First Trimester Gaze and Visual-Assisted Probe Motion Guidance for Obstetric Ultrasound Scanning

Elizaveta Savochkina

Linacre College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2023

Abstract

Pregnancy (or obstetric) ultrasound is routinely used worldwide for early detection of abnormalities in the fetus. Despite the possibility to reliably diagnose structural abnormalities in the first trimester (between 11 and 14 weeks of gestation), anomaly detection which makes anatomy visualisation easier is most commonly performed in the second trimester due to the larger size of the fetus and standardized anatomical imaging protocols.

The advancement in technology and ever evolving ultrasound image analysis and human-computer-interaction (HCI) research provides an opportunity to consider how to automate screening for abnormalities earlier in pregnancy. To effectively conduct routine first trimester ultrasound screening, a high level of expertise is required to navigate around the maternal womb, and record the necessary biometry planes.

In this thesis, a series of automatic image analysis algorithms are proposed to visually-assist and guide sonographers to important anatomical structures during a first trimester ultrasound examination. The overall research aim is to define tools to support the sonographer training and simplify clinical examination. The original contributions and medical data preparation procedures in this thesis:

1. Preparation of a multi-modal fetal ultrasound dataset involved curation of ultrasound videos and frames, IMU-assisted probe motion data and eye-tracking data. Each modality went through a number of data preparation procedures for a final multi-sensor alignment. Sensor data were also synchronised with technical and human annotations (where applicable). The curated dataset was used for work in this thesis by other lab members in their research [1].

2. A stochastic augmentation policy search method with two learned strategies, Random Augmentation (RA) and RA with Mixup (Mix.RA) for single frame saliency prediction was developed. The proposed policy improves the saliency map segmentation performance in the first trimester by predicting the eye gaze of a sonographer. In addition, the automated data augmentation reduces class imbalance and alleviates over-fitting, expands the medical training data and improves model generalisation. Using the learned policies, the augmentation strategies outperform the U-Net [2] baseline in all saliency metrics: KLD, SIM, NSS and CC (2.16, 0.27, 4.34 and 0.39 versus 3.17, 0.21, 2.92 and 0.28), see *List of Abbreviations* for metrics abbreviation.

3. A spatio-temporal convolutional network for video saliency prediction with stochastic augmentation is proposed. The architecture introduces a temporal dimension to analyse and improve saliency map segmentation in [3]. A video clip of 6 consecutive frames is found to record sufficient eye-gaze pattern variation to track changes in sonographer eye gaze trajectory. The proposed network outperforms the baseline approach from [3] on all saliency metrics: KLD, SIM, NSS and CC (2.08, 0.28, 4.53 and 0.42 versus 2.16, 0.27, 4.34 and 0.39). This model may be suitable for automatic guidance mechanism for real-time first trimester US scanning where the saliency predictions direct sonographer gaze to important anatomy. This is being investigated in an on-going translational project called the PURFECT study.

4. A prototype tracking system for freehand 3D ultrasound imaging that can visually-assist sonographers during an ultrasound examination is proposed. The prototype provides a digital representation of the fetus by means of 3D reconstruction. Using 2D ultrasound frames and IMU-assisted probe motion data, a fetal surface reconstruction in 3D space can be generated. The application of the proposed system can potentially reduce the time and effort required to mentally reconstruct, analyse and visualise a womb with solely visualizing 2D ultrasound scans, and help digitally pin-point the probe location during a scan.

The integration of these algorithms and prototype tracking systems into computer-assisted analysis tools may greatly reduce the time required to train sonographers, ease the mental workload with visual cues and assistive tools, and help in identification of anomalies in early stages of fetal development. Furthermore, the guidance algorithms will assist analysis and understanding of fetal structures by distilling 2D/3D information simultaneously from the womb.

First Trimester Gaze and Visual-Assisted Probe Motion Guidance for Obstetric Ultrasound Scanning

Elizaveta Savochkina

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

Acknowledgements

Personal

I am deeply grateful to my supervisor, Prof. Alison Noble, for granting me the opportunity to be a part of her meticulously curated research and to truly make a meaningful impact. Her unwavering encouragement, coupled with the freedom to work and explore in my unique way, has been invaluable. I extend my heartfelt thanks for her dedicated efforts and the countless hours she invested in assisting me with writing, research, and keeping sight of the bigger picture. It is widely acknowledged within the DPhil student community that a significant portion of a successful doctorate hinges on the guidance of the supervisor. Therefore, I am immensely proud and honored to have been mentored by Prof. Alison, and it is her invaluable guidance that has brought me to my current state of accomplishment. My gratitude knows no bounds.

I must also express my gratitude to the remarkable individuals in the Biomedical Image Analysis lab, which over time, seemed to naturally expand to encompass the entire floor of brilliant minds. I had the privilege of working and socializing with them, and they enriched my academic journey in countless ways. In particular, I want to acknowledge Dr. Mohammed Alsharid, my great friend, ardent supporter and cheerleader, as well as Dr. Robail Yasrab for his extraordinary friendship, Dr. Netzahualcoyotl Hernandez-Cruz for the endless fun, entertainment and positive attitude, Dr. Alex Gleed for our wonderful conversations and memorable violin concert visits, Dr. Clare Teng for her invaluable help and support even when unaware of it, Dr. He Zhao for the kind attitude and help in manuscript writing, Dr Yifan Cai for his guidance and his simplicity in explanations, Dr. Lee Lok Hin for his great assistance in coding, and Dr. Qianhui Men for the evenings spent decoding the intricacies of the server. I am thankful for the presence of dear friends like Divyanshu Mishra, Angus Nicolson, Pramit Saha, Jong Kwon, Yipei Wang, Dr. Richard Droste, Dr. Mourad Gridach, Dr. Zeyu Fu, Furat Aljishi, Mingze Yuan, Kangning Zhang, Dr. Md Mostafa Kamal Sarker, and Dr. Ping Lu in my life.

My sincerest thanks are due to my close friend and IT savior, Jamie Brett, for his prompt assistance, friendship, and all the important coffee breaks. A special appreciation goes to my best friend, Sogol Hosseini, who may not always see me but is always there for me when I need her. I would also like to extend my gratitude to

the superb MMA society, particularly coach James Halfhide, who made Brazilian Jiu Jitsu my favorite sport, and coach Josias Gomes, who demonstrated the profound benefits of Muay Thai for both the body and mind. Additionally, I want to thank my fellow members in the pool and snooker society, where I was able to hone my skills in my favorite non-contact sport, lead as a women's captain, participate in tournaments, and forge meaningful connections.

Lastly, my deepest appreciation goes to my family, which includes my parents, my partner, and my two furry friends. I want to express my gratitude to my father, Oleg Savochkin, who has been by my side every step of the way, providing both emotional and financial support throughout my PhD journey. My mother and grandmother, Marina Savochkina and Galina Dementeva, have consistently brought smile to my face. A special acknowledgment goes to my partner, Francis Lempp, who delved into every subject I encountered to find ways to support me and who held my hand through all the challenging moments over the past years. Last but not least, my little furry companions, bunnies Bennie and Maxwell, have been my most attentive listeners and unwavering sources of emotional support.

Institutional

I would like to acknowledge Linacre College for supporting my studies at Oxford and the PULSE project for providing valuable research data.

Abstract

Pregnancy (or obstetric) ultrasound is routinely used worldwide for early detection of abnormalities in the fetus. Despite the possibility to reliably diagnose structural abnormalities in the first trimester (between 11 and 14 weeks of gestation), anomaly detection which makes anatomy visualisation easier is most commonly performed in the second trimester due to the larger size of the fetus and standardized anatomical imaging protocols.

The advancement in technology and ever evolving ultrasound image analysis and human-computer-interaction (HCI) research provides an opportunity to consider how to automate screening for abnormalities earlier in pregnancy. To effectively conduct routine first trimester ultrasound screening, a high level of expertise is required to navigate around the maternal womb, and record the necessary biometry planes.

In this thesis, a series of automatic image analysis algorithms are proposed to visually-assist and guide sonographers to important anatomical structures during a first trimester ultrasound examination. The overall research aim is to define tools to support the sonographer training and simplify clinical examination. The original contributions and medical data preparation procedures in this thesis:

1. Preparation of a multi-modal fetal ultrasound dataset involved curation of ultrasound videos and frames, IMU-assisted probe motion data and eye-tracking data. Each modality went through a number of data preparation procedures for a final multi-sensor alignment. Sensor data were also synchronised with technical and human annotations (where applicable). The curated dataset was used for work in this thesis by other lab members in their research [1].

2. A stochastic augmentation policy search method with two learned strategies, Random Augmentation (RA) and RA with Mixup (Mix.RA) for single frame saliency prediction was developed. The proposed policy improves the saliency map segmentation performance in the first trimester by predicting the eye gaze of a sonographer. In addition, the automated data augmentation reduces class imbalance and alleviates over-fitting, expands the medical training data and improves model generalisation. Using the learned policies, the augmentation strategies outperform the U-Net [2] baseline in all saliency metrics: KLD, SIM, NSS and CC (2.16, 0.27, 4.34 and 0.39 versus 3.17, 0.21, 2.92 and 0.28), see *List of Abbreviations* for metrics abbreviation.

3. A spatio-temporal convolutional network for video saliency prediction with stochastic augmentation is proposed. The architecture introduces a temporal dimension to analyse and improve saliency map segmentation in [3]. A video clip of 6 consecutive frames is found to record sufficient eye-gaze pattern variation to track changes in sonographer eye gaze trajectory. The proposed network outperforms the baseline approach from [3] on all saliency metrics: KLD, SIM, NSS and CC (2.08, 0.28, 4.53 and 0.42 versus 2.16, 0.27, 4.34 and 0.39). This model may be suitable for automatic guidance mechanism for real-time first trimester US scanning where the saliency predictions direct sonographer gaze to important anatomy. This is being investigated in an on-going translational project called the PURFECT study.

4. A prototype tracking system for freehand 3D ultrasound imaging that can visually-assist sonographers during an ultrasound examination is proposed. The prototype provides a digital representation of the fetus by means of 3D reconstruction. Using 2D ultrasound frames and IMU-assisted probe motion data, a fetal surface reconstruction in 3D space can be generated. The application of the proposed system can potentially reduce the time and effort required to mentally reconstruct, analyse and visualise a womb with solely visualizing 2D ultrasound scans, and help digitally pin-point the probe location during a scan.

The integration of these algorithms and prototype tracking systems into computer-assisted analysis tools may greatly reduce the time required to train sonographers, ease the mental workload with visual cues and assistive tools, and help in identification of anomalies in early stages of fetal development. Furthermore, the guidance algorithms will assist analysis and understanding of fetal structures by distilling 2D/3D information simultaneously from the womb.

Contents

List of Figures

List of Abbreviations

US	Ultrasound
PULSE	Perception Ultrasound by Learning Sonographic Experience
FASP	UK Fetal Anomaly Screening Programme
1D, 2D, 3D	One- or two- or three-dimensional
ML	Machine Learning
DL	Deep Learning
CRL	Crown-Rump Length
NT	Nuchal Translucency
ROI	Region Of Interest
GT	Ground Truth
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
cLSTM	convolutional Long Short-Term Memory
cLSTMU-Net	convolutional Long Short-Term Memory with U-Net
RNN	Recurrent Neural Network
NN	Nearest Neighbour
KLD	Kullback-Leibler divergence
SIM	Similarity
NSS	Normalized Scanpath Saliency
CC	Pearson's Correlation Coefficient
FF	Freeze Frame
IMU	Inertial Measurement Unit
VSP	Video Saliency Prediction
RA	Random Augmentation
Mix.RA	Random Augmentation with Mixup

AHRS . . . . Attitude and Heading Reference System

DCM . . . . . Direction Cosine Matrix

NED . . . . . North-East-Down

The beginning is the most important part of the work.

— Plato

1 Introduction

Contents

1.1 Clinical Motivation

Pregnancy (or obstetric) ultrasound is routinely used worldwide for early detection of abnormalities in the fetus. In the UK, routine pregnancy ultrasound scans are standardized by the National Health Service under a Fetal Anomaly Screening Programme (FASP) [4]. There are three types of scans: a dating scan in the first trimester (between 11^{+2days} to 14^{+1days} weeks of gestation) to estimate gestational age, an anomaly scan in the second trimester (between 18 and 20^{+6days} weeks) that detects anomalies during fetal development, and a growth scan in the third trimester. The second trimester screening is a standard part of prenatal care, it offers a higher visibility due to a bigger size of the fetus.

For decades, ultrasound has been the standard modality to assess fetal gestational

age and overall fetal health. Ultrasound is economical, safe for mother and fetus due to no exerted radiation, and allows real-time visualisation [5]. Since the first reported attempts to image the uterus at 14 weeks' gestation made in 1958 [6], the modality has progressively become an indispensable obstetric care tool and plays an important role in neonatal care as well.

The newer generation ultrasound systems became simpler and more user-friendly, offering features like fewer dropdown menus, less keystrokes, faster processing times and the automation or semi-automation of measurements [7]. US machines can be equipped with a dual-probe technology as well [8] where transducers allow for the acquisition of real-time volumes of tissue in multiple planes simultaneously, i.e. the transverse and sagittal dimensions. Artificial Intelligence (AI) models are being integrated in the backend of ultrasound systems to enhance workflows, enable visual mapping, support anatomical annotation [9, 10], and hands-free operation [11] using AI-voice recognition controls.

Advances in technology have improved the spatial resolution and detail of ultrasound imaging in the first trimester, enabling detailed assessments of early fetal development. Karim et al. [12] performed a review and meta-analysis of relevant literature on detection of congenital fetal anomalies prior to 14 weeks' gestation. The paper argued the importance of the first-trimester ultrasound screening in early detection of fetal anomalies, recording anomaly detection rates of 32% in low-risk groups to more than 60% in high-risk groups. It is argued that the introduction of detailed anatomical protocols with standard anatomical views will prompt greater sensitivity and optimize first-trimester anomaly detection where structural abnormalities can be diagnosed reliably between 11 and 14 weeks. However, achieving consistent image quality remains a challenge, especially in suboptimal conditions due to patient movement, probe misalignment, or artifacts. Addressing the lack of high-quality images involves using ML models trained on diverse datasets, including images of varying quality. These models can generalize well and maintain robust performance, effectively identifying key structures under challenging conditions [13]. Preprocessing techniques such as noise reduction (i.e.

speckle filtering) [14, 15], and super-resolution [16, 17] can be embedded within ultrasound systems to enhance image quality in real-time, allowing for better downstream analysis and interpretation.

High level of expertise is required to perform routine ultrasound screening scans [18]. There is an active academic and commercial interest in developing automated image analysis algorithms to assist a sonographer in correctly identifying standardized planes in real-time during ultrasound image acquisition. While a large body of the academic computer-based image interpretation literature is based on single image analysis [19–22], the Oxford-based project PULSE - Perception Ultrasound by Learning Sonographic Experience - aims to capture expert level skills combining "machine learning, ultrasound video image analysis and human-computer-interaction (HCI) research" [23]. Data collected from eye tracking, IMU-assisted probe motion, and recorded sonographer speech are synchronized to create comprehensive guidance systems. This integration is essential for building robust training tools and ensuring real-time applicability. This alignment, however, poses challenges related to data synchronization for modalities with different sampling rates (high-frequency probe motion data must be downsampled to match the frame rate of ultrasound videos), buffering problems, real-time processing, and maintaining model performance when integrating multimodal inputs (discussed in detail in section 3.5.5).

The integration of AI has significant implications for enhancing sonographer proficiency and improving clinical outcomes. ML-assisted systems that provide real-time guidance and feedback can support sonographers by optimizing probe positioning and orientation, facilitating the acquisition of standardized imaging planes. Eye-gaze prediction models, which analyze historical gaze data from expert sonographers, offer another layer of assistance by predicting and overlaying the next focal points on the ultrasound video in real-time. These approaches help trainees develop an intuitive understanding of probe manipulation and attention focus, while experienced practitioners benefit from reduced cognitive load and improved navigation during scans (see Chapter 2 for a detailed literature review).

Despite advancements, aligning different modalities such as probe feedback, eye-gaze tracking, and ultrasound video requires meticulous calibration to ensure consistency and accurate interpretation. This is crucial to maintain algorithm reliability across various image qualities and validate models rigorously, preventing diagnostic errors. Continued refinement of training datasets and algorithmic improvements is essential for deploying these systems effectively in clinical settings, where inherent variability can challenge model performance.

The ultimate goal is to automate parts of ultrasound, perform image analysis and guide sonographers to important anatomies. The difficulty of acquiring desired planes and key anatomical structures during an ultrasound scan are well-documented in several studies [24–28]. These challenges persist for several reasons:

1. Inconsistency in ultrasound image. Factors such as acoustic shadowing, attenuation and artifacts lead to varied image quality.

2. Fetal size in the first trimester. Limited size affects the visibility of anatomical structures.

3. Variable orientation and position. Fetal movement can cause anatomical structures to be captured from different and unpredictable angles.

4. Probe maneuvering complexity. Acquireing standard planes requires significant training, and inter- and intra-operator variation contributes greatly to ultrasound image quality [29].

To address the challenges raised above, this thesis integrates image and video analysis with machine learning to develop methodology aimed at visually-assisting and guiding sonographers during first trimester ultrasound examinations. The research leverages sonographer gaze in the form of gaze-tracking data, XIO probe motion-tracking data and explores techniques such as sonogarpher eye-gaze prediction and the virtual pinpointing of the probe's location on a 3D reconstructed fetus.

1.2 Contributions and Thesis Structure

The primary contributions of this thesis are a series of automatic image and video analysis tools to assist sonographers during ultrasound scans. The research

aims to bridge the gap between technical development and clinical application, ultimately providing real-time guidance to help sonographers identify key anatomical structures while freely navigating the maternal womb.

Chapter 2 introduces the clinical motivation for each subsequently developed method. It elaborates challenges and limitations that exist in current fetal ultrasound examination. Next, the chapter describes possible image and video analysis solutions and reviews progress made that is reported in related literature.

Chapter 3 provides a detailed description of data used in this thesis. It describes the data acquisition, data preparation, annotation (if any), the subsets of data used in each chapter, and the detailed train/validation/test sets partitioning. Data visualisation is also provided for better understanding of the dataset.

Chapter 4 investigates the benefits of data augmentation in medical imaging to combat class imbalance and shortage of data for a specific class, increase model generalization, and expand training data. First, the chapter presents a single frame saliency prediction method for first trimester ultrasound images. Next, a stochastic augmentation policy search method is implemented to improve model segmentation performance when predicting where a sonographer will look next. Using a simple grid search and different types of augmentation transformations, ultrasound images and ground truth saliency maps are transformed producing new artificial examples. A common approach is to use augmentation transformations that result in training images that look natural and realistic. However, the hypothesis is formulated such that the addition of distorted images in training can improve model generalization performance, where the augmentation process generates artificial images with additional fetal appearance and position characteristics as well as ultrasound artifacts/noise that inevitably appear during the scan. The proposed augmentation policy search method includes two strategies, Random Augmentation (RA) and RA with Mixup (Mix.RA). Strategies are evaluated using benchmark saliency metrics [30]. The results show that a stochastic augmentation policy

search method increases segmentation accuracy and alleviates over-fitting in first trimester ultrasound saliency prediction.

Chapter 5 builds on the findings of Chapter 4 by utilizing naturally temporal ultrasound data and gaze tracking to inform saliency prediction across sequences of frames. This chapter focuses on differentiating between fast and slow-moving temporal segments to refine gaze prediction on unseen ultrasound frames. It introduces the spatio-temporal convolutional LSTMU-Net neural network (cLSTMU-Net), designed for video saliency prediction with stochastic augmentation. The architecture design is built to find an optimal number of consecutive video frames that capture variations in eye-gaze patterns over time and better track sonographer gaze shifts. Evaluation indicates that optimal performance is achieved when using 6-frame sequences, balancing between capturing meaningful gaze dynamics and computational efficiency. The model benefits from incorporating temporal information from preceding frames, which enhances the accuracy of predicted saliency maps.

This chapter also discusses the application of this approach in automatic guidance systems that direct sonographer gaze to key anatomical structures in real time. Ongoing clinical studies, such as the PURFECT study, are assessing the model's impact on diagnostic accuracy and workflow efficiency, providing tangible evidence of its clinical utility. The trained model is embedded in an ultrasound machine, where real-time eye-gaze predictions are overlaid on the ultrasound video. Sonographers have the option to adjust the intensity of these gaze predictions during scans. Sonographers assess the accuracy of predictions, noting whether the guidance aligns with their preferred scanning practices (i.e. if the predicted eye-gaze is in a correct place) and helps highlight important anatomy, or if it becomes a distraction or is not beneficial.

