shin. "3D freehand ultrasound reconstruction using a piecewise smooth Markov random field". In: *Computer Vision and Image Understanding* 151 (2016), pp. 101–113.

[160] Pak-Hei Yeung, Ana IL Namburete, and Weidi Xie. "Sli2Vol: Annotate a 3D Volume from a Single Slice with Self-Supervised Learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2021, pp. 69–79.

[161] Madeleine K Wyburd, Nicola K Dinsdale, Ana IL Namburete, and Mark Jenkinson. "TEDS-Net: Enforcing Diffeomorphisms in Spatial Transformers to Guarantee Topology Preservation in Segmentations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2021, pp. 250–260.

[162] Linde S Hesse, Moska Aliasi, Felipe Moser, the INTERGROWTH-21st Consortium, Monique C Haak, Weidi Xie, Mark Jenkinson, and Ana IL Namburete. "Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning". In: *NeuroImage* 254 (2022), p. 119117.

[163] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. "Voxelmorph: a learning framework for deformable medical image registration". In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pp. 1788–1800.

[164] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. "Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[165] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 405–421.

[166] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. "NeRF−−: Neural Radiance Fields Without Known Camera Parameters". In: *arXiv preprint arXiv:2102.07064* (2021).

[167] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. "BARF: Bundle-Adjusting Neural Radiance Fields". In: *International Conference on Computer Vision (ICCV)*. 2021.

[168] Qing Wu, Yuwei Li, Lan Xu, Ruiming Feng, Hongjiang Wei, Qing Yang, Boliang Yu, Xiaozhao Liu, Jingyi Yu, and Yuyao Zhang. "IREM: High-Resolution Magnetic Resonance Image Reconstruction via Implicit Neural Representation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2021, pp. 65–74.

[169] Liyue Shen, John Pauly, and Lei Xing. "NeRP: Implicit Neural Representation Learning With Prior Embedding for Sparsely Sampled Image Reconstruction". In: *IEEE Transactions on Neural Networks and Learning Systems* (2022), pp. 1–13.

[170] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. "On the spectral bias of neural networks". In: *International Conference on Machine Learning (ICML)*. PMLR. 2019, pp. 5301–5310.

[171] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[172] Maria Kuklisova-Murgasova, Gerardine Quaghebeur, Mary A Rutherford, Joseph V Hajnal, and Julia A Schnabel. "Reconstruction of fetal brain MRI with intensity matching and complete outlier removal". In: *Medical Image Analysis* 16.8 (2012), pp. 1550–1564.

[173] Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images". In: *IEEE Transactions on Medical Imaging* 39.4 (2019), pp. 1064–1072.

[174] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. "An information fidelity criterion for image quality assessment using natural scene statistics". In: *IEEE Transactions on Image Processing* 14.12 (2005), pp. 2117–2128.

[175] Felipe Moser, Ruobing Huang, INTERGROWTH-21st Consortium, Bartłomiej W Papież, and Ana IL Namburete. "BEAN: Brain Extraction and Alignment Network for 3D Fetal Neurosonography". In: *NeuroImage* (2022), p. 119341. URL: `https://www.sciencedirect.com/science/article/pii/S1053811922004608`.

[176] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 694–711.

[177] Seyed Ali Amirshahi, Marius Pedersen, and Stella X Yu. "Image quality assessment by comparing CNN features between images". In: *Journal of Imaging Science and Technology* 60.6 (2016), pp. 60410–1.

[178] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. "On the use of deep learning for blind image quality assessment". In: *Signal, Image and Video Processing* 12.2 (2018), pp. 355–362.

[179] Zhou Zheng, Xuechang Zhang, Huafei Xu, Wang Liang, Siming Zheng, and Yueding Shi. "A unified level set framework combining hybrid algorithms for liver and liver tumor segmentation in CT images". In: *BioMed Research International* 2018 (2018).

[180] Amir Hossein Foruzan and Yen-Wei Chen. "Improved segmentation of low-contrast lesions using sigmoid edge model". In: *International Journal of Computer Assisted Radiology and Surgery* 11.7 (2016), pp. 1267–1283.