Chapter 6 presents an initial prototype tracking framework for freehand 3D ultrasound imaging that leverages existing 2D ultrasound data, as commonly used in clinical settings. This research was inspired by discussions with sonographers,

who emphasized the need for a guidance system that could display the fetus during a scan with the probe's location pinpointed on it, aiding in the effective teaching of novice sonographers. This approach is designed to work entirely within the constraints of a standard 2D ultrasound probe without additional external hardware, ensuring practical applicability for hospitals that predominantly use 2D probes. The chapter combines IMU-based orientation sensor technology with a transducer to determine positional and orientation coordinates of an ultrasound probe at every time stamp. Fetal mask edges 2D ultrasound frames and probe motion data are synchronised and combined to derive a fetal surface reconstruction in 3D space. This reconstruction is refined through Delaunay triangulation to achieve a final surface reconstruction and visualisation. By displaying the virtual position of the probe on the reconstructed 3D fetus, this prototype aims to visually assist sonographers, improving their spatial awareness and making probe navigation more intuitive during fetal imaging. The contribution also addresses synchronization challenges inherent to combining motion tracking with 2D ultrasound, facilitating a more seamless integration into existing clinical workflows.

Future clinical studies can demonstrate the significance of this prototype by assessing its impact on diagnostic accuracy, workflow efficiency, and cognitive load reduction. Comparative trials with standard 2D imaging and sonographer feedback will provide validation. Motion correction features, as discussed in the conclusion, would further solidify its role in enhancing real-time navigation and training in fetal imaging.

Chapter 7 concludes the thesis by discussing the broader implications and potential applications of the developed methods. It also outlines future research directions, including further clinical evaluations and enhancements to the proposed tools to support advanced real-time guidance and image analysis.

1.3 List of Publications

1.3.1 Peer-Reviewed Conference Proceedings Papers

Elizaveta Savochkina, Lok Hin Lee, Lior Drukker, Aris T. Papageorghiou and J. Alison Noble. First Trimester Gaze Pattern Estimation Using Stochastic Augmentation Policy Search for Single Frame Saliency Prediction. In an Annual Conference on Medical Image Understanding and Analysis (MIUA) (pp. 361-374), 2021 - Oral presentation. See Chapter 3.

Elizaveta Savochkina, Lok Hin Lee, He Zhao, Lior Drukker, Aris T. Papageorghiou and J. Alison Noble. First Trimester Video Saliency Prediction Using cLSTMU-Net with Stochastic Augmentation. At the IEEE International Symposium on Biomedical Imaging (ISBI), 2022. See Chapter 4.

1.3.1.1 Published work led by others that builds on the thesis work

Mourad Gridach, **Elizaveta Savochkina**, Lior Drukker, Aris T. Papageorghiou and J. Alison Noble. Self-Knowledge Distillation for First Trimester Ultrasound Saliency Prediction. In a workshop on Simplifying Medical Ultrasound (ASMUS) (pp. 117–127), 2022. Self-knowledge distillation (SKD) is a novel machine learning technique that allows a shallow student network to distill its own knowledge, reducing computational complexity. In the context of medical imaging and predicting sonographer gaze, using a multi-modal ultrasound and gaze dataset an original SKD framework with a Wide Feature Distillation module and ReSL loss demonstrates superior performance compared to existing models in [1].

1.3.2 Journal Articles

Robail Yasrab, **Elizaveta Savochkina**, Lior Drukker, Aris T. Papageorghiou and J. Alison Noble. Automating the human action of first-trimester biometry measurement from real-world freehand ultrasound. At the IEEE Transactions on Medical Imaging (TMI) (under second review).

2

Review of Multi-modal Fetal Ultrasound Image and Probe Motion Assisted Analysis

Contents

This literature review chapter outlines the key challenges and progress made in fetal ultrasound image analysis and probe motion analysis. From the clinical motivation discussed in section 1.1, the application focus of our research is on the first trimester ultrasound screening. First, the chapter reviews ways in which image and video analysis, human gaze and saliency prediction can help guide a sonographer to important fetal anatomical structures. Next, the review discusses literature on 3D US fetal volume reconstruction using probe motion and ultrasound image data. The order of discussion is arranged accordingly following the focus of each contribution chapter presented in this thesis. For each topic, the theory and the applications are first described. Next, an image analytic tool is proposed to address the problem, followed by a review of strengths and weaknesses of established methods. The review also briefly describes relevant work in the fields of computer vision, robotics and automation to place the thesis contributions in context.

2.1 Image and Video Analysis

In this section we review relevant literature on automated image and video analysis. In recent years this has been dominated by artificial intelligence (AI), specifically, deep learning (DL) which is discussed in section 2.1.1. Section 2.1.2 reviews data augmentation techniques and strategies. Sections 2.1.3 and 2.1.4 provide an introduction to semantic, object detection and instance image segmentation, respectively.

2.1.1 Deep Learning

Artificial intelligence and deep learning (DL) have the potential to fundamentally alter the way medicine is practiced. DL involves the development and use of optimisation algorithms to perform tasks that normally require human intelligence, such as visual perception, pattern recognition, decision-making and problem solving, at a similar or improved level of performance [31]. Deep learning could have particularly transformative applications in fetal ultrasound given the multifaceted

and highly technical nature of this field of medicine with a heavy reliance on recognition and computer software.

Deep learning is a subset of machine learning (ML) and has been on a rise since its first successful application to solve voice recognition and image analysis problems a decade ago [32]. Figure 2.1 shows the distinction between classic machine learning and deep learning.

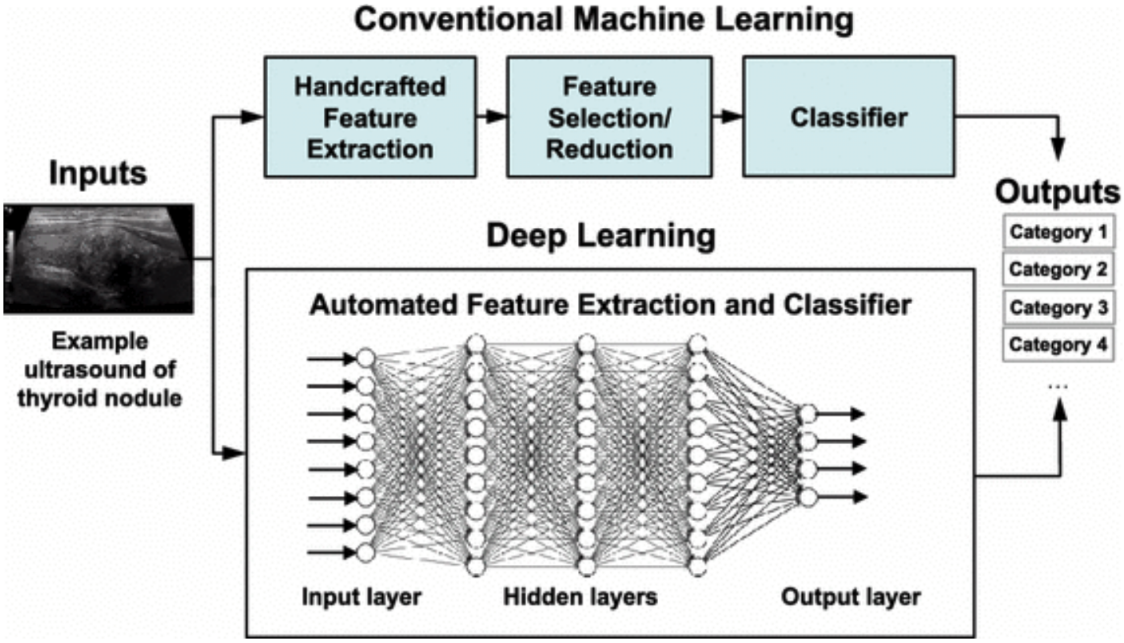


Figure 2.1: Work pipeline change for classic Machine Learning and Deep Learning [33].

As shown in Figure 2.1, early ML algorithms were predicated on rule-based reasoning performed by a computer system according to a set of steps and procedures defined by human experts. However, the generalizability of these methods on variation of the input data and task scope is often limited due to out-of-distribution cases which are not explicitly described in the ML pipeline. Over the past decade, Artificial Neural Network (NN) algorithms powering the automation of image-based tasks have dominated the medical field.

Deep learning has no need a hard core feature extraction [33], i.e human defined rules, as it learns incrementally through it's hidden layer architecture (illustrated with a neural net in Fig. 2.1). A DL algorithm learns from the training data provided and makes predictions; this approach has demonstrated high accuracy

and state-of-art results in a variety of applications [34–36]. Based on the presumed understanding of how neurons work in the human brain, NNs are implemented as representations of chains of neurons where each neuron takes an input and produces an output, depending on the conditions initially provided.

In general and in theory, DL methods can be highly effective when a model is trained on a large, curated and representative of all the cases dataset. However, in medical applications, there can be a limited number of available samples and little to no expert labeling (human annotations). Annotated data aids in the analysis of data distribution and its characteristics. Consequently, one of the main challenges in applying deep learning to medical images is in building deep models that do not suffer from over-fitting, i.e. unable to generalize to unseen examples.

When talking about deep learning, many refer to Convolutional Neural Networks (CNN) [37], i.e AlexNet [38]. However, there are also other DL algorithms, such as Auto-Encoders [39] and Deep Boltzmann Machines [40], GAN [41], LSTM [42] and others.

A simple CNN is made up of neurons with corresponding learnable weights and biases; each neuron performs a dot product of its input and its own parameters, optionally followed by the application of a non-linear transformation; the whole network receives input at one end, expresses a score function, and outputs a score vector as output; and the network can be learnt end-to-end [43]. Unlike NNs, CNNs are well-suited for image analysis because they leverage both spatial hierarchies of features (by capturing relationships between pixels) and parameter sharing (using the same filters across different regions). This allows CNNs to efficiently learn and detect features that are spatially related within an image, making them highly effective for visual tasks.

Although, AlexNet popularized CNNs in computer vision, the network was over-fitting due to a high number of parameters. Authors introduced Data Augmentation and Dropout to help reduce the differences in training and test data distributions.

2.1.2 Data Augmentation

As mentioned in section 2.1.1, deep learning models such as CNNs or others are highly effective when a model is trained on a large, curated dataset which represents all classes equally (i.e. no bias in a model towards a majority class with higher number of samples). In a classic computer vision image processing phase, networks must overcome issues of viewpoint, lighting, occlusion, background, scale, and more. In a medical application, medical data requires rights to be accessed (patient privacy and confidentiality) which limits public access. The medical image data also presents with higher occlusion, undesirable noise, shadows and artifacts.

Many application domains do not have access to big data and medical image analysis is one of them. Not only is there a need for high number of samples for model training but also the need for expert labeling, i.e. human annotations if supervised learning is needed. Annotations allow to classify the dataset into groups or classes to analyse how balanced and diverse the dataset in question is. Human annotations are expensive and labor-intensive to collect which causes the neural networks to develop biases toward certain patterns or classes that are overrepresented in the data.

Data Augmentation (DA) encompasses a suite of techniques that enhance the size and diversify training datasets which allows researchers to build efficient deep learning models. DA can be seen as a solution to the problem of limited data as well as a mechanism which can diversify a dataset with under-represented classes.

DA is not the only technique proposed to reduce over-fitting; other strategies for increasing generalization performance modifying focus on the model architecture itself. Functional solutions such as dropout regularization, batch normalization, transfer learning, and pre-training have been developed to try to extend DL for application on smaller datasets. Over-fitting solutions and a complete survey on regularization methods for DL models has been compiled by Santos and Papa [44].

In contrast to the techniques mentioned above, data augmentation approaches over-fitting from the root of the problem, the training dataset. The augmentations artificially inflate the training dataset by either data warping or oversampling. The taxonomy of image data augmentation techniques is covered in Fig. 2.2.

Data warping augmentations transform existing images preserving their label, i.e. geometric and color transformations, random erasing, adversarial training, and neural style transfer. Oversampling augmentations create synthetic instances and add them to the training set. This includes mixing images, feature space augmentations, and generative adversarial networks (GANs) [45].

Figure 2.2: A taxonomy of Image Data Augmentation techniques. The colored lines in the figure depict which data augmentation method the corresponding meta-learning scheme uses [45]. For example the work in Chapter 4 is mainly covered by the Basic Image Manipulations (with addition of an improved AutoAugment method, RandAugemnt [46]) which are displayed on the left hand side of the taxonomy classification system, i.e. Color Space and Geometric Transformations, Mixing Images and Kernel Filters for creation of saliency maps.

Classic augmentation transformations used in medical image analysis include random transformations such as scaling, rotation and translation that make training images look natural and realistic [47]. Other simple transformations include horizontal flipping, color space augmentations, and random cropping. Whilst data warping

augmentations transform existing images preserve their label, oversampling augmentations create new artificial examples and show top end performance [46, 48–50].

In ultrasound (US) imaging, specific augmentations are particularly beneficial due to the modality's unique characteristics. For instance, rotation invariance is critical, as sonographers often rotate the probe to capture optimal anatomical views. Including rotated frames during training enables models to recognize structures like the nuchal translucency or palate regardless of orientation, improving generalization across various probe angles. US-specific augmentations can also focus on intensity transformations that mimic common ultrasound artifacts and variations. These include speckle noise augmentation, which simulates the grainy texture often seen in ultrasound images, and brightness or contrast adjustments, reflecting the natural variability in image quality due to factors like patient body composition or probe positioning.

Mixing images falls under augmentations which do not look like a useful transformation to a human observer. However, Ionue et al. [51] showed how pairing data samples reduces error rate from 8.22% to 6.93% on image classification task. Two images are randomly cropped, horizontally flipped and mixed by averaging the pixel values for each of the RGB channels. The label assigned to the new image is the same as the first randomly selected image. Authors also found that better results were obtained when mixing images from the entire training set rather than from instances exclusively belonging to the same class. Summers and Dinneen [50] further looked at using non-linear methods to combine images into new training instances. Takahashi et al. [52] mixed images by randomly cropping them and concatenating the output together to form new images. Similar to SamplePairing and mixed-example augmentation, Smart Augmentation [53] combines existing examples and produces one. However, the mechanism is more sophisticated and uses an adaptive CNN to derive new images instead of averaging pixels.

Previous works on automated augmentation strategies [49, 54–56], at the time, demonstrated top results in image classification and object detection. However, these methods require a separate and expensive search phase and make it challenging

to adopt a policy on a large-scale. The RandomAugment (RA) paper [46] introduces a search space with the regularization strength that can be tailored based on model and dataset size. RA was found to significantly reduce a search space and matched or outperformed predictive performance of augmentation methods mentioned above. RA is one of the augmentation strategies that is used in Chapter 4.

There are many other oversampling techniques which include GAN-based [57–59] and meta-learning [60] augmentations, however, an exhaustive list exceeds the scope of this review.

2.1.3 Semantic Segmentation

A number of approaches have used CNNs [61–67] for semantic segmentation, in which each pixel is labelled with the class of its enclosing object or region. Long et al. [68] modified these networks for segmentation specifically. The final classification layer was removed as well as fully-connected layers replaced with convolutional layers instead. The author proposed a fully convolutional network (FCN), trained end-to-end, pixels-to-pixels on semantic segmentation which, in turn, was the first semantic segmentation network (model displayed in Figure 2.3).

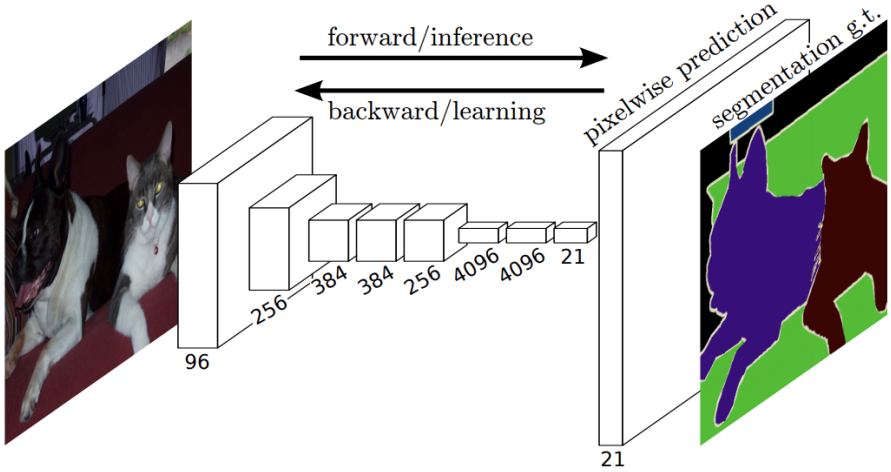


Figure 2.3: A fully convolutional image segmentation network [68].

U-Net [2] was proposed by Ronneberger based on a FCN and has become a popular architecture in the medical imaging community for image segmentation as

it is better suited than an FCN to smaller training datasets (relative to general computer vision) typically seen in medical imaging. A FCN uses deconvolution to restore image size and features. The goal of U-Net is to identify the location and shape of different objects in the image by classifying every pixel with a label.

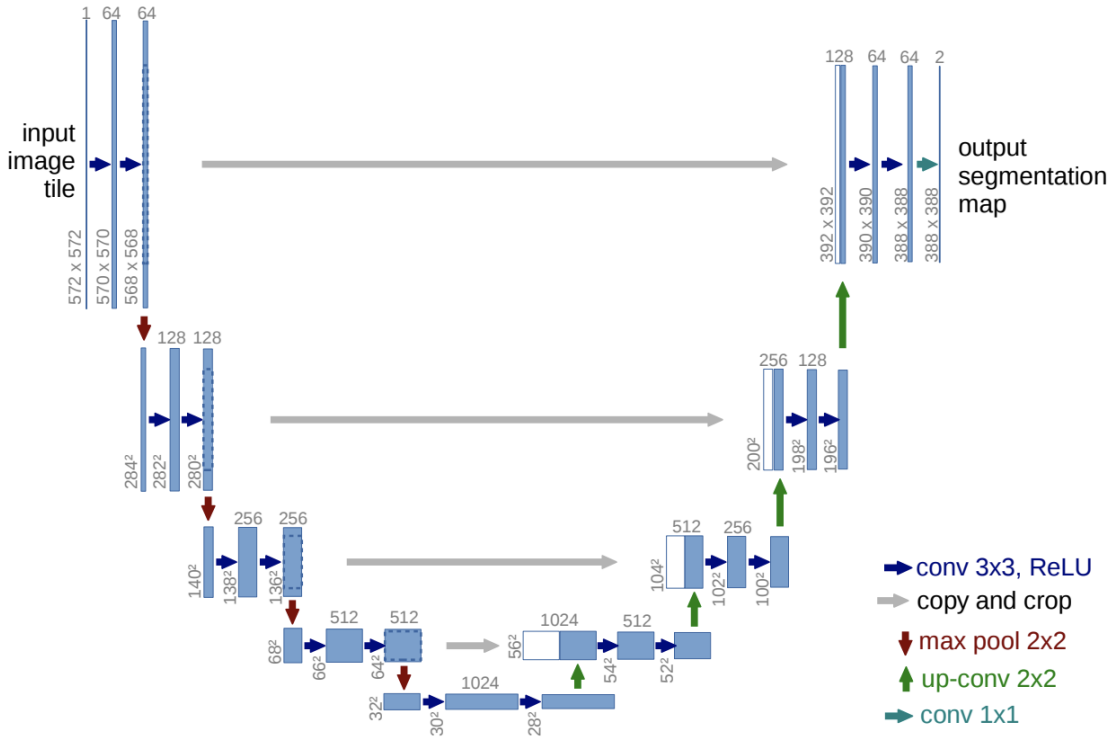


Figure 2.4: U-Net architecture. The blue boxes denote feature map blocks with their indicated shapes[2].

From Figure 2.4, the U-Net architecture consists of a contracting path to capture context by continuously merging the convolution layers to extract features, and a symmetric expanding path that enables precise localization. The symmetry of layers makes a network look like a "U" shape.

The step of the encoder reducing the image size is called down-sampling, and the step of the decoder reducing the image details and size is called up-sampling. Before up-sampling of the encoder and the down-sampling of the decoder, concatenation to splice the feature maps is implemented, followed by deconvolution [69]. U-Net adopts the skip connection strategy of splicing applied to shallow feature information of all scales to make full use of the features of the encoder to be used for up-sampling.

Many encoder-decoder networks benefit from skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low level, fine-grained feature maps from the encoder sub-network, and have proven to be effective in recovering fine-grained details of the target objects [70–72], even for a complex background [73, 74]. There are a number of other semantic segmentation methods [75–78] and U-Net variants [79, 80] but an exhaustive list will exceed the scope of this review.

2.1.4 Object Detection and Instance Segmentation

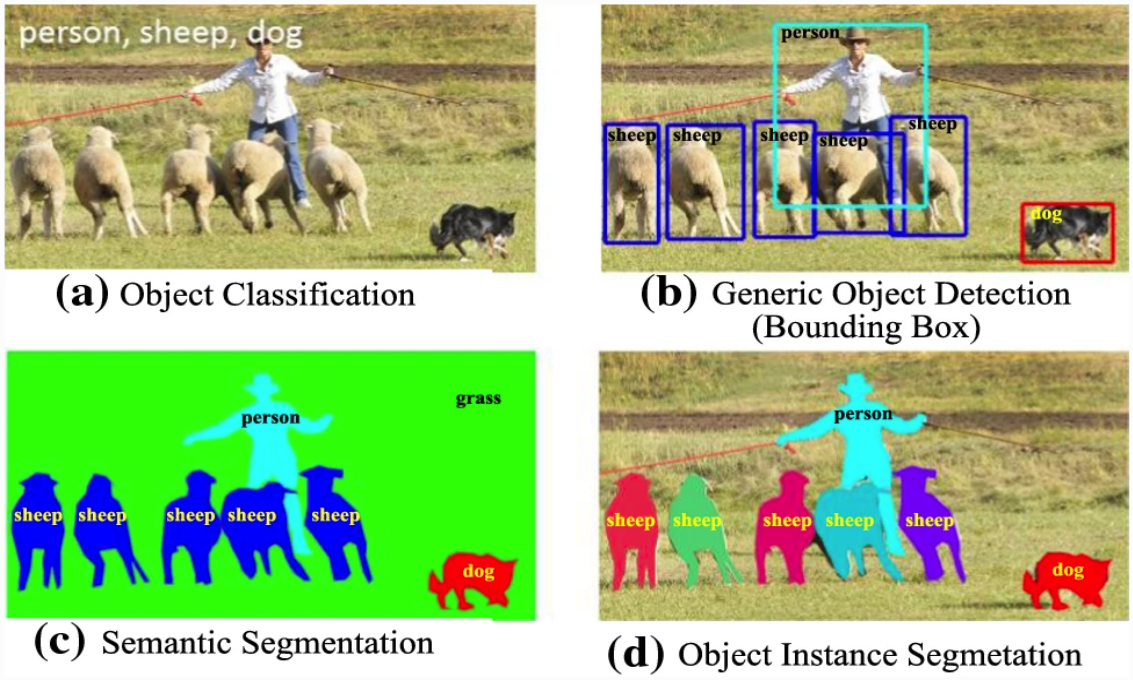


Figure 2.5: Recognition problems: a) image level object classification, b) bounding box level generic object detection, c) pixel-wise semantic segmentation, d) instance level semantic segmentation [81].

Figure 2.5 shows the progress from coarse to fine inference of recognition problems. Object detection is improved on object classification by detecting instances of semantic objects of a certain class. Classic object detection places a rectangular bounding box corresponding to each class in an image, however, it provides no information about the shape of an object. Semantic segmentation solves problems object classification cannot, providing the classes as well as their spatial location

(Fig 2.5c). Instance segmentation goes beyond localization/object detection to distinguishing different instances of the same object class. In Figure 2.5d, instance level semantic segmentation distinguishes between the different sheep which are of the same category.

Girshick et al. [82] extended an ordinary CNN to a regional CNN (R-CNN) improving the mean average precision of second-best result by over 30% on VOC 2012 test detection of average precision. The R-CNN and its extensions (Fast R-CNN [82], Faster R-CNN [83], Masked-RCNN [84]) have proven successful in ultrasound object detection applications. An example of a Faster R-CNN is presented below.

The Faster R-CNN [83] architecture developed for object detection uses a region proposal network (RPN) to replace selective search. The RPN extracts a Region of Interest (RoI), and a RoIPool (maxpooling) layer computes feature maps from these proposals in order to infer the bounding box coordinates and the class of the object. Masked-CNN is an extension on Fast R-CNN and is one of the top-performing methods on the COCO instance segmentation task [84]. Masked-CNN detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance by adding an extra output branch for object mask prediction. There are many other instance segmentation models (such as R-FCN [85] and YOLO [86]) that have been developed based on R-CNN but an exhaustive list exceeds the scope of this review.

2.2 Gaze, Saliency, Attention and Ultrasound Image Analysis

This section reviews literature on gaze prediction, saliency and attention. Sections 2.2.1 and 2.2.2 provide an introduction to human gaze and saliency prediction, respectively. Section 2.2.3 reviews visual attention in computer vision applications, whilst Section 2.2.4 discusses gaze in medical imaging-based diagnostics.

2.2.1 Human Gaze

Human eye movement has been extensively studied in cognitive science and experimental psychology. In order to better understand and explain the human visual system, a number of computational models of visual attention [87–89] have been proposed and developed.

There are distinctions in terminology of scene perception where the definitions of fixations and saccades are studied [90]. *Gaze* is the direction of sharp central vision (fovea) of the eye's receptive field towards a location in the scene. The location in the scene is the gaze point. *Fixations* are points of gaze stability that can only be acquired due to saccadic suppression [91]. *Saccades* are the rapid movements of gaze points between fixations. In addition, we will specify overt and covert attention in the subsequent discussion. *Overt* attention shifts are saccadic eye movements to a target, whilst *covert* attention corresponds to attending a target without fixation [92].

Land and Hayhoe [93] found that the eye movements are 'forward planning', i.e. seeking out objects for future use and lead each action as opposed to responding to stimuli. It was noted that the gaze precedes action by a fraction of a second. In addition to that, Henderson [94], identified the gaze as the main component in comprehension and exploration of visual perception. One of the reasons for this was that eye movements are nearly all associated with task-relevant objects. Therefore, fixations are directed towards important scene regions. Another reason is related to eye movements revealing the overt attention where the eye is focused, and covert attention where they are not. Finally, "eye movements provide an unobtrusive, sensitive, real-time behavioral index of ongoing visual and cognitive processing" which can be tracked [94].

2.2.2 Saliency Prediction

As discussed in section 2.2.1, human gaze is interesting for computer vision since fixations provide important information about a scene and can be measured via eye tracking. Studies of gaze on images and patterns of perception led to a concept of saliency in images – saliency maps. In the context of visual processing, saliency

maps are a topographical representation of unique features of an image that depict visually alluring locations [95].

Using saliency maps, one can predict the probability that the user is likely to fixate on a given region. The predictive features could be spatial or spatio-temporal, local or global, and inherent or task-related. It is of importance to note that covert attention refers to attention allocation without eye movement [96] with some [97] furthering the claim and stating that covert attention is independent of overt saccadic shifts. Whereas, others debate a connection between covert attention and microsaccades [98, 99]. However, up until now, there has been no reliable method for covert shifts detection and regardless of the true role of covert attention, its existence poses a difficulty for treating human fixation data as the ground truth for the allocation of human attention [96].

2.2.2.1 Image Saliency Prediction

A lot of work has been done on image saliency prediction (prediction of salient areas in images) for natural images [100, 101]. Most saliency models are validated on images of natural scenes using ground truth representations in the form of human fixations or saliency maps (discrete fixations converted into a continuous distribution). Motivated by deep learning results in vision tasks, Vig et al. [102] compared 32 state-of-art models that predict human fixation location and scan path sequences. Subsequently, Kummerer et al. [12, 103] proposed two deep saliency prediction networks: DeepGaze I (based on the AlexNet) [104] and DeepGaze II [105] built upon the VGG-19 [106]. Deep Gaze I network was developed to predict human eye fixations by reusing existing neural networks that have been pre-trained on task of object recognition, specifically fixation prediction. The model was built in a transfer learning manner where instead of implementing a deep search for parameters to fit the model, Deep Gaze learns weights for the linear combination of responses utilizing each layer of AlexNet.

Static saliency has been studied using CNN-based models [107–115] that yield superior performance, rapidly becoming state of the art for static saliency predic-

tion. Large-scale eye-fixation datasets that contributed to the progress of deep saliency predictions are SALICON and MIT1003 [116], with corresponding saliency benchmarks SALICON and MIT seen as a gold standard for evaluating single-image saliency models.

2.2.2.2 Video Saliency Prediction

Saliency prediction in videos differs from single-image prediction due to temporal correlation between fixations [117]. In static images only spatial cues need to be considered, whilst in video saliency prediction (VSP) temporal cues add a layer of complexity with motion scenarios across frames. As fixations are affected by subjective consciousness of the eye where the attention changes due to camera motion, Chen et al. [118] states that an effective spatio-temporal feature representation is key to VSP as it can "guide attention orientation shift prediction under various scene transition".

Depending on a task, the static saliency models are extended to the video scenario by incorporating motion features [119–121]. Wang et al. [119] proposed a two-model architecture to exploit spatio-temporal features: the first module performs frame-level saliency prediction; the second module, instead, takes pairs of frames with saliency predicted by the first module, and generates a dynamic saliency map. SalEMA [122], instead, proposes a 2D encoder-decoder architecture with a recurrent module added to the bottleneck for integrating temporal information provided by the previous frames. A CNN-based approach by Chaabouni et al. [123] detects salient areas in natural videos by adding optical flow as an additional input channel to a single-frame CNN. Droste and Jiao [124] integrate four novel domain adaptation techniques into an encoder-RNN-decoder-style network and achieve state-of-art on all video saliency datasets DHF1K, Hollywood-2 and UCF-Sports, and on par with the image saliency datasets SALICON and MIT300.

Motion cues have also been included in saliency prediction through either convolutional recurrent networks or RNNs applied to spatial feature encodings. Jiang et al. [125] argues that human attention is attracted by moving objects which

led to development of an object-to-motion CNN (OM-CNN) to predict intra-frame saliency. OM-CNN is a dual-stream network that extracts spatial and temporal features using YOLO [86] and FlowNet [126], whose respective objectness and motion features are then combined via a two-layer ConvLSTM. Similarly, ACLNet [127] performs static saliency prediction through attention module that performs a global spatial operation on learned features. These features are then given to a ConvLSTM to model temporal information. Wu et al. [128] proposed a SalSAC architecture where multi-level features are extracted by an encoder and input into a shuffled attention mechanism. A self-attention module (mechanism which uses different positions of a sequence to compute representation of the same sequence) is adopted from [127, 129] to better model spatial and dynamic saliency. Correlation features between multi-level features and shuffled attention on the same features are fed into a ConvLSTM to learn temporal cues.

2.2.3 Visual Attention in Computer Vision

This section focuses on the application of visual attention to computer vision tasks such as classification or segmentation. Selective attention demonstrates that objects or pixels are not treated equally. Attention in deep learning localizes the most important information in an image or video.

There are different types of so-called visual attention; in particular, hard (stochastic) attention, and soft (deterministic) attention. In order to use neural networks with integrated visual attention, one can select the region of interest (or foveation) for downstream processing to be human vision mimicking. In contrast, soft attention, weights values of *all* inputs used in attention calculation, the value of which is annuled by the loss function, hence, is differentiable. Deterministic attention uses attention maps [130] applied to the feature maps via element-wise multiplication. Larochelle and Hinton [131], presented a model inspired by the human retina with the resolution that can only cover a small area of the image, therefore, the model is forced to decide upon the sequence of fixations and their "glimpse" at each fixation. That paper discusses a fixation point strategy incorporated in a

restricted Boltzmann machine (RBM) and demonstrates a system that can make use of a variable resolution retina containing very few pixels. Chen et al. [132] presented a visual attention-based CNN which combines human multi-glance and an attention mechanism which is fed into a VGG-Net [106]. The architecture shows improvements in robustness and proficiency in fine-grained classification. Mnih et al. [133] argue that a CNN is computationally expensive and instead, proposes RNN which adaptively selects a sequence of regions, processes only them with high resolution and extracts information. Luo et al. [134] introduced a dual-stream model which is half classification model, and half saliency prediction model trained with human fixations. The second part of the model improved multilabel image classification performance.

2.2.4 Gaze in Medical Imaging-Based Diagnostics

Image perception is seen as a prior in understanding of clinical decision making [135, 136]. Nodine and Kundel [137], looked into the visual search strategy of radiologists in lung tumor detection. Voisin et al. [138] found that diagnostic errors in mammography can be predicted to a good extent by leveraging the radiologists' gaze patterns and image content. Tourassi et al. [139] also looked at the mammography domain and linked image content, gaze, and cognition and claimed valuable insights if using machine learning techniques. Bertram et al. [140] performed expert-novice-comparisons on multi-slice CT images, arguing that radiologists needed fewer fixations on the relevant areas than the radiographers. Wang et al. [141] process radiologists' gaze to supervise the DNN's attention via an attention consistency module to asses knee X-rays.

Cai et al. [27] presented SonoEyeNet where he exploited eye movement data of a sonographer in automatic interpretation. Patra et al. [142] proposed PeTRA, a teacher-student knowledge transfer framework, where a combined gaze tracking and US video information are an input into a large teacher model. The teacher model transfers knowledge to train compact models performing frame classification for the fetal abdomen, head, and femur. Droste et al. [143] studied the safety

of ultrasound when scanning pregnant women. Tracking the sonographers' eye-movements, the authors were able to detect whether clinicians were paying attention to the safety indices or not. That paper showed that the displayed bioeffect safety indices were looked at in 27 routine scans (4.2%), by 4 of the 17 operators. In all 27 scans, safety indices were checked once. Further related work on video saliency prediction is discussed in Chapter 5.

2.3 Motion Tracking using Inertial Sensors

The latest advancements in tracking technologies have enabled conventional medical devices to be equipped with more advanced functions. In biomedical US imaging, object tracking technologies are key to locate US probes and other medical tools for precise operation and intuitive visualization [144]. There is a number of commonly used tracking technologies, however, this section will solely focus on inertial sensors such as Inertial Measurement Unit (IMU). The underlying physical principles behind the inertial tracking, sources of error, calibration and relevant literature are reviewed below.

2.3.1 Overview of Position and Orientation Estimation

Inertial tracking systems are based on an inertial measurement unit (IMU), which is a small, lightweight, cost-effective sensor enabled by micro-electromechanical systems (MEMS) [145]. It is a technology that can be easily integrated into engineering systems with minimal burden or hardware dependency.

Inertial sensors can be used to provide information about the pose of any object that they are rigidly attached to and its orientation. An IMU sensor with 9 axes that integrates accelerometers, gyroscopes, and magnetometers is commonly used for 6 DOF object tracking. The accelerometer measures the external specific force acting on the sensor. The specific force consists of both the sensor's acceleration and the earth's gravity. The gyroscope measures the angular rates of the sensor when rotated and the magnetometer detects the magnetic field strength of the sensor location.

As a result, after calibration and compensation for drifts and errors (discussed in section 2.3.2), the position and orientation of the target can be determined [146]. Specifically, relative rotation and orientation are derived using angular velocity of a gyroscope, whilst relative position is derived using linear acceleration of an accelerometer.

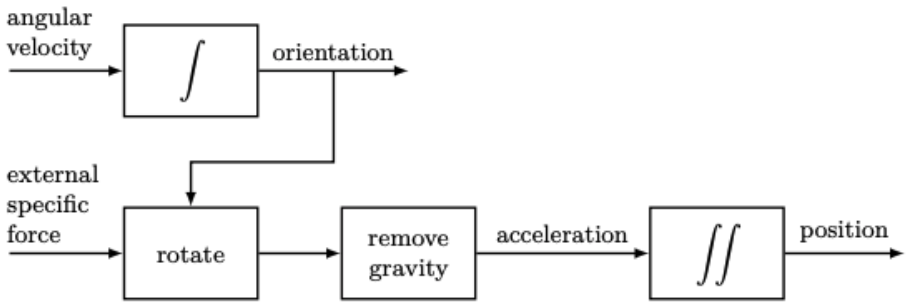


Figure 2.6: Schematic illustration of dead-reckoning, where the accelerometer measurements (external specific force) and the gyroscope measurements (angular velocity) are integrated to estimate position and orientation.

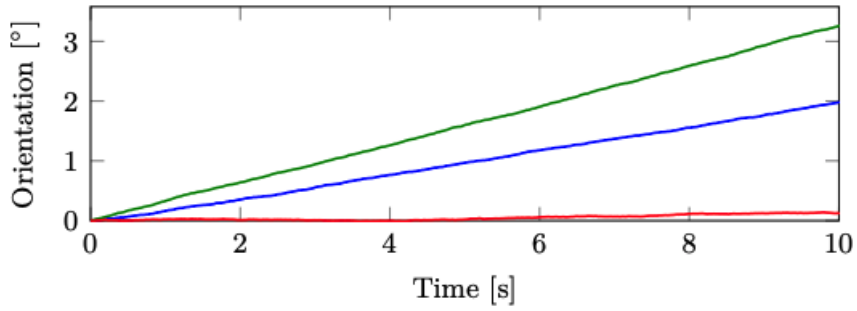
Figure 2.6 illustrates the schematics of how the relative orientation and position are estimated. Integration of the gyroscope measurements provides information about the orientation of the sensor. After subtraction of the earth's gravity, double integration of the accelerometer measurements provides information about the sensor's position. To be able to subtract the earth's gravity, the orientation of the sensor needs to be known. Hence, estimation of the sensor's position and orientation are inherently linked when it comes to inertial sensors. The process of integrating the measurements from inertial sensors to obtain position and orientation information, often called *deadreckoning* (refer to Fig. 2.6).

2.3.2 Sensor Errors

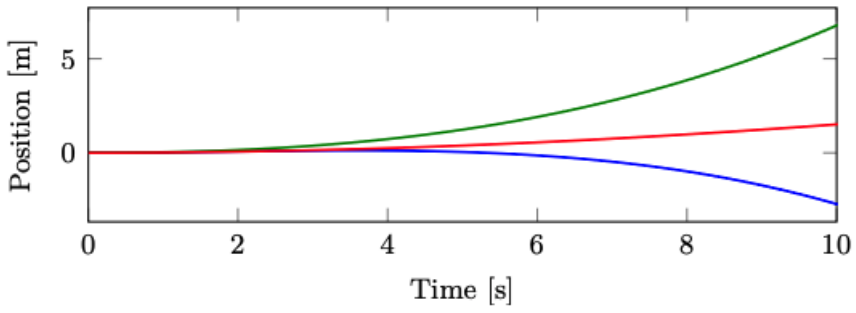
If the initial pose would be known and perfect models for the inertial sensor measurements existed, the process illustrated in Figure 2.6 would lead to 100% accurate pose estimates [147]. In practice, however, the inertial measurements are noisy and biased. Because of this, the integration steps from angular velocity to rotation and from acceleration to position introduce integration drift, illustrated in Fig. 2.7.

Bhardwaj et al. [148] finds that using the accelerometer, gyroscope, and magnetometer alone may produce poor estimations in terms of accuracy or robustness due to various sources of error. Hence, different sensor fusion models are introduced to better track motion [149]. On the other hand, Valldeperes et al. [150] solely uses wireless IMU device system to quantify the sway area of healthy subjects directly calculated using real displacement (position) data.

Errors in the measurements have a large impact on the quality of the estimated position and orientation using inertial sensors only. This is particularly the case for position, which relies both on double integration of the acceleration and on accurate orientation estimates to subtract the earth's gravity. Hence, parametarization of orientation and position need to account for the signal error, bias and the earth's gravity.



(a) Integrated orientation for the position in x- (blue), y- (green) and z-direction (red).



(b) Integrated position for rotation around the x-axis (blue), the y-axis (green) and the z-axis (red).

Figure 2.7: Position and orientation estimates based on dead-reckoning of the stationary inertial sensors [147]: a) the orientation estimates drift over 10s, with different sensor bias in the different axes. b) The position (double integrated acceleration) drifts several meters over 10s. Note: the position drift encompasses signal noise from double integration as well as the *leaked* gravity which needs to be subtracted from the orientation estimates.