[181] Changyang Li, Xiuying Wang, Stefan Eberl, Michael Fulham, Yong Yin, Jinhu Chen, and David Dagan Feng. "A likelihood and local constraint level set model for liver tumor segmentation from CT volumes". In: *IEEE Transactions on Biomedical Engineering* 60.10 (2013), pp. 2967–2977.

[182] Benoit M Dawant, Rui Li, Brian Lennon, and Senhu Li. "Semi-automatic segmentation of the liver and its evaluation on the MICCAI 2007 grand challenge data set". In: *3D Segmentation in The Clinic: A Grand Challenge* (2007), pp. 215–221.

[183] Guotai Wang, Maria A Zuluaga, Rosalind Pratt, Michael Aertsen, Anna L David, Jan Deprest, Tom Vercauteren, and Sebastien Ourselin. "Slic-Seg: slice-by-slice segmentation propagation of the placenta in fetal MRI using one-plane scribbles and online learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2015, pp. 29–37.

[184] Simon Hermann and René Werner. "High accuracy optical flow for 3D medical image registration using the census cost function". In: *Pacific-Rim Symposium on Image and Video Technology*. Springer. 2013, pp. 23–35.

[185] Stephen L Keeling and Wolfgang Ring. "Medical image registration and interpolation by optical flow with maximal rigidity". In: *Journal of Mathematical Imaging and Vision* 23.1 (2005), pp. 47–65.

[186] Sergiu Mocanu, Alan R. Moody, and April Khademi. "FlowReg: Fast Deformable Unsupervised Medical Image Registration using Optical Flow". In: *Machine Learning for Biomedical Imaging* 1 (September 2021 issue 2021). URL: https://melba-journal.org/papers/2021:015.html.

[187] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. "MRF-based deformable registration and ventilation estimation of lung CT". In: *IEEE Transactions on Medical Imaging* 32.7 (2013), pp. 1239–1248.

[188] Gunnar Farnebäck. "Two-frame motion estimation based on polynomial expansion". In: *Scandinavian Conference on Image Analysis (SCIA)*. Springer. 2003, pp. 363–370.

[189] Zihang Lai and Weidi Xie. "Self-supervised Learning for Video Correspondence Flow". In: *British Machine Vision Conference (BMVC)*. 2019.

[190] Zihang Lai, Erika Lu, and Weidi Xie. "MAST: A Memory-Augmented Self-Supervised Tracker". In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6479–6488.

[191] Bram Van Ginneken, Tobias Heimann, and Martin Styner. "3D segmentation in the clinic: A grand challenge". In: *MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge*. Vol. 1. 2007, pp. 7–15.

[192] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation". In: *Medical Image Analysis* 69 (2021), p. 101950.

[193] L Soler, A Hostettler, V Agnus, A Charnoz, J Fasquel, J Moreau, A Osswald, M Bouhadjar, and J Marescaux. "3D image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database". In: *IRCAD, Strasbourg, France, Tech. Rep* (2010).

[194] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. "A large annotated medical image dataset for the development and evaluation of segmentation algorithms". In: *arXiv preprint arXiv:1902.09063* (2019).

[195] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, Joshua Dean, Michael Tradewell, Aneri Shah, Resha Tejpaul, Zachary Edgerton, Matthew Peterson, Shaneabbas Raza, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. *C4KC KiTS Challenge Kidney Tumor Segmentation Dataset*. 2019. URL: `https://wiki.cancerimagingarchive.net/x/UwakAw`.

[196] Holger Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M. Summers. *A new 2.5 D representation for lymph node detection in CT*. 2015. URL: `https://wiki.cancerimagingarchive.net/x/0gAtAQ`.

[197] Holger Roth, Amal Farag, Evrim B. Turkbey, Le Lu, Jiamin Liu, and Ronald M. Summers. *Data From Pancreas-CT*. 2016. URL: `https://wiki.cancerimagingarchive.net/x/eIlXAQ`.

[198] *Data Science Bowl Cardiac Challenge Data*. URL: `https://www.kaggle.com/c/second-annual-data-science-bowl`.