2.3.3 Parametrizing Orientation

In this section we introduce different ways of parametrizing orientations. The different parametrizations of the orientation need to be considered due to the nonlinear estimation problems [151, 152], each with its own specific properties. Careful modeling and a careful choice of algorithms can improve the accuracy of the estimates.

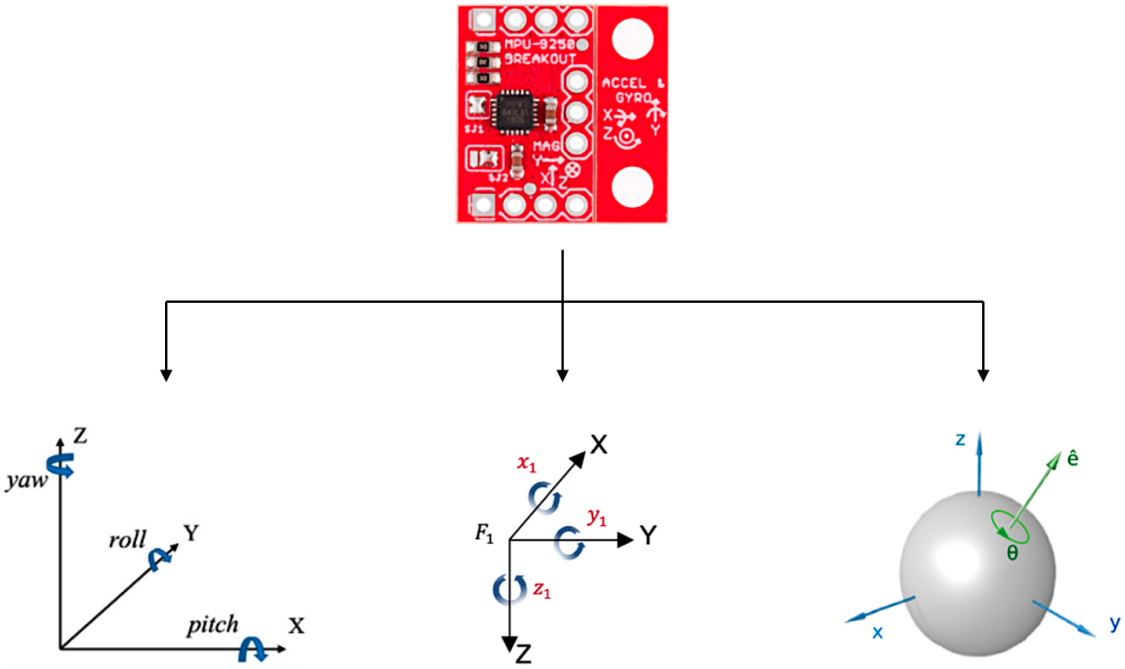


Figure 2.8: IMU module coordinate systems and its conventions. Left to right: Pitch-Roll-Yaw (PRY) Convention, Euler (XYZ) Convention and Unit quaternions (θ, e_x, e_y, e_z).

The parametrizations can describe the same quantity, can be used interchangeably and converted to one another. There are differences which can be found in the number of parameters used in the representation, the singularities and the uniqueness.

Sensor orientation can be defined as a consecutive rotation around three axes in terms of so-called *Euler angles* which can be represented using (X, Y, Z) or Pitch-Roll-Yaw conventions. *Unit quaternions* uniquely describe any three-dimensional rotation about an arbitrary axis and do not suffer from gimbal lock (described in 2.3.3.1). Figure 2.8 illustrates different representations of IMU sensor rotations and their conventions.

2.3.3.1 Euler Angles

We use Euler angle convention (x, y, z) to define rotation. A rotation of ψ radians about the x-axis is defined as

$$R_x(\psi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix}. \tag{2.1}$$

Subsequently, A rotation of an angle θ radians around the y-axis is defined as

$$R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}. \tag{2.2}$$

Finally, a rotation of φ radians about the z-axis is defined as

$$R_z(\varphi) = \begin{bmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.3}$$

The ψ, θ, φ angles are often referred to as pitch, roll, yaw, respectively. A general rotation matrix has the form

$$R = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix}. \tag{2.4}$$

The matrix can be thought of a sequence of 3 rotations, one about each principle axis. The rotation matrix R is expressed as

$$R = R_x(\psi) \cdot R_y(\theta) \cdot R_z(\varphi) \tag{2.5}$$

$$= \begin{bmatrix} 0 & \sin(\psi - \varphi) & \cos(\psi - \varphi) \\ 0 & \cos(\psi - \varphi) & \sin(\psi - \varphi) \\ -1 & 0 & 0 \end{bmatrix}. \tag{2.6}$$

Euler angle representations are not unique descriptions of a rotation for two reasons. First, due to wrapping of the Euler angles, the rotation $(0, 0, 0)$ is for instance equal to $(0, 0, 2\pi k)$ for any integer k. Second, from (2.5), setting $\theta = \frac{\pi}{2}$ leads to only the rotation of $\psi - \varphi$. That way, for example the rotations $(\frac{\pi}{2}, \frac{\pi}{2}, 0)$, $(0, \frac{\pi}{2}, -\frac{\pi}{2})$, $(\pi, \frac{\pi}{2}, \frac{\pi}{2})$ are all three equivalent. This is called *gimbal lock*. To simplify,

gimbal lock is the loss of one degree of freedom, the axes of two of the three gimbals are driven into a parallel configuration.

2.3.3.2 Unit Quaternions

Another commonly used parametrization of orientation is that of unit quaternions. Quaternions were first introduced by [153] and are widely used in orientation estimation algorithms [154, 155] (a full review in sections 2.3.6 and 2.3.7). A unit quaternion use a 4-dimensional representation of the orientation according to

$$q = (q_0 \quad q_1 \quad q_2 \quad q_3)^T = \begin{pmatrix} q_0 \\ q_v \end{pmatrix}, \quad q \in \Re^4, \quad ||q||_2 = 1 \tag{2.7}$$

A unit quaternion is not a unique description of an orientation. If q represents a certain orientation, then $-q$ describes the same orientation. For a vector v, its quaternion representation is given by

$$\overline{v} = \begin{pmatrix} 0 \\ v \end{pmatrix}, \tag{2.8}$$

The rotation of a vector v by a unit quaternion q can be written as

$$q \odot \overline{v} \odot q\,. \tag{2.9}$$

2.3.4 Measurement Models

In the past two sections, we have focused on how orientations can be parameterized. This section discusses probabilistic models for the position and orientation estimation problems.

The tri-axial measurement of the accelerometer are accelerations of each axis, where $a = [a_x, a_y, a_z]^T = [\frac{d^2x}{dt^2}, \frac{d^2y}{dt^2}, \frac{d^2z}{dt^2}]^T$. Gyroscope measurements indicate the angular rates of the sensor when rotated, where $\omega = [\omega_x, \omega_y, \omega_z]^T = [\frac{\varphi^2 x}{dt^2}, \frac{\theta^2 y}{dt^2}, \frac{\psi^2 z}{dt^2}]^T$.

2.3.4.1 Gyroscope Measurement Models

As discussed in section 2.3.1, the gyroscope measures angular velocity ω_t at each time instance t. However, from section 2.3.2, its measurements are corrupted by a slowly time-varying bias $\delta_{\omega,t}$ and noise $e_{\omega,t}$. Hence, the gyroscope measurement model is

$$y_{\omega,t} = \omega_t + \delta_{\omega,t} + e_{\omega,t}\,. \tag{2.10}$$

2.3.4.2 Accelerometer Measurement Models

Recall from Figure 2.6, the accelerometer measures the specific force f_t at each time instance t. The accelerometer measurements are typically assumed to be corrupted by a bias $\delta_{a,t}$ and noise $e_{a,t}$ as

$$y_{a,t} = f_t + \delta_{a,t} + e_{a,t}\,. \tag{2.11}$$

The specific force measure by the accelerometer is given by

$$f = R(a - g)\,, \tag{2.12}$$

where g is Earth's gravity, $g = [0, 0, 9.8]^T$. Since the accelerometer measures both the local gravity vector and the linear acceleration of the sensor, it provides information both about the change in position and about the inclination of the sensor. For orientation estimation, only the information about the inclination is of concern. Since the accelerometer measurements are typically dominated by the gravity vector, a commonly used model assumes the linear acceleration to be approximately zero

$$y_{a,t} = -R_t\, g + \delta_{a,t} + e_{a,t}\,. \tag{2.13}$$

2.3.4.3 Magnetometer Measurement Models

Magnetometers measure the local magnetic field, consisting of both the earth magnetic field and the magnetic field due to the presence of magnetic material. The (local) earth magnetic field is denoted m (illustrated in Fig. 2.9).

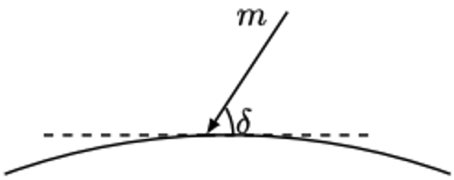


Figure 2.9: Dip angle: part of the earth where the local earth magnetic field m makes an angle δ with the horizontal plane.

Its horizontal component points towards the Earth's magnetic north pole. The ratio between the horizontal and vertical component depends on the location on the earth and can be expressed in terms of the so-called dip angle δ. Orientation can be estimated based on the direction of the magnetic field. Hence, the earth magnetic field can be modelled as

$$m = (\cos\delta \quad 0 \quad \sin\delta)^T \tag{2.14}$$

Assuming that the magnetometer only measures the local magnetic field, the measurement model is

$$y_{m,t} = R_t\, m + e_{m,t}\,, \tag{2.15}$$

where $e_{m,t}$ represents the magnetometer measurement noise and model uncertainty.

2.3.5 Orientation and Position Probabilistic Models

In this section, we describe the modeling choices for the pose estimation problem and for the orientation estimation problem. Prior to the final modelling, the state and the dynamics of the model are selected based on a task. Dynamic models describe how the state changes over time. The basis of dynamic models can be seen in relation of the position p, velocity v and acceleration a, where

$$v = \frac{dp}{dt}, \quad a = \frac{dv}{dt}. \tag{2.16}$$

Depending on the orientation parametrisation, the dynamics of the position and velocity can be expressed in terms of acceleration using Euler discretisation.

Similarly, the dynamics of the orientation can be expressed in terms of unit quaternions or rotation matrices.

2.3.5.1 Pose Estimation

For pose estimation, we model the accelerometer and gyroscope measurements as inputs to the dynamics. Hence, the state vector consists of the position p_t, the velocity v_t and a parametrization of the orientation. We use the inertial measurements in combination with position measurements to estimate the pose. The state space model for pose estimation

$$\begin{pmatrix} p_{t+1} \\ v_{t+1} \\ q_{t+1} \end{pmatrix} = \begin{pmatrix} p_t + T\, v_t + \frac{T^2}{2}\Big(R_t(y_{a,t} - \delta_{a,t}) + g + e_{p,a,t}\Big) \\ v_t + T\,\Big(R_t(y_{a,t} - \delta_{a,t}) + g + e_{v,a,t}\Big) \\ q_t \odot \exp_q\Big(\frac{T^2}{2}(y_{\omega,t} - \delta_{\omega,t}) - e_{\omega,t}\Big) \end{pmatrix} \tag{2.17}$$

The process of noise on the position and velocity states is modelled in terms of the accelerometer noise, with different realization enforced. We assume that the inertial measurement are properly calibrated. Hence, we assume that their biases $\delta_{a,t}$ and $\delta_{\omega,t}$ are zero.

2.3.5.2 Orientation Estimation

For orientation estimation, the state vector only consists of a parametrization of the orientation. The inertial sensors in combination with the magnetometer measurements are used to estimate the orientation. The magnetometer measurements are modelled as in (2.15). We use the model (2.13) where it is assumed that the linear acceleration is approximately zero. This leads to the following state space model for orientation estimation,

$$q_{t+1} = q_t \odot \exp_q\big(\frac{T^2}{2}(y_{\omega,t} - \delta_{\omega,t}) - e_{\omega,t}\big)\,, \tag{2.18a}$$

$$y_{a,t} = -R_t\, g + e_{a,t}\,, \tag{2.18b}$$

$$y_{m,t} = R_t\, m + e_{m,t}\,, \tag{2.18c}$$

where (2.18a) describes the dynamics while (2.18b) and (2.18c) describe the measurement models. We assume that the inertial measurement are properly calibrated with zero bias $\delta_{\omega,t}$.

2.3.6 Inertial Sensor-Based Motion Tracking Models

In almost all applications, the magnetic and inertial measurement units (MIMUs) (comprised of a tri-axial accelerometer, gyroscope, and magnetometer), are used to track the displacement and orientation of a rigid body in real-time [156]. Applications of both inertial trackers (IMUs) and MIMUs are discussed in [157]. MIMU orientation must be first calculated using a sensor fusion algorithm (SFA), also known as attitude and heading reference system (AHRS) [149].

The fundamental problem of inertial sensor fusion has been solved by a large number of previously proposed inertial orientation estimation (IOE) algorithms. Extensive overview of existing methods can be found in [149, 158]. However, using the accelerometer, gyroscope, and magnetometer alone may yield poor estimations in terms of accuracy or robustness due to various sources of error [148]. As suggested by recent literature, the filter parameters play a central role in determining the orientation errors.

The large majority of the published sensor fusion algorithms [159–169] can be grouped in two main classes: Kalman filters (KF) [170] and complementary filters (CF). The orientation estimation problem is seen as the most important pars of the pose estimation problem, since most complexities lie in the parametrization of the orientation and in the nonlinear nature of the orientation. A number of formulations can be highlighted which include different mathematical orientation representations (i.e. quaternion, rotation matrix and Euler angles) (refer to section 2.3.3), different Kalman filter formulations (direct or indirect, linear, extended, unscented, etc.), and different strategies to fuse the signal information (algebraic or optimization) [166]. Despite the large number of algorithms proposed for sensor fusion [158, 159, 163–166, 171–174] the results are contradictory, sometimes inconclusive and lack accuracy [158].

Several optimisation methods and models used to estimate position and orientation include 'smoothers' [175–180] and Kalman filters mentioned above [175, 181] and their extensions [182]. Smoothers help compensate the errors which may occur during the initial motion estimation and Kalman filters are used for estimation of

unknown variables based on the measurements observed over time. Optimisation methods are closely related to Kalman smoothers [183, 184].

For pose estimation, inertial sensors are often combined with measurements from cameras [185–188]. For orientation estimation, they are often used in combination with magnetometers, which measure the direction of the magnetic field [160, 189].

Sun et al. [181] used Kalman filters to build a simple distance estimation algorithm using inertial sensors and a mono camera. Next, Sun et al. [175] implemented an off-line smoother to improve foot motion estimation, making it a quadratic optimization problem. First, Kalman filter was used to obtain the initial foot position and the signal error which occured during the estimation was compensated with a smoother. Duong et al. [176] proposed a suboptimal smoothing algorithm that had a lower computational burden compared to [175] for attitude, velocity and position estimation. Indeman [177] proposed a non-linear optimization for processing IMU input using factor graph-based incremental smoothing. The model automatically determines the number of states to recompute at each step acting as an adaptive information fusion. The method was compared against a full non-linear batch optimization and to a conventional extended Kalman filter (EKF).

2.3.7 IMU-Assisted Ultrasound Probe Motion Tracking

There are a number of methods that visually-assist sonographers during an ultrasound examination by guiding ultrasound probe movement [146, 190–194]. Droste et al. [190] proposed a multi-modal US-GuideNet which receives US video and motion signal provided by IMU attached to an US probe. The model does not utilize IMU for translation, it guides sonographers to standard planes by predicting the rotation towards the standard plane position from a current plane and the general hand movement (probe rotation) that an expert operator might perform. Grimwood et al. [5] proposed a a probe movement guidance framework posed as a high-level command classification problem for prostate external beam radiotherapy. Zhao et al. [191] proposed a virtual fetal model to provide global position visualization

cues to the operator during US scanning by automatically retrieving anatomical landmarks using US image with probe poses.

Pagoulatos et al. [195] proposed a fast calibration method for 3D tracking; authors attached a DC magnetic sensor on a 2D phased-array probe to scan a custom-built phantom. To calibrate irregular movements of the probe during scanning, Kin et al. [194] used three gimbal assisted ultrasonic distance sensors to extract relative position data and IMU module for direction tracking. Cai et al. [146] developed IMU-assisted ultrasonic tracking system that enables accurate US probe localization. Another technology which uses IMU is Micro Electro-Mechanical Systems (MEMS) which make sit possible to enclose accelerometers and gyroscopes into a small chip. Rahni et al. [196] attempts to avoid the orientation drift from the accelerator by housing a sensor into a cube of known size and rolling it across a plane; that way the calculation of translation from change in orientation is more accurate.

2.4 3D Ultrasound Volume Reconstruction

In general, human beings are good at approximating the size and rough geometry of 3D objects as well as predicting how an object would look from another view point based on the prior knowledge. Previously seen views of an object enable us to develop mental models of what objects look like in 3D. With little prior knowledge and expertise in fetal ultrasound scanning, novice sonographers may struggle to mentally reconstruct and analyse a fetal womb from just 2D ultrasound scans. Since each 2D US scan only represents a cross-sectional view of the three-dimensional (3D) scanned structure, it requires significant mental workload to convert 2D images into a 3D mental representation.

The main advantages of 3D US vs 2D US are summarized in surveys [197–205]. 2D ultrasound does not allow to completly capture a region of interest (ROI) and without an automatic frame of reference, the orientation of an object is generally unknown; the operator may manually enter a label on each saved image or cine loop to aid later interpretation, however, such labeling is time-consuming, inconsistent, and may be inaccurate.

While many protocols and conventions for 2D ultrasound sonography have been established (e.g., standard views, selective capture of 2D image frames required to make a diagnosis, and image landmarks to aid measurements and localize probe position and orientation) [190, 206–208], there remains a significant variability in the quality and diagnostic utility of images acquired by different users.

Freehand 3D ultrasound reconstruction has emerged as a promising approach to overcoming these limitations. This method can be divided into sensorless (image-based) and sensor-based approaches. Sensorless techniques rely on intrinsic image features for position estimation [209, 210], while sensor-based techniques utilize external tracking devices to record the orientation and position of 2D scans [197]. Although sensorless methods offer flexibility, they may be less accurate due to variability in scanning conditions. Conversely, sensor-based methods provide more reliable positional data but require careful synchronization and calibration [211]. Further discussion can be found in section 2.4.3.

The principles of sensor-based freehand ultrasound reconstruction are integral to the approach proposed in Chapter 6, which incorporates IMU-assisted probe motion tracking to precisely capture orientation and position data. By synchronizing this data with real-time ultrasound frames, the research aims to achieve a 3D reconstruction of fetal anatomy, ultimately reducing the cognitive load on sonographers and enhancing scanning efficiency (see section 6.1.1).

In medical imaging it is important for physicians to visualize the anatomy of the patient for a correct diagnosis and analysis purposes. It is a difficult task to interpret anatomical features of a fetus whilst navigating between biometry planes. Consequently, sonographer skills may vary which leads to a non-uniform skill-set and inter-operator variability which in turn may affect the accountability of the diagnosis [212–216].

A visualisation of a 3D ultrasound volume can enhance the understanding of physicians of a scanned ROI without spending too much on mental workload. A 3D ultrasound volume visualization can be achieved via a 3D ultrasound reconstruction

process, which is the generation of 3D ultrasound volume from a series of 2D ultrasound images [217].

2.4.1 3D Ultrasound Imaging Systems

There are several stages required for successful reconstruction of a 3D ultrasound volume of a fetus. The 3D ultrasound reconstruction pipeline includes the data acquisition, data processing, reconstruction method and 3D visualisation/rendering. The data acquisition depends on the type of US system used in a clinical setting and the type of examination the clinicians are trying to perform; A 3D US system may include 2D array scanning, mechanical scanning, or position tracking-based freehand and untracked-based (sensorless) freehand scanning (see Fig. 2.10). The data collected are generally comprised of 2D ultrasound images and their relative spatial information.

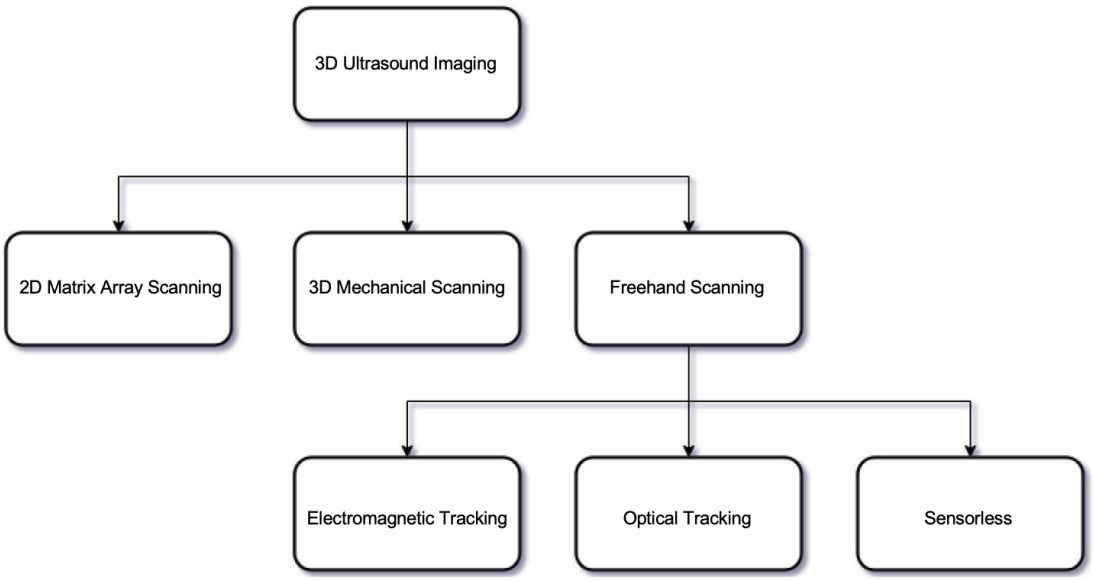


Figure 2.10: Types of 3D ultrasound imaging [217].

3D ultrasound is less operator-dependent as the recorded 2D images can be re-evaluated and re-processed at a later date using the generated volume rendering, whilst 2D ultrasound only acquires 2D images [218]. Moreover, 3D US images allow to further analyse morphology and dimension which helps in identification of

characteristic patterns; such systems allow for orientation-independent visualization and can provide a measurement of quantitative attributes in 3D [219–222].

Despite the convenience and benefits of 3D ultrasound, 3D systems are much more expensive than traditional ultrasounds [223]. As 3D reconstruction depends on 2D image quality, a surface rendered image is only as good as the quality of the original 2D data. The probe and the fetus must remain stationary during image acquisition to avoid movement artifacts. Artifacts that simulate pathology are common in 3D, e.g. boundaries can lead to apparent defects such as a hole in the head or missing limbs [224].

Figure 2.10 underlines the types of 3D US systems designed and implemented to obtain 3D-images from which volumetric images are reconstructed. The reconstruction is the technique used to interpolate classical 2D slices to 3D-like structures. In most of the cases, 3D ultrasound is acquired by sweeping a 2D probe over the area of interest where the resulted 2d B-scans are 'stacked up' to form a 3D volume. The US systems differ in how the motion of a probe is constrained and how the position of each B-scan is determined [201]. The most used types of acquisition protocols in US setting are described below.

2.4.2 3D Ultrasound Acquisition Protocols

Transducers are commonly categorized as "matrix" array probes (with electronic beam steering in elevation and azimuth), mechanical "wobbler" probes (with internal motorized translation of a 1D array), and freehand acquisition probes [199, 225] (see Fig. 2.11). Transducers with *2D phased arrays* electronically steer the ultrasound beam across a pyramidal volume and the received echoes are processed to integrate 3D ultrasound images in real time [226–230]. Hence, probes with 2D phased arrays acquire 3D information via electronic scanning which happens almost instantaneously. *Mechanical 3D probes* contain a slowly rotating mechanism to sweep a predefined ROI where multiple 2D images are acquired when the motor is activated [199]; the probes use a conventional linear array transducer which is motored to rotate, tilt, or translate within the probe [229, 231]. Finally, in

a *freehand scanning* the probe is moved by hand in an arbitrary manner, and the resulting sequence of B-scans is located in 3D space by either intrinsic (image-based) or extrinsic (position-sensing) means [232]. This is the only technique that gives the clinician complete freedom to guide the probe along the path of the anatomy.

The data for 3D scanning can be acquired using phantoms [195, 211, 233–236], in-vitro (in the laboratory) [237–239] or in-vivo (in the body) [240–248].

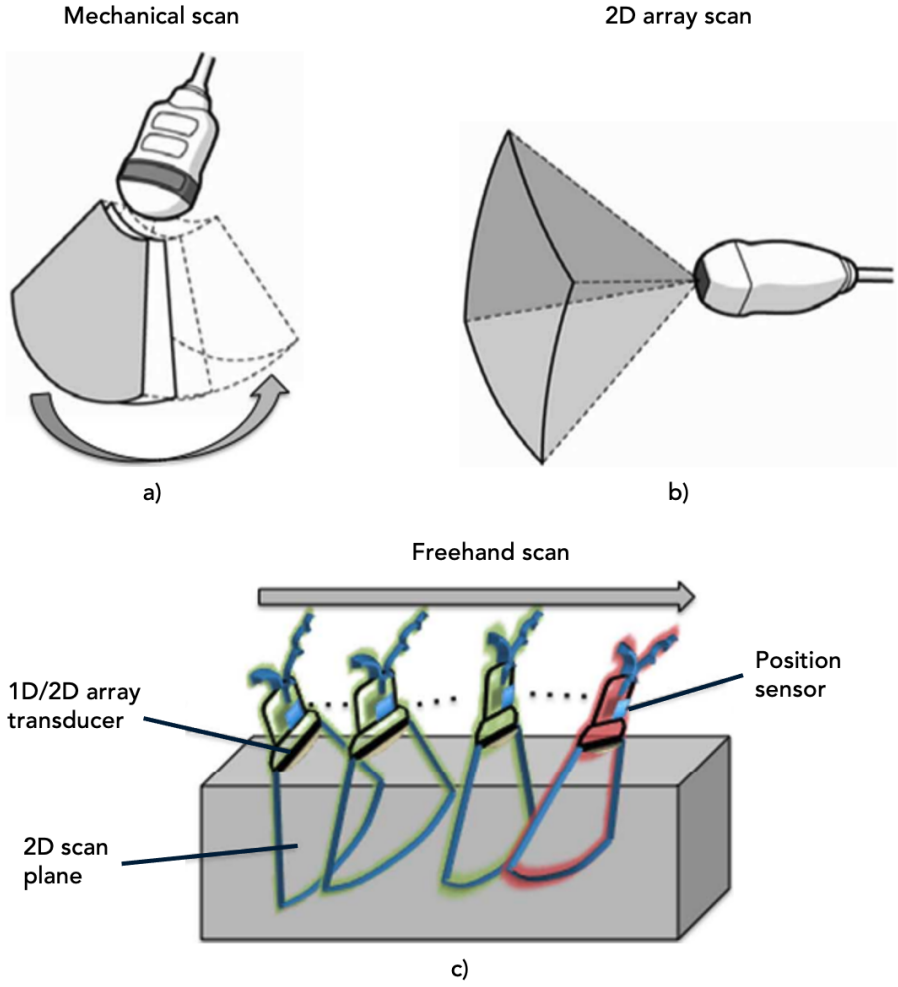


Figure 2.11: Types of transducers used for 3D US acquisition: a) mechanical 3D probes, b) 2D matrix-array transducers, and c) freehand 3D acquisition using a conventional 1D/2D array probes with position sensor [249].

Both matrix arrays and 3D mechanical transducers allow volumetric acquisition but with different acquisition speed and using different coordinate systems [201]. Despite the benefits in penetration depth (higher probe frequency), the systemic load or the electrical connection of 2D phased array probes is higher than 1D array

probes, making it harder to process the elements of an ultrasound beam [250]. Mechanical probes provide 3D volume data in near real-time, however, they are expensive and have limitation on scanning large volume organs, i.e. limited field of view (matrix array) [251]. The freehand method has advantages over use of the transducers mentioned above, especially in the application of ultrasound-guided interventions, as it is low in cost, can provide inherent flexibility, large field of view (FOV) and high in-plane resolution [197, 252–254]. Section 2.4.3 further investigates sensorless and tracked freehand 3D reconstruction methods.

2.4.3 Freehand Scanning Methods: Image-Based & Sensor-Based

Freehand 3D ultrasound reconstruction can be divided into sensorless and tracked freehand scanning protocols (see Fig. 2.12). Sensorless scanning or image-based sensing approaches do not use any tracking devices to record orientation and position. Instead such intrinsic approaches extract the relative positions by analyzing the image features instead of depending on position sensors [209]. Position-sensing methods require integration of ultrasound imaging with real-time tracking of probe pose which records position and orientation of the 2D B-scans.

In most of the freehand scanning methods (incl. sensorless), the 3D reconstruction and quantitative/qualitative analysis is performed using the ground truth in a form of a tracking device with position and orientation data [255–257]. Hence, in both cases, the image data and its corresponding locations need to be be synchronised using a temporal calibration and a filtering step, in which redundant data are removed [211]. Additionally, the tracking sensor specifies only the position and orientation of the probe itself rather than each pixel's location, therefore, spatial calibration is needed prior to translation of each pixel in a B-scan to the corresponding voxel in the 3D volume [197]. Depending on the data used and a reconstruction task, other methods of calibration should be used (see Fig. 2.12 to the left of *Acquisition*). However, this exceeds the scope of this review.

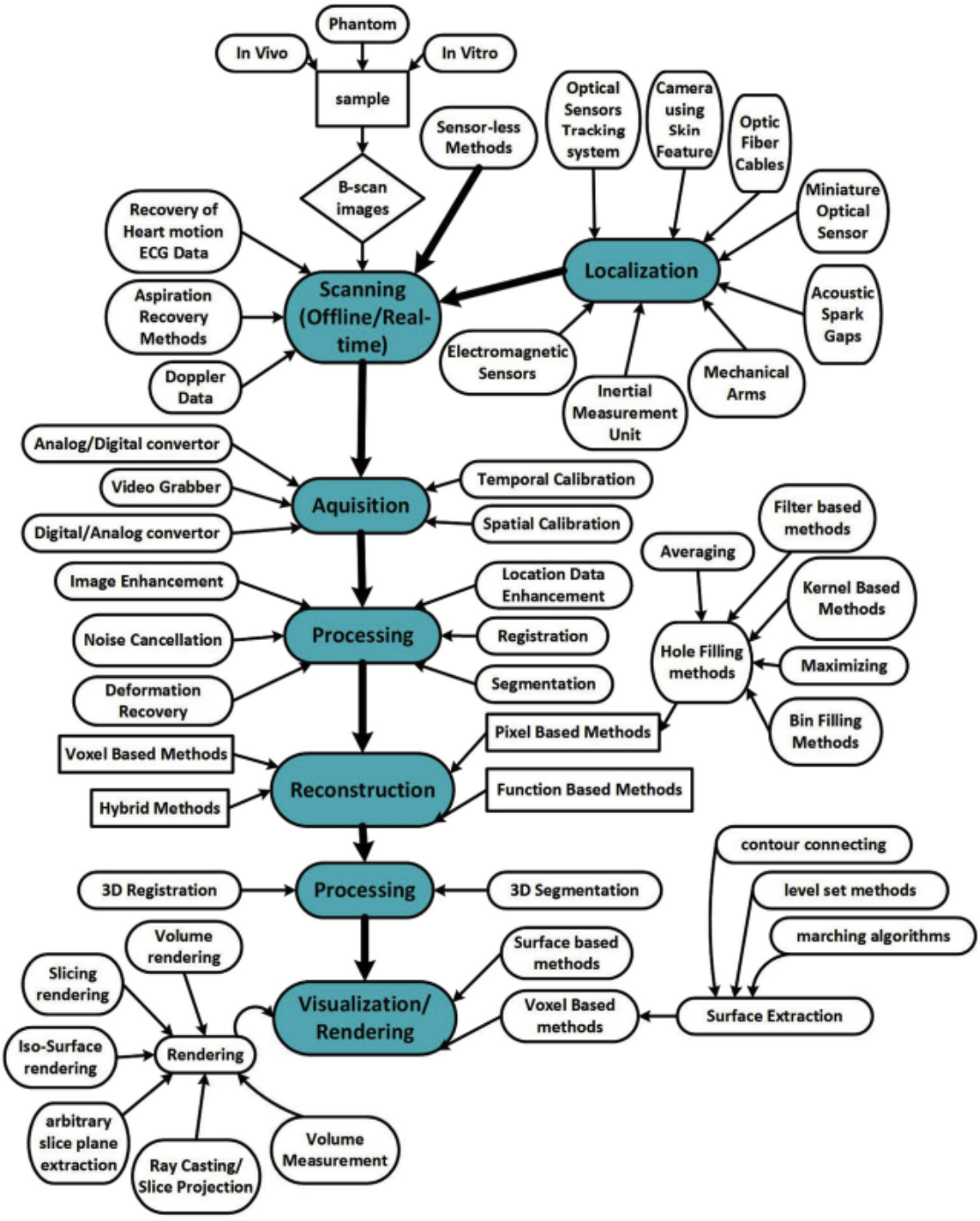


Figure 2.12: Overview of freehand tracked and sensorless 3D ultrasound reconstruction systems [217].

In a freehand 3D ultrasound scanning, the data is sparse and irregularly sampled which makes sensorless and tracked 3D reconstruction a difficult task [258]. Therefore, the reconstruction methods or algorithms are researched and developed in order to solve the stated problem.

To use a sensorless scanning protocol, an operator moves the transducer at a constant velocity in a linear or rotational manner to guarantee appropriate intervals. Such an approach may lack accuracy and can be difficult to implement [231]. Hence, researchers attempt to tackle the problem using a number of techniques discussed next.

2.4.3.1 Image-Based Freehand 3D Reconstruction

Sensorless reconstruction methods formulate 3D reconstruction as a recognition problem. Based on the increasing availability of large training data sets, image-based methods attempt to recover the 3D geometry and structure of objects from one or multiple RGB images without the complex camera calibration process [259]. There are a number of deep learning-based 3D reconstruction networks which are usually represented by an encoder-decoder architecture, i.e. Variational Auto-Encoders as well as CNNs [192, 260, 261], LSTMs [255], GANs and others. Some image-based algorithms estimated 3D geometry of an object from images or decompose the reconstruction into sequential steps where each step predicts an intermediate representation. Other methods include surface-based representations such as meshes [262] and point clouds [263] . Volumetric representation techniques are adopted to parameterise 3D shapes using regular voxel grids. That way, 2D convolutions used in image analysis can be easily extended to 3D. The speckle decorrelation techniques estimate the relative elevational distance of two neighboring frames, i.e. the relative difference of position and orientation between neighboring US images to the correlation of their speckle pattern [210, 264, 265].

Recent methods show a possibility of 3D US volume reconstruction using no external location-tracking device. In these techniques, the relationship between the statistical information on the speckles in different images, such as linear regression, frame distance estimation and decorrelation, is used for localization of the images [197]. Image-based methods are innovative and show a promising alternative for the future of 3D US reconstruction, although if relied solely on methods like speckle decorrelation, models render unreliable results [266, 267].

From section 2.4.3, most of the sensorless methods [192, 253, 260, 268, 269] require numerical comparison (labels) of a model created against the ground truth represented by a motion-tracking sensor. In other cases, the ground truth is represented by a 3D volume data which is then sliced into 2D frames such that the interval between slices represent a freehand scanning.

Prevost et al. [260] first used a convolutional neural network (CNN) to estimate the relative inter-frame motion; two consequitive frames and a generated optical flow field between them was stacked as an input into a CNN to estimate relative rotations and translations between the framees. The work was extended [270] to estimate the trajectory (transducer's motion path) using IMU orientation coordinates of the probe, similar to Housden et al. [271] who used an IMU to estimate the orientation of an US probe.

Guo et al. [255] proposed a deep contextual learning network (DCL-Net) to extract the correlation information of US video clips using sequences of 2D frames and attention mechanism for speckle-rich image regions to improve on [260] out-of-plane motion analysis. A transrectal (TRUS) probe was connected to an electromagnetic (EM) tracking device that moves the probe in a linear, rotational or tilting path near the ROI (prostate). In each scan the position and orientation are known and are used for 3D reconstriction. An EM device serves as the GT label in the training phase. The authors improved on previous works [255, 272] by integrating a contrastive feature learning method (make similar data points closer and dissimilar data points farther apart in a learned feature space) to estimate the spatial movement and self-focused attention module for US speckle information extraction [273].

Yeung et al. [192] proposed a CNN model that predicts the 3D position of 2D US frames in a 3D fetal brain volume; 2D slices are sampled from aligned 3D volumes and augmented to train as part of a self-supervised model where 2D frames and their 3D volume positional coordinates are used as the ground truth. Next, Yeung et al. [256] proposed sensorless (in training and inference phase) 3D reconstruction pipeline based on deep implicit representation, tested on volume-sliced and 2D US images.

Luo et al. [274] designed an online learning framework (OLF) that improves reconstruction performance by utilizing consistency constraints and shape priors. The authors sliced 3D volume in different ways (representing rotation, fast and slow scan etc) in attempt to replicate freehand scanning and used the 2D slices as the ground truth. Next, to reduce drift errors whilst estimating elevational displacement between images, Luo et al. [257] integrated images and used IMU sensor with motion data as weak labels in a deep motion network.

2.4.3.2 Sensor-Based Freehand 3D Reconstruction

In biomedical US imaging, object tracking technologies are key to locate US probes and other medical tools for precise operation and intuitive visualization [275, 276]. With the recent advancement of position tracking technology, the tracked freehand ultrasound scanning method has improved in terms of imaging quality, accuracy, effectiveness, portability, and reliability [217]. A position sensor is attached to a 1D/2D transducer (refer to Fig. 2.11c), and image and sensor data are acquired simultaneously. The 3D US volumes are then visualized following geometric registration of arbitrary 2D slices into 3D Cartesian coordinates (or similar) and reconstruction.

Position sensors for 3D US reconstruction

The tracking system can consist of one of the following or a combination [197, 277, 278] (Figure 2.13 depicts some of the position sensor examples): magnetic or EM sensors [238, 279–283] , acoustic (spark gaps) [284–286] or optical sensors [287, 288], mechanical arms [289], fiber-optic cables, inertial measurement units (IMUs), skin feature tracking using cameras or laser sensors and any other systems that can track the US transducer position and orientation accurately with six degrees of freedom (6 DOF).

The conventional and most popular methods of obtaining position data in a freehand US scanning system include magnetic and optical tracking systems, followed by acoustic, articulated-arm and inertial tracking systems. Huang et al. [231] and Peng et al. [144] summarized sensor-based freehand 3D US imaging systems.