[199] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, Joshua Dean, Michael Tradewell, Aneri Shah, Resha Tejpaul, Zachary Edgerton, Matthew Peterson, Shaneabbas Raza, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes". In: *arXiv preprint arXiv:1904.00445* (2019).

[200] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. "The liver tumor segmentation benchmark (LiTS)". In: *arXiv preprint arXiv:1901.04056* (2019).

[201] Catalina Tobon-Gomez, Arjan J. Geers, Jochen Peters, Jürgen Weese, Karen Pinto, Rashed Karim, Mohammed Ammar, Abdelaziz Daoudi, Jan Margeta, Zulma Sandoval, Birgit Stender, Yefeng Zheng, Maria A. Zuluaga, Julian Betancur, Nicholas Ayache, Mohammed Amine Chikh, Jean-Louis Dillenseger, B. Michael Kelm, Saïd Mahmoudi, Sébastien Ourselin, Alexander Schlaefer, Tobias Schaeffter, Reza Razavi, and Kawal S. Rhode.

"Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets". In: *IEEE Transactions on Medical Imaging* 34.7 (2015), pp. 1460–1473.

[202] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[203] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18 (2021), pp. 203–211.

[204] Mubashir Ahmad, Danni Ai, Guiwang Xie, Syed Furqan Qadri, Hong Song, Yong Huang, Yongtian Wang, and Jian Yang. "Deep belief network modeling for automatic liver segmentation". In: *IEEE Access* 7 (2019), pp. 20585–20595.

[205] Song-Toan Tran, Ching-Hwa Cheng, and Don-Gey Liu. "A Multiple Layer U-Net, Un-Net, for Liver and Liver Tumor Segmentation in CT". In: *IEEE Access* (2020).

[206] Seisyou Kou, Luis Caballero, Raluca Dulgheru, Damien Voilliot, Carla De Sousa, George Kacharava, George D. Athanassopoulos, Daniele Barone, Monica Baroni, Nuno Cardim, Jose Juan Gomez De Diego, Andreas Hagendorff, Christine Henri, Krasimira Hristova, Teresa Lopez, Julien Magne, Gonzalo De La Morena, Bogdan A. Popescu, Martin Penicka, Tolga Ozyigit, Jose David Rodrigo Carbonero, Alessandro Salustri, Nico Van De Veire, Ralph Stephan Von Bardeleben, Dragos Vinereanu, Jens-Uwe Voigt, Jose Luis Zamorano, Erwan Donal, Roberto M. Lang, Luigi P. Badano, and Patrizio Lancellotti. "Echocardiographic reference ranges for normal cardiac chamber size: results from the NORRE study". In: *European Heart Journal–Cardiovascular Imaging* 15.6 (2014), pp. 680–690.

[207] Kevin Cheng, M Monaghan, Antoinette Kenny, Bushra Rana, Rick Steeds, Claire Mackay, and D van der Westhuizen. "3D echocardiography: benefits and steps to wider implementation". In: *British Journal of Cardiology* 25 (2018), pp. 63–68.

[208] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. "Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations". In: *International Conference on 3D Vision (3DV)*. IEEE. 2021, pp. 972–981.

[209] Shujaat Khan, Jaeyoung Huh, and Jong Chul Ye. "Pushing the Limit of Unsupervised Learning for Ultrasound Image Artifact Removal". In: *arXiv preprint arXiv:2006.14773* (2020).

[210] Mohammad Jafari, Hany Younan Azer Girgis, Nathan Woudenberg, Nathaniel Moulson, Christina Luong, Andrea Fung, Shane Balthazaar, John Jue, Micheal Tsang, Justin Bribe, Kenneth Gin, Robert Rohling, Purang Abolmaesumi, and Teresa Tsang. "Cardiac point-of-care to cart-based ultrasound translation using constrained CycleGAN". In: *International Journal of Computer Assisted Radiology and Surgery* 15.5 (2020), pp. 877–886.

[211] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. "Pixelnerf: Neural radiance fields from one or few images". In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 4578–4587.

[212] Cheng Sun, Min Sun, and Hwann-Tzong Chen. "Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction". In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.