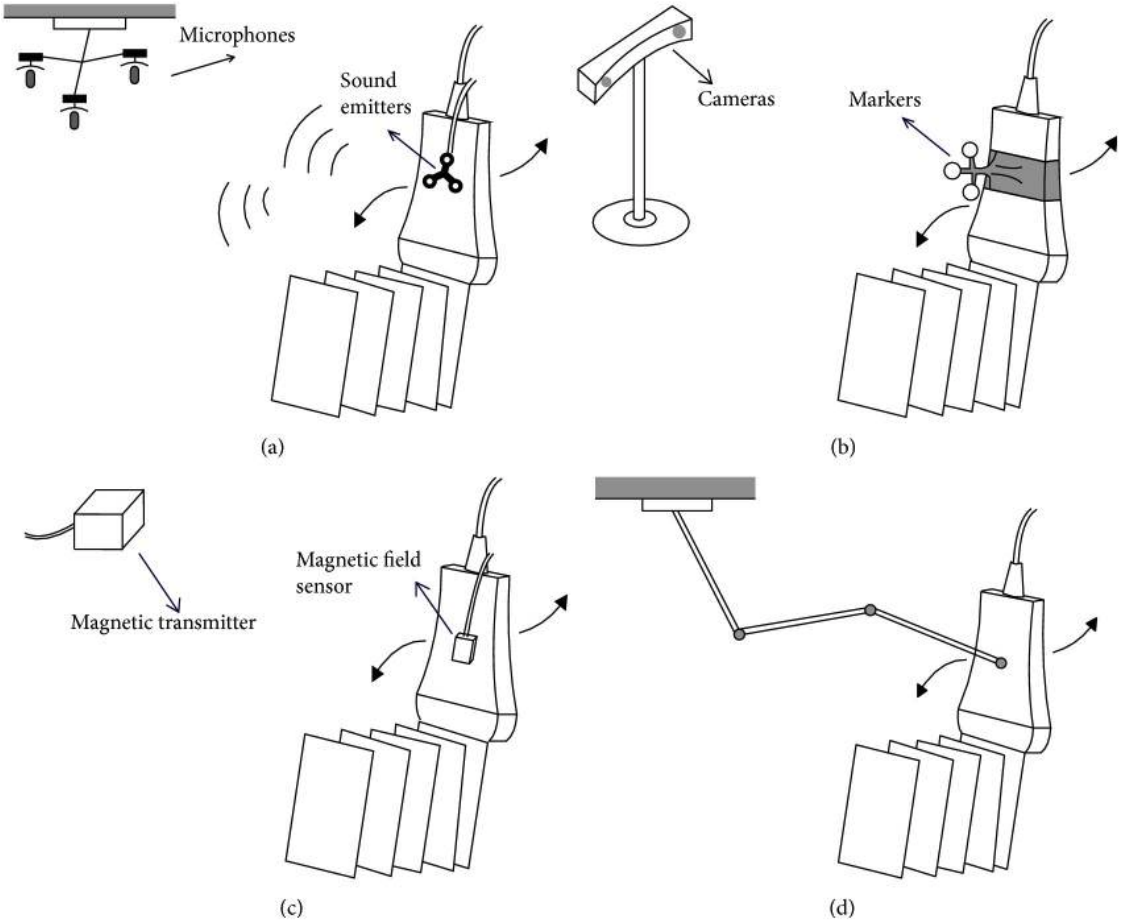


Figure 2.13: Schematic structure examples of position sensors: a) acoustic sensor, b) optical sensor, c) magnetic field sensor, and d) mechanical arm sensor [197]. Note: linear array transducer is used as an example.

Prager et al. [211] implemented real-time volume measurement and visualization using a freehand US system with a magnetic sensor; authors were able to reslice a 3D volume at 10Hz using optimized sequential algorithms. Dai et al. [290] used a concave probe and an EM sensor to semi-automatically determine the reconstruction volume and assign an optimal viewing direction. Yu et al. [205] used a phantom for cardiac image analyis and estimated the optimal segmentation parameters using an electromagnetic position and orientation measurement device (EPOM). Daoud et al. [253] and Chen et al. [291] developed a freehand 3D US imaging system using an EM sensor and a linear probe; authors [291] obtained volume reconstruction and visualization during data acquisition at real-time level.

Welch et al. [292] proposed a real-time freehand 3D US system for image-guided

surgery using a linear probe and an optical sensor to render 3D volumes. Chung et al. [269] used optical motion tracking to detect a carotid artery contour for carotid atherosclerosis diagnosis. The 3D motion tracking system consisted of 8 Eagle digital CCD cameras for motion detection in 3D space and 4 passive fluorescent markers attached to an US probe, showing spatial and temporal resolutions.

Herickhoff et al. [293] used a 3D matrix-array probe and orientation readings of IMU from a sweep about a single axis to perform a low cost voxel-based US 3D volume reconstruction (under USD 250). Only a single degree-of-freedom pivot sweep of the probe was performed. The work was extended and applied to a phantom [294].

Despite the usefulness of each tracking device, EM tracking accuracy [269] and 3D reconstruction (geometric distortion) [199] suffer from the interference of magnetic signals, i.e working next to metal instruments and power cables [217]. The cameras of optical based sensors should not be obstructed [295] which is difficult to use in a clinical setting, as well as the markers that are mounted on a probe may restrict movement [269]. Limitations of acoustic and mechanical arm sensors are trivial. Hence, the magnetic trackers and optical based sensors are not used in this thesis. The IMU positional tracking suffers from drifting [147] due to the low signal-to-noise ratio and the required double integration. However, The drift and errors can be resolved with sensor fusion. In Chapter 6, we further explore the use of IMU in ultrasound imaging for 3D fetal reconstruction.

2.4.4 3D Ultrasound Volume Reconstruction Algorithms

The selection of a correct volume reconstruction method is based on the goal of reconstruction which defines the preference for the final visualisation, the computational capabilities and the data available. First, the coordinate system and volume grid of a future reconstruction are established, including volume size, axes of volume, origin of axes, and the size of a voxel [296]. Next, to analyze the sequences of B-scans, two types of approaches can need to be considered: the reslicing (without reconstruction) or the true 3D reconstruction which includes interpolation step [297]. Methods producing various slices or approximating surfaces

based on the irregularly distributed B-scans have been investigated by several researchers [211, 251, 283, 298–302].

The sequence of B-scans can be arbitrarily resliced and distance/volume measurements are performed without reconstruction. Prager et al. [211] created a Stradx system that deal solely with raw 3D data (B-scan images and positions) without constructing voxel arrays. The worked enabled the analysis of the data without reconstruction which can be powerful for manual analysis of 3D datasets. However, 3D isotropic reconstruction is still necessary in the clinical context when automatic segmentation or registration procedures are required.

The second approach is based on the interpolation of the information within the B-scans. The approach has been analysed in several research papers [195, 205, 254, 258, 297, 302, 303] and surveys [197, 259, 298, 304, 305]. Based on the acquired spatial information from the tracking system, interpolated data is used to fill a regular 3D lattice and thus create a volumetric reconstruction. Due to irregularly spaced 2D B-scans, a conversion of 3D data into a 3D volume grid is computationally expensive and may lead to poor quality, i.e. image distortion and introduction of geometrical artifacts during reconstruction. To avoid damaging or losing the underlying shape of the data, Rohling et al. [298] reviewed a number of often-used methods for resampling B-scan data into 3D image data. There are volume reconstruction methods that can be grouped into three categories: pixel-based method (PBM), voxel-based method (VBM), and also function-based method (FBM).

2.4.4.1 Pixel-Based Methods

The PBM category includes pixel nearest-neighbour (PNN) interpolation [306] where the pixels from the acquired B-scans are traversed and the pixel intensity value is assigned to its nearest voxel. Multiple contributions to the same voxel are usually averaged. The basic interpolatioin algorithm used to fill gaps in the voxel arrays consists of two stages: bin-filling stage and hole-filling stage [307]. Despite the effective methods of interpolation, the gaps in the reconstructed volume data may still occur due to the sampled B-scan images being too far

apart. Sethian [308] introduces a fast marching method (FMM) to interpolate empty voxels in the hole-filling stage.

2.4.4.2 Volume-Based Methods

The VBM category includes voxel nearest-neighbour (VNN) interpolation [211, 309], distance-weighted (DW) interpolation [288, 300] recent methods such as kernel regression [258], median-filter reconstruction [302]. For the VNN method, the intensity of the nearest pixel in a voxel neighbourhood is assigned to this voxel. VNN tends to produce reconstruction errors on the image slices, particularly when the scanning has big gaps. Unlike the VNN and PNN methods, the DW interpolation computes each voxel value by assigning the weighted average of a set of pixels falling into a predefined 3D region centred about that voxel [298]. Finally, the average value of those pixels is placed on the voxel. The DW method is able to suppress speckle noise [254], whilst smoothing the boundaries of 3D reconstructed volume, causing the loss of some information on the original 2D ultrasound frames [258]. The implementation of kernel regression can also help estimate the whole voxels in a volume [258]. Huang et al. [302] introduced a median-filter-based reconstruction method to improve the quality of volume reconstruction by utilizing median filter to reduce speckle noise. Authors continued the research [251, 310–312] by improved DW algorithm and introduced using methods like adaptive Gaussian distance weighted (AGDW) and a square distance-weighted (SDW) which helps improve the quality of reconstructed image.

2.4.4.3 Feature-Based Methods

The FBM category attempt to introduce functional interpolation for 3D US reconstruction. Once a function is selected (i.e a a polynomial), the method determines the its coefficients utilizing the pixel values and relative positions. Next, the functions are evaluated at regular intervals to produce the voxel array. Rohling et al. [298] proposed the Radial Basis Function (RBF) that used an estimate function to compute a spline that passes through the pixels that form a shape in the 2D ultrasound frames.

2.4.5 3D Volume Visualization

After volume reconstruction, the 3D visualization method is used to display the volume data from the volume grid for the sonographers to see the result of ultrasound scanning. The choice of the visualisation method depends on the clinical task and the problem that needs to be solved using the reconstruction. The visualization of the scanned anatomy can be used to analyse the acquired scan data and assist in diagnosis, i.e. in image-guided surgery or to ease sonographer scanning. In many applications, visualization of the ROI surface is important to measure organ volume or any quantitative properties. For example, in obstetrics and gynecology, surface rendering of the fetus and its organs from the US volume is beneficial for the assessment of many prenatal diseases and antenatal treatments [199].

The 3D visualization process is also the last step to complete the fully functioning 3D ultrasound reconstruction [217]. The common rendering algorithms for 3D visualization are multiplanar reformatting, volume rendering, and surface rendering [313].

2.4.5.1 Multiplanar Reformatting

The multiplanar reformatting method is a visualization technique where 2D ultrasound planes, also known as resliced 2D images, are extracted from the 3D ultrasound data and displayed to the user with 3D impressions [314]. There are three approaches to display multiplanar reformatted rendering, which are the planar cross-sectional images, the cube view, and the orthogonal planes [199].

Figure 2.14 illustrates planar cross-sectioinal images and cube view rendering. The planar cross-sectioinal result can be viewed on three orthogonal slice views, which are represented by traverse, coronal, and sagittal planes [316]. The limitation of planar viewing is a loss of information due to the complex shape of ROI, especially when viewing spinal curvature (see Fig. 2.15). Therefore, the non-planar volume rendering method can compensate for such limitation [315]. The multiplanar reformatting method has been implemented by many researchers due to its simplicity and low computational requirement; for example, Prager et al. [211] used Gouraud technique to render slices of the scanned volume.

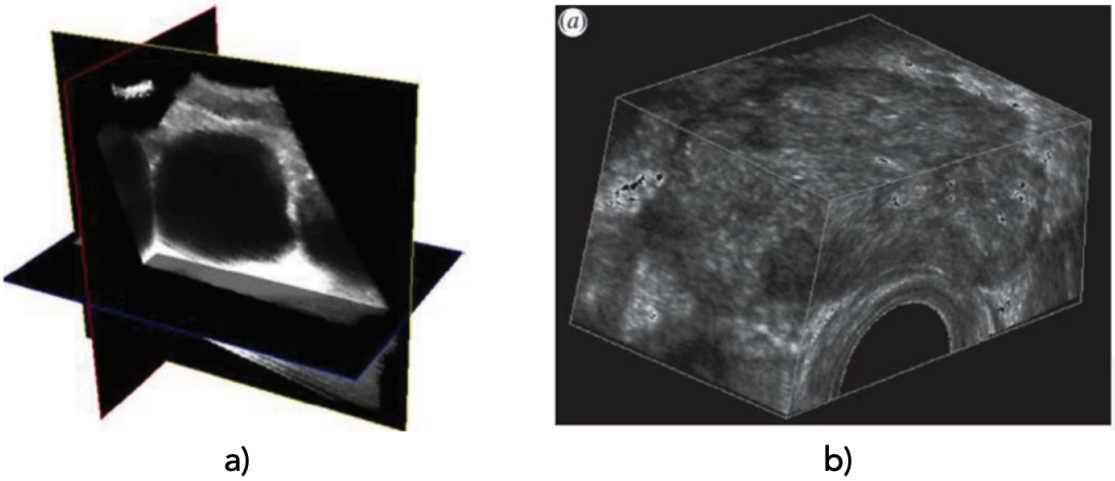


Figure 2.14: Multiplanar reformatting: a) Planar cross-sectional images of reconstructed volume data [254] and b) cube view of reconstructed volume data [199].

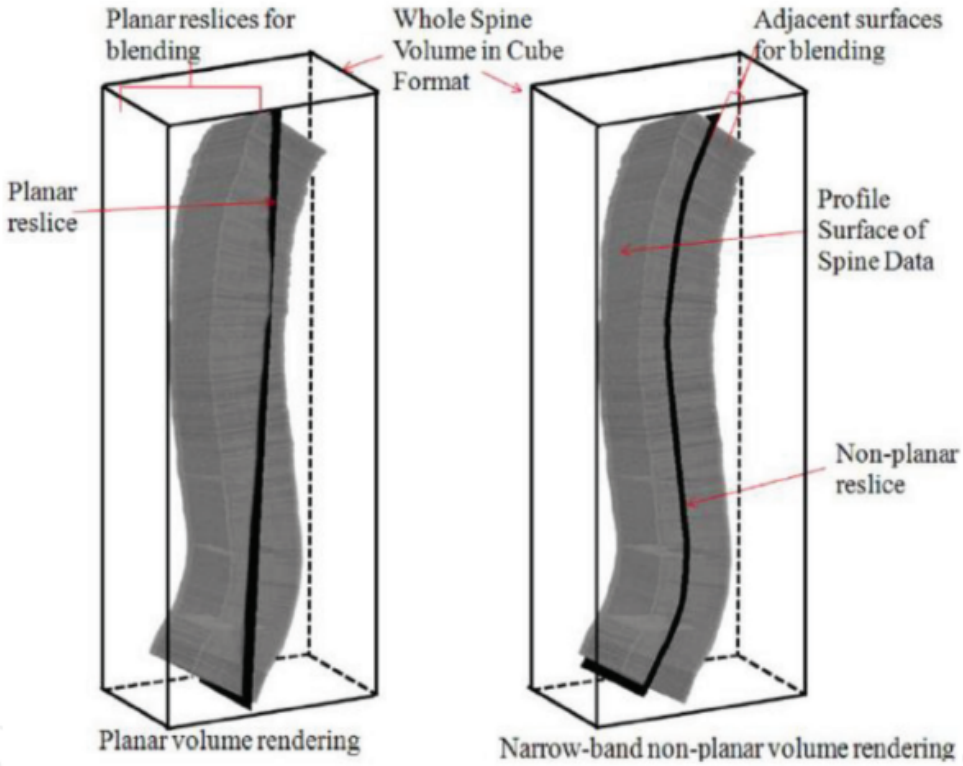


Figure 2.15: The difference of planar and nonplanar volume rendering in the assessment of scoliosis [315].

2.4.5.2 Volume Rendering

Volume rendering involves a ray-casting or ray-marching technique where the change of light that went through the 3D volume data is projected as the output

visualization results for the operator to view [252]. The light absorption principle [317] is implemented in the volume rendering technique where every voxel has the attributes such as brightness, transparency, and color [314]. Figure 2.16 illustrates ray-casting technique in volume rendering.

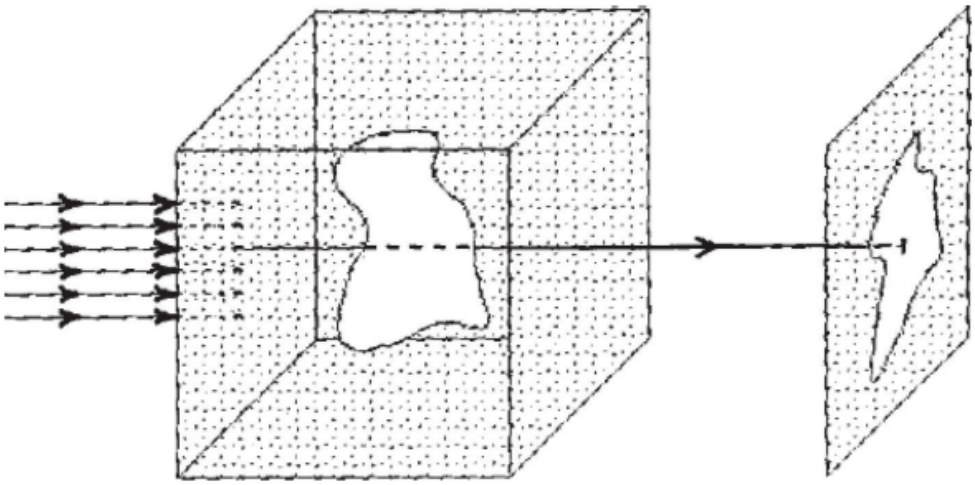


Figure 2.16: The volume rendering technique involves several rays passing through 3D volume data. The synthesis methods can be applied to each voxel value that the ray passed to produce specific effects, such as transparency and maximum intensity projection of certain objects [252], i.e. tissues, blood vessels and others.

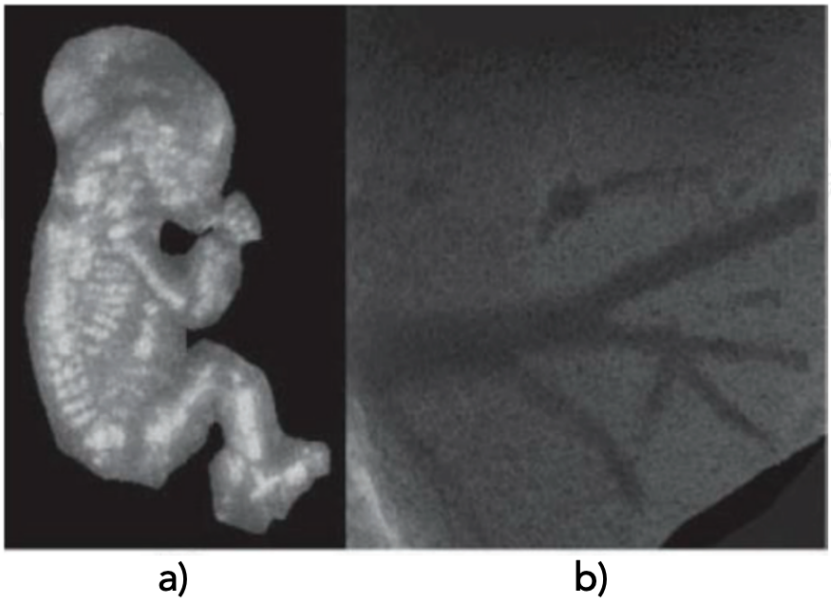


Figure 2.17: Different volume visualisation techniques for 3D ultrasound imaging: a) the maximum intensity projection of a fetus and b) the minimum intensity projection of blood vessels in the liver [314].

There are several approaches used for the volume rendering visualization which include maximum intensity projection [318] and translucency rendering [199]. Translucency rendering allows physicians to freely choose the opacity values to highlight or dim particular features of the anatomy, which improves the diagnostic accuracy [231]. Figure 2.17 shows the example of volume rendering that uses maximum and minimum intensity projection.

The volume rendering preserves all the 3D information, making it the most common technique for 3D display [316]. Furthermore, it can distinguish between tissue and fluids very well, and hence, it is suitable to view 3D ultrasound fetal image [199, 252]. However, the volume rendering is CPU-intensive and is not suitable to view the soft tissues details [199].

2.4.5.3 Surface Rendering

The surface rendering produces a 3D surface based on the segmented boundary data points by generating the surface triangles or polygons associated with standard surface-rendering interpolation [300]. The surface rendering can help visualise tissue surfaces, improve the interpretation of data sets [234] by showcasing the shapes of 3D objects making the topography and 3D geometry easier to visualise [319]. It can be used for ultrasound image analysis as a guidance and visualisation tool for novice sonographers to navigate around the fetal womb.

The surface rendering technique can be classified into indirect and direct surface rendering. In the direct surface rendering is a technique that renders a 3D model from medical images without intermediate geometric representations [320], setting thresholds or using object labels to define a range of voxel intensities to be viewed [321]. The transparency and colors are used for the better 3D visualization of volumes [322].

Indirect surface rendering is a technique that generates intermediate 3D model representations from HRCT and MRI scans. The technique requires the segmentation or classification steps executed prior to rendering itself, to identify the surfaces of relevant structure boundaries within the volume [321]. Some features of the data

can be lost during the segmentation, making the method sensitive to noisy data. Figure 2.18 shows the 3D visualization using indirect surface rendering (on the left) and direct surface rendering technique (on the right).

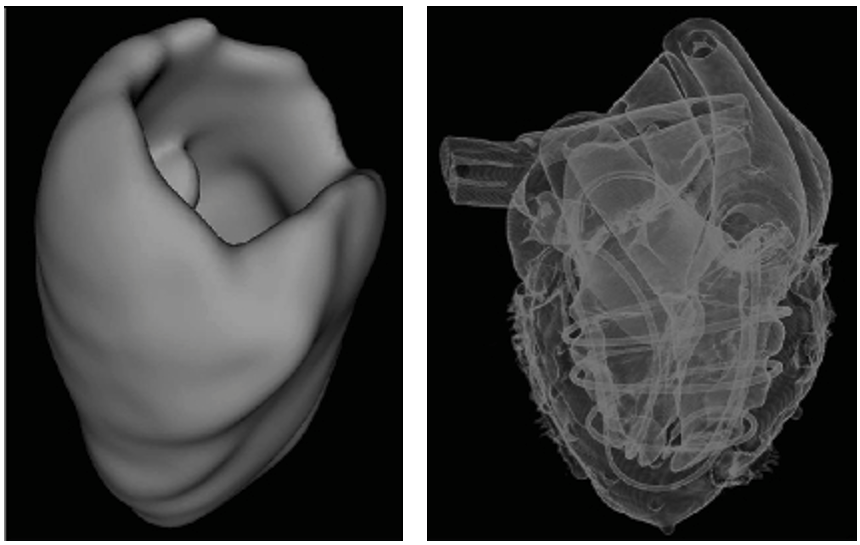


Figure 2.18: Surface rendering techniques: a) the indirect surface rendering of cardiac structure and b) the direct surface rendering of an MR heart phantom [321].

The examples of indirect surface rendering include contour filtering. Contour filtering decides how contours of two successive slices need to be connected where the vertices of the assigned contours should be connected to form triangular mesh [323] (see Fig 2.18a).

The marching cubes algorithm [151] and its improvements [324] are commonly used as part of direct volume rendering to display high-quality surface rendering for medical 3D volume data. The marching cubes algorithm divides the 3D volume into many voxel cubes that form a voxel array [325]. It calculates a triangle mesh that represents the surface of an object, based on a threshold t, the so-called isovalue. The algorithm marches through the entire volume along the voxels and forms a cube with eight voxels each [326]. The surface rendering of different parts of the medical data can be distinguished and visualized as shown in Fig. 2.18b.

2.5 Summary

This chapter introduced clinical motivation for each subsequently developed tool with the main application focus on the first trimester ultrasound screening. Specifically, how image and video analysis, human gaze and saliency prediction can help guide a sonographer to important fetal anatomical structures. Next, the review discussed literature on 3D fetal ultrasound reconstruction using probe motion and ultrasound image data, to visually assist, help guide sonographers and ease the navigation around the fetal womb. Finally, the chapter elaborated on challenges and limitations that exist in current fetal ultrasound examination and described possible image analysis solutions and reviewed progress made in related literature.

3 Datasets

Contents

3.1 Introduction

This chapter provides a detailed description of data used in this thesis. We present the data acquisition, data preparation, the subsets of data used in each chapter,

and the detailed train/validation/test sets partitioning. Data visualisation is also provided for better understanding of the dataset.

3.2 PULSE Fetal Ultrasound Dataset Curation

We use a novel dataset of clinical fetal ultrasound exams with real-time sonographer gaze tracking and probe motion data. The acquired dataset includes ultrasound videos, probe motion, sonographer eye gaze and audio [327]. The exams were performed on a GE Voluson E8 ultrasound machine (General Electric, USA) equipped with standard curvilinear (C2-9-D, C1-5-D) and 3D/4D (RAB6-D) ultrasound transducers used to perform all the ultrasound scans used in this work. Probe motion tracking is achieved using an IMU (Inertial Measurement Unit) attached to the probe and is recorded at 400 Hz. Further details about the probe motion data are discussed in section 3.5

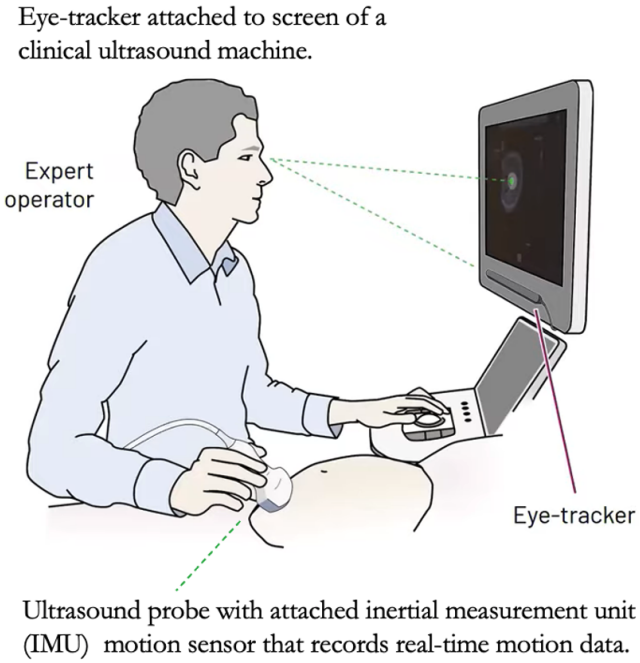


Figure 3.1: PULSE set up.

The sonographers used an LCD monitor with 1920 × 1080 pixels resolution and a refresh rate of 60 Hz. The video signal of the ultrasound machine monitor is recorded at 30 Hz while eye gaze is simultaneously recorded at 90 Hz with a

Tobii Eye Tracker 4C (Tobii, Sweden). The eye tracker records the point-of-gaze data (relative x, y-coordinates with corresponding timestamp) and 3D eye position of each eye, effectively recording 3 gaze points per frame. The eye tracker was rigidly attached under the display area (see Fig. 3.1) with a magnetic mounting bracket as per the instruction of the product. The eye tracker was calibrated for each sonographer following a 9-point calibration protocol. Sonographers were free to adjust the height of the chair and the inclination of the monitor to avoid an eye tracker restricting the view.

Ethics approval was obtained for data recording and data are stored according to local data governance rules. This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051) and the ERC ethics committee. Written informed consent was given by all participating pregnant women and sonographers. Figure 3.2 provides a visualisation of the PULSE[1] dataset used in this work.

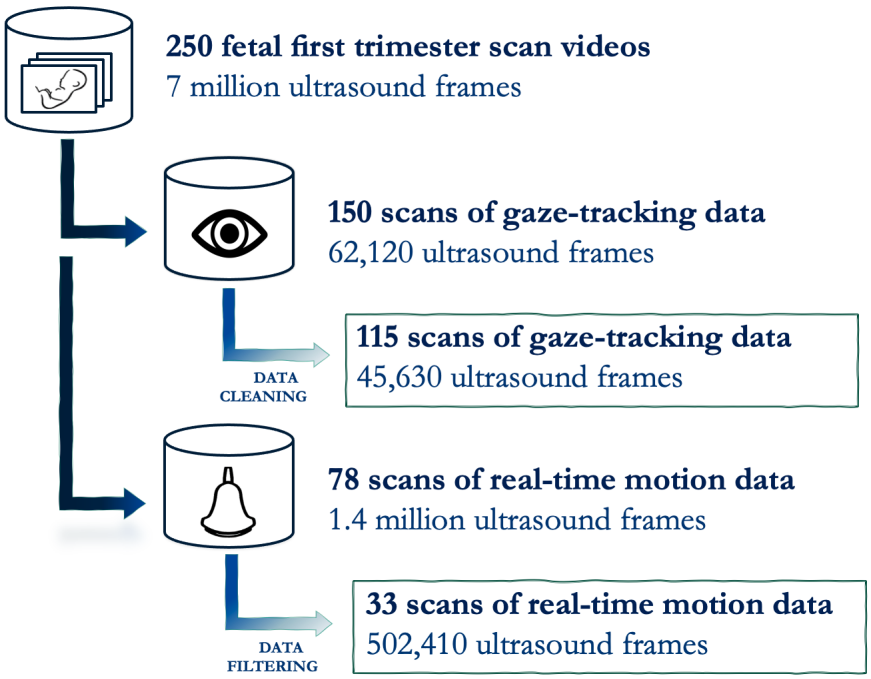


Figure 3.2: PULSE first trimester dataset description with further breakdown justification in sections 3.3.2 and 3.5.2.

The PULSE project data includes pregnancy scans of the first trimester from

[1]PULSE: Perception Ultrasound by Learning Sonographic Experience, http://eng.ox.ac.uk/pulse/

11^{+0} to 13^{+6} weeks of gestational age. In total, the dataset consists of 250 first trimester full-length ultrasound (US) video scans ($7,050,130$ frames) illustrated in Fig. 3.2. The scans represent healthy pregnancies, with no specific count of abnormalities documented in the dataset, reflecting the typical variety of pregnant women attending routine scans. For the tasks and constraints described in sections 3.3 and 3.4, we extract real-time gaze tracking data and end up with 150 scans of of short video clips which translate to a total of 62,120 frames. For the task described in section 3.5, we end up with 78 scans of real-time motion data. A further breakdown of first trimester gaze-tracking and motion data with corresponding US video clips are discussed in sections 3.3 and 3.5, respectively. There are 5 biometry planes that are captured during a first trimester ultrasound scan: crown-rump length (CRL), nuchal translucency (NT), fetal brain (BPD), abdomen (ABD), heart (HR) plus "other" class that contain other anatomy structures such as the placenta and femur. The six planes are illustrated in Figure 3.3. The US clinical workflow in terms of distinguishing successive scanning of biometry planes is discussed below.

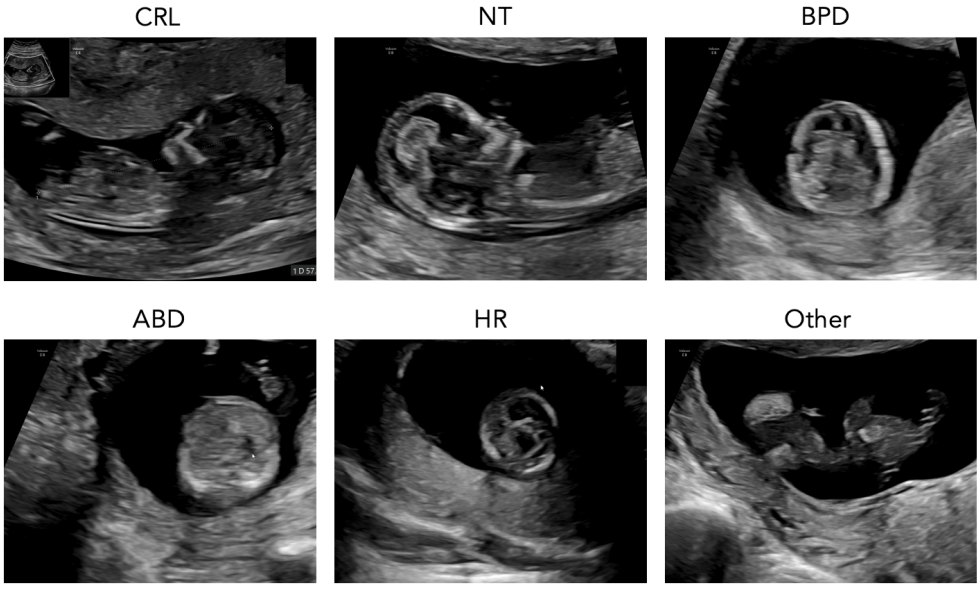


Figure 3.3: Illustration of 6 fetal planes that occur in a first trimester scan.

In an ultrasound scan, the main goal is to find standard anatomical planes of the fetus to allow their diagnostic examination. The fetal structures can be captured in

various acquisition planes (axial, coronal and sagittal) using different ultrasound modes (2D, Doppler, and 3D) and methods (transabdominal and transvaginal).

Following a defined protocol (FASP[2]), a sonographer manipulates a transducer whilst searching for a satisfactory view of a fetal anatomy (denoted as live B-mode on a diagram in Fig. 3.4). Once the fetal anatomical plane is found, a clinician may freeze the frame (denoted in blue and displayed as a snowflake on the ultrasound machine used in this work) where a series of two-dimensional standard imaging planes are automatically stored [327]. Finally, the measurement of the best imaging plane is taken, analysed (denoted in red) and buffered into the ultrasound machine. The search-freeze procedure may be repeated if a satisfactory standard plane is not found. In practice, due to a small size and frequent movement of a fetus, there is a high likelihood that a sonographer may re-visit the same anatomy multiple times to record the best standard imaging plane to increase the accuracy of anatomical assessment and measurement [328].

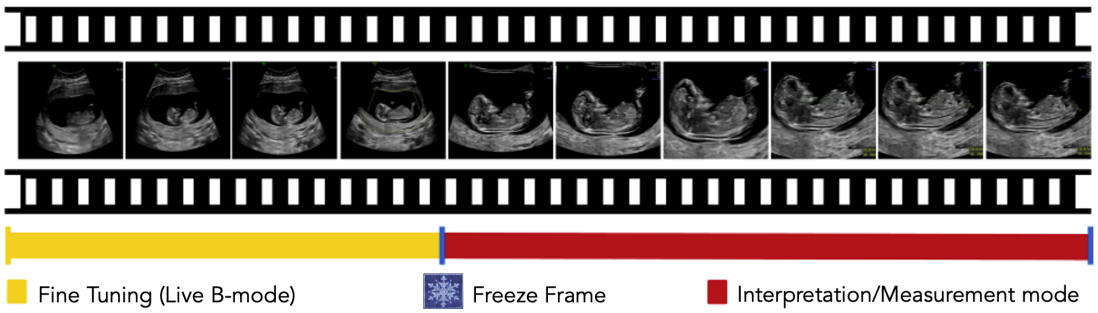


Figure 3.4: Scanning mode of a first trimester ultrasound examination in three stages: exploration in live B-mode (yellow), freezing the frame (blue) and analysing/taking measurements of the best image (red).

During an ultrasound examination, the proportion of time spent performing different anatomical tasks varies. Automated US clinical workflow analysis has been reported by Yasrab et al. [329]. Using semantic anatomical description of a first trimester ultrasound scan (shown in Fig. 3.5) the authors assessed the amount of time spent searching for the anatomies of interest and analysed the order in which these anatomies are scanned. Figure 3.5 shows some typical

[2]FASP: Fetal Anomaly Screening Programme, protocol which provides guidance for sonographers to routine standardized US scans.

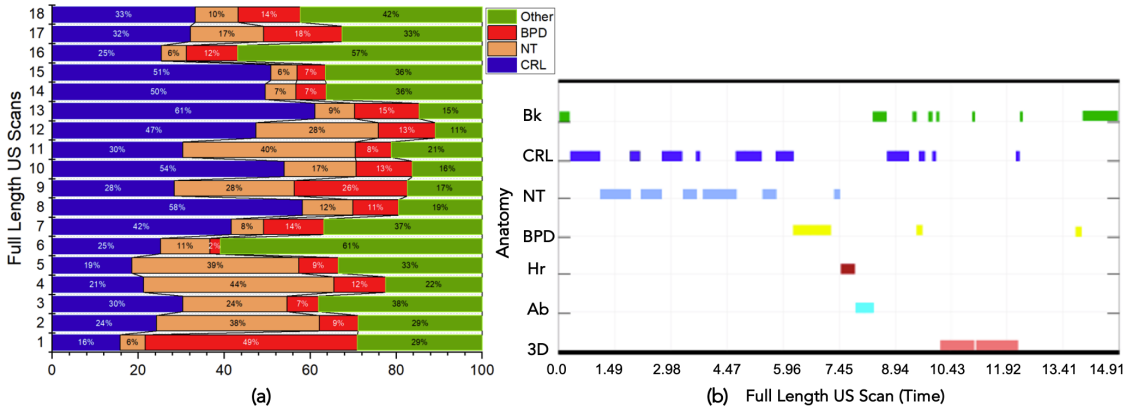


Figure 3.5: First trimester US clinical workflow analysis. a) The percentage of time spent on key anatomical tasks (CRL, NT, BPD). b) A sample full length US scan with labelled anatomy appearing throughout the US examination [329]. Note: Bk and Ab represent classes "other" and abdomen, respectively.

samples of labelled videos, where each video is partitioned into key anatomical tasks carried out during the first-trimester scan. Figure 3.5(a) shows the percentage of time spent on key anatomical tasks and (b) shows the order of each anatomy appearing throughout a scan.

Based on 250 first trimester US videos, authors [329] found the mean duration of an ultrasound scan to be 12.4 minutes (interquartile range (IQR) 9.6-19.5 minutes). In addition to 6 anatomy planes that can be seen in the first trimester scan, 3D-mode which is not an anatomical structure, is also used to assess the fetus. On average, 5.2 minutes (IQR 4.6-9.8) were dedicated to CRL measurement, 2.0 minutes (IQR 1.8-2.0) to NT measurement and 1.5 minutes (IQR 1.9-2.1) to BPD measurement. This shows that the CRL measurement prevails over the other anatomical planes taking up approximately 42% of a full-length ultrasound scan.

A first trimester fetal examination time is limited to approximately 20 minutes per scan and only two biometric measurements, Nuchal Translucency (NT) and Crown-Rump Length (CRL), are taken as standardised imaging planes in FASP protocol (see Fig. 3.6). CRL and NT are the only compulsory image components that should be analysed and scored using an image guidance table (from FASP) which consists of 12 components that describe how measurements need to be taken. If all 12 components are present in a biometry plane, an image receives a 'good' score which

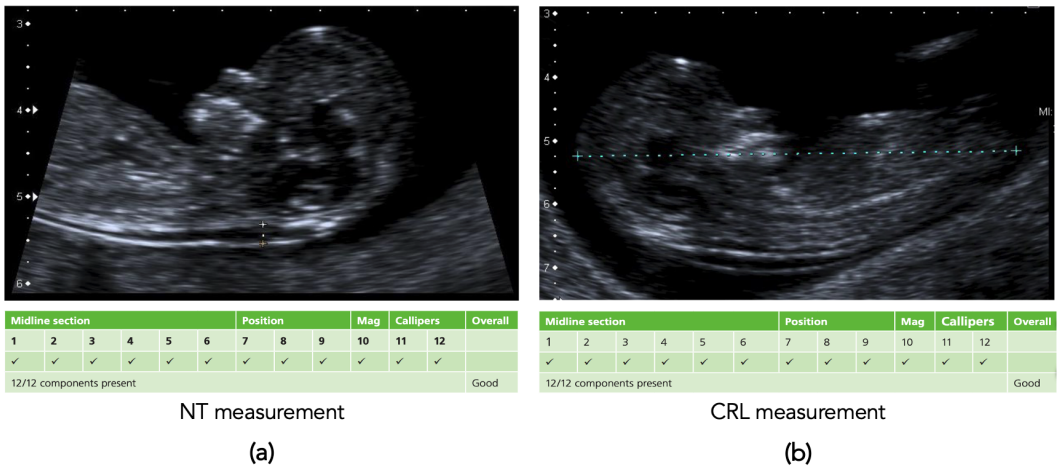


Figure 3.6: a) Nuchal Translucency (NT) and b) Crown-Rump Length (CRL) measurement. Sonographers analyse each biometry plane using a FASP guidance table and score each image (i.e. good, acceptable or poor) by the number of components present in an image.

demonstrates evidence of good clinical practice. Consequently, a sonographer's main focus falls on the measurement of CRL and NT. The measurement of both planes happens in the mid-sagittal (or more accurately median) view of the fetal profile.

Each US video used in this study is a full-length video of a session acquired through screengrab. This means that it includes screen time when the operator recorded and saved personal details of the subject; and times when the US probe shows no activity during the scanning session or the human gaze cannot not be tracked. This may happen, for instance, when the sonographer is speaking to the subject or colleagues and is looking away from the screen and/or stops operating the probe. Both scenarios of missing data will be discussed in the relevant sections 3.3 and 3.5.

Throughout a scan, there are also significant portions of video unrelated to a final view of a fetal anatomy where sonographers spent time exploring and searching for a high-quality view of a fetal structure through the noise, artifacts and an inadequate view of a fetus. Hence, for the task discussed in sections 3.3 and 3.4 we fully focus on the fetal frames which are directly related and lead to a frozen segment (standard planes) and exclude frames acquired during exploration.

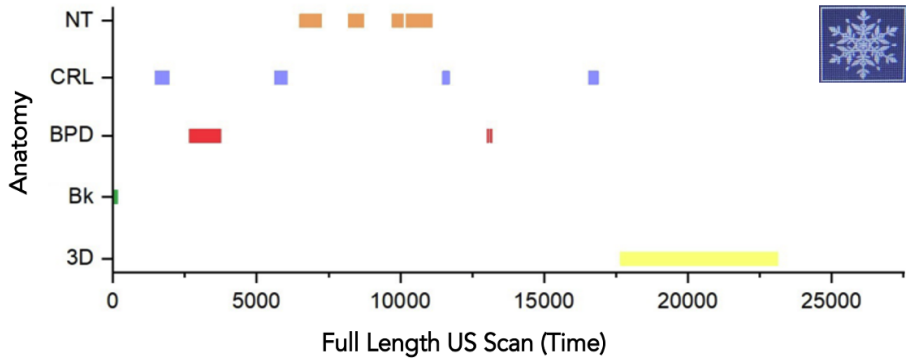


Figure 3.7: An example of frozen segments for different fetal anatomies appearing throughout a full-length first trimester scan.

A freeze frame (FF) video segment is recorded when a sonographer is satisfied that it is a standard plane, i.e. the most clear and representative image of a fetal anatomy. The average length of each FF segment is 31.8 seconds. A complete first-trimester video scan can be divided into different FF categories; CRL, NT, Brain, Heart and Abdomen scans of different subjects which are identified by optical character recognition (OCR) and extracted (displayed in Fig. 3.7).

In practice, each first trimester video contains between 2 to 7 frozen segments. All frozen segments are manually annotated by a sonographer according to the FASP protocol and saved into a memory buffer for further analysis. Freeze frames are automatically detected using OCR (snowflake illustrated in Fig. 3.7) from the full-length videos. Each video is processed to remove user interface information and to anonymize the sonographer ID.

3.3 Chapter 4: Data Augmentation Dataset

This section presents the subset of PULSE data used in Chapter 4. For this study, we used two modalities, synchronised first trimester ultrasound scan videos and real-time gaze tracking data. No human annotations were used in this work except for the technical annotations input by sonographers as part of the routine ultrasound scan.

3.3.1 Data Partitioning

Recall from Figure 3.1, the PULSE dataset contains 250 first trimester scan videos and 150 scans of real-time gaze tracking data. It can be noted that the quantity of US video frames exceeds the number of gaze data points; whilst there is a continuous recording of an US machine monitor, an eye gaze can be lost due to a sonographer looking away from the screen when performing an ultrasound examination (refer to section 3.3.2.2). For this reason, we solely utilise frames which have a corresponding eye gaze data point. For model training, the US data used in Chapter 4 consists of frames that appear 3 seconds before the first freeze-frame (described in Fig. 3.4).

A 90-frame interval was chosen for two main reasons. First, visual confirmation with sonographers indicated that the key anatomical features relevant to the freeze-frame decision were consistently captured within this 90-frame interval. Second, the 3-second interval was found to provide an uninterrupted sequence of 90 US frames with corresponding sonographer gaze data leading up to the freeze frame. Extending beyond this interval often introduced frame gaps in some scans, resulting from instances where sonographers looked away from the screen or due to necessary data cleaning.

After data cleaning, for our experiments, we use 115 fetal first trimester videos and their corresponding gaze points; the selection and boundary conditions are discussed in detail in section 3.3.2.

The data are split into training, validation and test datasets, with 70/17/29 training/validation/test split by a number of US scans (videos). We chose to split the data on a video level to keep the temporal relation of frames intact.

Further data split is performed when preparing the eye gaze data for temporal smoothing (TS) of saliency maps (displayed in Fig. 3.8 and further described in Fig. 3.12). The total number of US videos translates to $45,630$ US frames and their corresponding gaze points. Dataset division results in 70 videos ($29,250$ frames) for training, 17 videos ($7,290$ frames) for validation and 29 videos ($9,126$ frames) for testing.

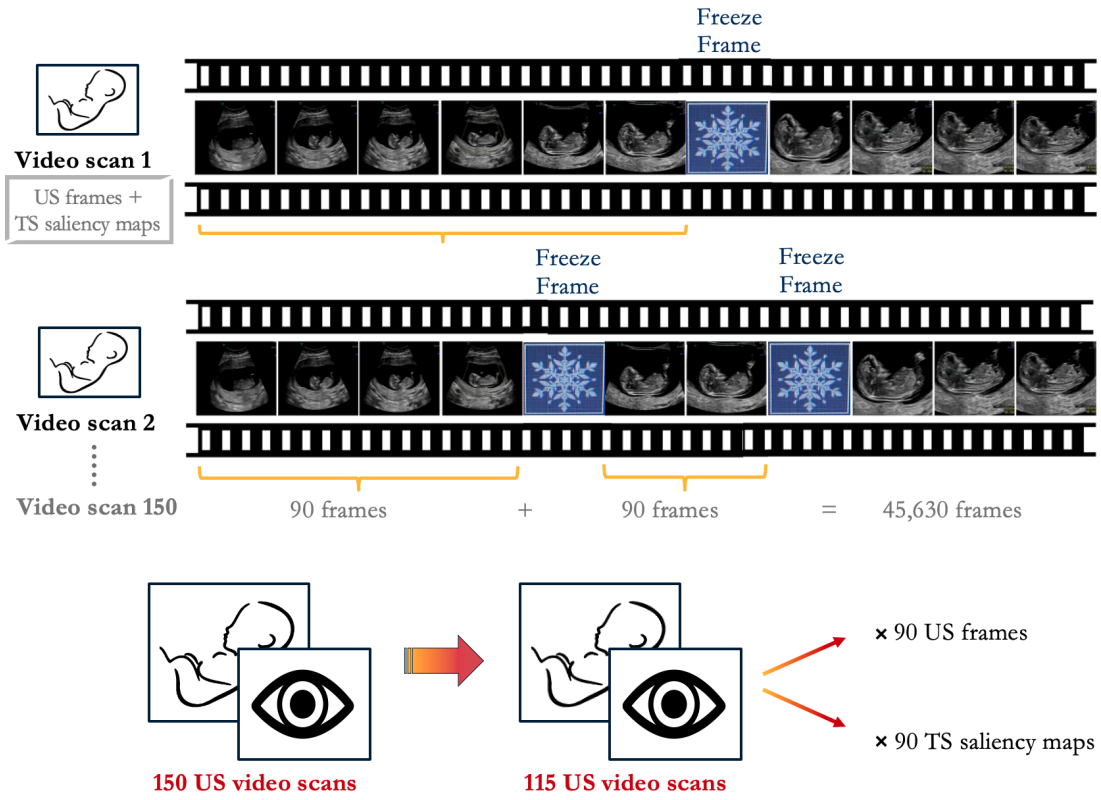


Figure 3.8: Further breakdown of a PULSE first trimester dataset, to prepare for eye gaze manipulation. Initially, the dataset contained 150 US video scans. For each scan, we analyse the data that appears 3 seconds before the first freeze frame which translates to 90 frames (at 30Hz US video rate). First, the binary gaze points are smoothed to generate saliency maps (see Fig. 3.10 and section 3.3.2.2 for details). Second, using 102 saliency map frames before a freeze frame, 90 saliency maps are temporally smoothed (TS) (see Fig. 3.12). Finally, US frames and TS saliency maps are equal in size where each video scan contains a multiple of 90 frames (90 frames * 107 batches = 46,630 frames) and the quantity of video scans is 115.

3.3.2 Data Preparation

This section is divided into two parts: manipulation of ultrasound video frames and manipulation of gaze data.

3.3.2.1 Ultrasound Frame Manipulation

Discarding irrelevant frames All video frames that did not correspond to 2D B-mode live scanning (e.g. Doppler, 3D/4D or frozen frames) or had no gaze data were discarded to ensure consistency in image quality and clinical relevance. US videos that contained less than 1800 frames (1 min) were discarded. This decision was based on typical first-trimester ultrasound scan durations which have a mean

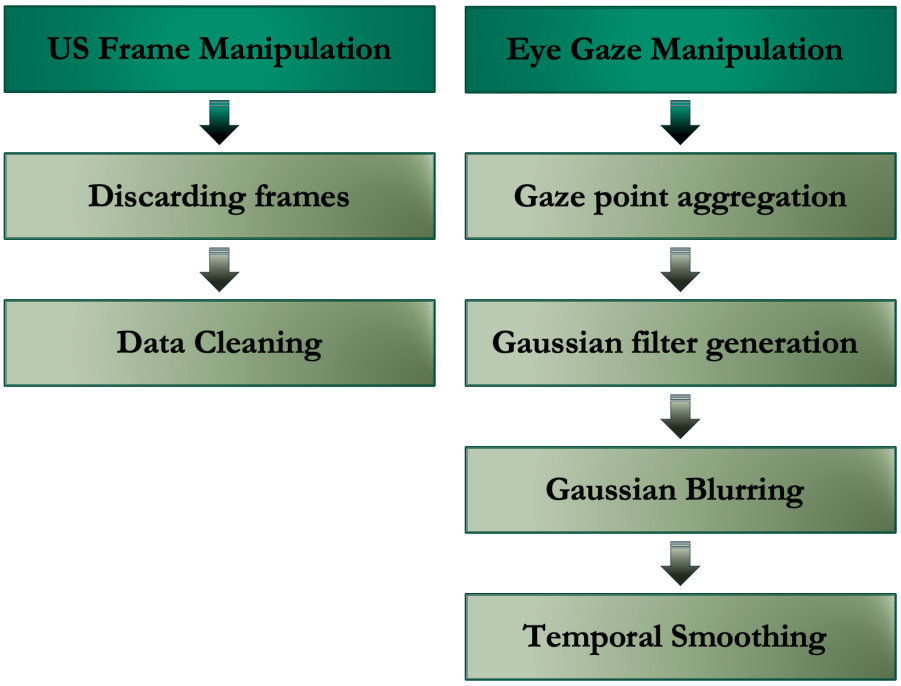


Figure 3.9: A summary of the data preparation procedure for two modalities: US video frames and eye gaze data.

of 12.4 minutes (see Figure 3.5). Setting a minimum duration of 1 minute ensures that even the shortest videos in our dataset contain sufficient temporal information to capture meaningful scanning patterns and sonographer gaze dynamics

Data Cleaning All 150 fetal US videos were manually inspected. Frames that were corrupted or displayed two ultrasound views simultaneously in a split-screen mode were manually deleted. The split-screen mode allows a sonographer to compare two views side by side but complicates the analysis of individual frames. After the removal of such frames, the final dataset included 115 US videos and a total of 45630 individual US frames.

3.3.2.2 Eye Gaze Manipulation

Recall from section 3.2 the eye tracker used in this work records the point-of-gaze data (relative x, y-coordinates with corresponding timestamps). Using (x,y) gaze point coordinates, a binary gaze map is generated (described in Fig. 3.10). Each gaze point or a binary gaze map represents a salient point in an image, i.e a pixel

which attracts human attention. The probability that each pixel in the image will attract human attention is represented by a gaze tracking heat map or a gray-scale heat map, hereafter referred to as *saliency map*. Saliency maps are used as a tool to describe the visual attention of a human eye within a certain visual field, hence, the maps help us in predicting where sonographers may look next. Using Gaussian blurring (illustrated in Fig. 3.11), saliency maps smooth the image which makes it easier to analyse and indicate what region (or point) is salient or not [330]. The Gaussian blurring technique and generation of saliency maps are discussed below.

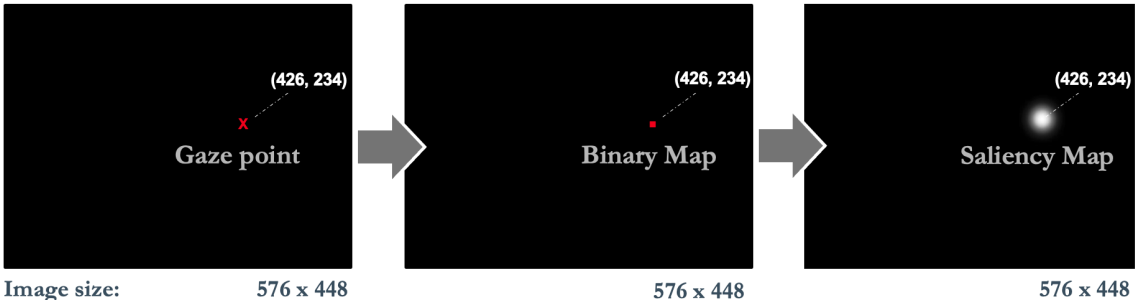


Figure 3.10: Illustration of how eye gaze data is transformed to train a saliency map prediction model. Left to right: A single gaze point with (x,y) coordinates is labelled as 1 and the background space as 0, creating a binary map. Gaussian kernel (illustrated in Fig. 3.11) is convolved with a binary map to generate a Gaussian (saliency) map for saliency prediction.

Gaze point aggregation Prior to generation of saliency maps, gaze points that were located outside of the ultrasound image fan were discarded.

Gaussian kernel/filter generation To make predictions using eye gaze data, the binary gaze maps need to be transformed to represent a distribution of gaze points where some pixels attract more human attention than others. For better analysis of eye gaze probability distributions, the images with gaze maps are smoothed with a Gaussian filter.

Sonographer's eye gaze attention is represented in a form of a heat map rather than a binary point, where a binary map was smoothed assigning higher weights (degree of importance) to pixels that have a higher probability of attention (closer to the centre). A higher value pixel receives the heaviest weight (the highest Gaussian weighted value) and neighboring elements receive smaller weights as their distance

from the centre increases. This creates a more informative saliency map that highlights regions of likely attention with a smooth, gradual decay from the peak.

A Gaussian filter is a type of low-pass filter where a Gaussian blur smooths uneven pixel values, blurs sharp edges and high frequency features in an image by cutting out the extreme outliers [331].

The Gaussian filter was chosen for practical and conceptual reasons: Gaussian distributions are widely used in saliency modeling due to their symmetry, simplicity, and capability to approximate human attention distribution. By applying a Gaussian blur to each gaze point, we simulate the focus and peripheral spread of attention, making regions closer to the gaze point more salient while gradually reducing importance with distance.

Gaussian smoothing also helps mitigate the limitations of binary maps, which would otherwise produce only one salient point with sharp, unnatural edges. Smoothing these maps yields a continuous probability distribution, facilitating better comparison between predicted saliency maps and ground truth data. This distributional approach enables the model to learn nuanced variations in attention, reflecting both focal points and surrounding areas with diminishing importance.

The properties of a Gaussian kernel are detailed below and can be visualised in Fig. 3.11.

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{3.1}$$

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.2}$$

The Gaussian distribution in the 1D and 2D cases are shown in Equations 3.1 and 3.2, respectively. The standard deviation, σ, determines the *width* or the *degree of smoothing* of the Gaussian kernel (discussed in detail below).

In our work, the Gaussian filter (illustrated in Fig. 3.11) has a standard deviation, $\sigma = 13.5$. The standard deviation of the Gaussian is equivalent to ca. 1° visual angle to account for the radius of visual acuity and the uncertainty of the eye tracker measurements [332]. The larger the standard deviation is, the stronger the

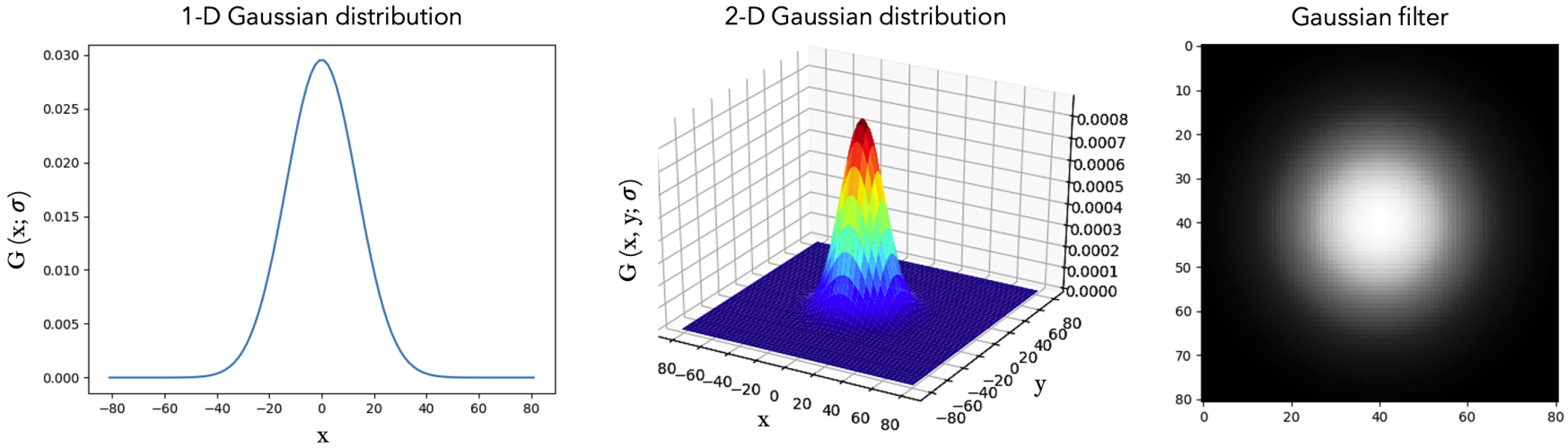


Figure 3.11: Illustration of a single Gaussian filter/kernel. Left panel: 1D Gaussian distribution. Middle panel: General form of 2D Gaussian with zero mean. Right panel: depiction of Gaussian filter in grayscale (white = high, black = low value)

effect of smoothing on the binary gaze map. As mentioned above, the numerical value of a window size or the width of a Gaussian kernel is $x = y = 6\sigma$ which equates to 81 pixels (seen in Fig. 3.11 on each grid of Gaussian distributions). It can also be noted that the Gaussian distribution has symmetrical tails due to a mean, $\mu = 0$. The Gaussian kernel specified above is isotropic, which means that the behaviour of the function is in any direction the same. For 2D this means the Gaussian function is circular [333]. Hence, the kernel is rotationally symmetric with no directional bias where $x = y$.

The standard deviation also determines the amplitude or the peak of a Gaussian kernel. Recall, from the Equations 3.1 and 3.2, the terms $\frac{1}{\sqrt{2\pi}\sigma}$ and $\frac{1}{2\pi\sigma^2}$ in front of 1D and 2D Gaussian kernels, respectively, are the normalization constants. For every σ, the Gaussian kernel needs to be normalised to unity (total area under the curve is 1) as it averages the pixel weights of a kernel. This means that increasing σ of the kernel reduces its amplitude. The peak value of the Gaussian PDF can be calculated using the Gaussian density (where $x = y = \mu$) which is the same as the normalisation constant. Note, from Figure 3.11 1D Gaussian distribution peak is approximately 0.03 and the 2D Gaussian peak is 0.0008. When convolving 1D Gaussian in the x direction with 1D Gaussian in the y direction, the kernel becomes broader and the binary map gets more blurred.

Gaussian Blurring Once a suitable kernel has been calculated, the Gaussian smoothing can be performed using standard convolution methods which represent sonographer's eye gaze attention. First, the horizontal direction is filtered by taking each pixel in an image, centering the filter on that pixel and multiplying the pixel values by the weight at each filter location. When converting the Gaussian's continuous values into the discrete values needed for a kernel, the sum of the values is not 1. The image would darken or get brighter. To avoid this, the values are normalized by dividing each term in the kernel by the sum of all terms in the kernel. This process is then repeated vertically on the horizontally processed image to create the final image.

Recall, the 2D Gaussian filter used in our work is rotationally symmetric and hence, isotropic. That is, instead of performing a 2D convolution at once, the kernel is separable where each dimension can be processed separately but run in parallel. All the properties of a separable 2D Gaussian kernel make the computation time more efficient.

A final 2D Gaussian blurring of a gaze map can be seen in Figures 3.10 (last frame) and 3.12, where each frame (in yellow) represents the eye gaze attention of a sonographer.

Temporal Smoothing The dataset in Chapter 4 is used for *single frame saliency prediction* where a sonographer's visual attention is modelled on static ultrasound video frames through prediction of saliency. The model learns static visual attention by treating each US video frame as an independent image. To represent the eye gaze distribution and capture a gaze pattern a sonographer might have over the similar US frames, a concept called *temporal smoothing* was explored.

Temporal smoothing of eye gaze maps or saliency maps was used to analyse the spatial changes within successive frames. The idea is to sum the preceding gaze points (or generated saliency maps) for each frame to represent the gaze distribution that a sonographer might have over the similar frames (resented for visualisation in Fig. 3.12 with description of the process below).

There are two ways to perform temporal smoothing. One way is to aggregate all the binary maps, temporally smooth the gaze points and smooth the eye gaze maps using Gaussian Blurring. In contrast, the order of steps can be changed where saliency maps are generated first, then aggregated and finally, temporally smoothed. The latter option was implemented.

In Chapter 4, we estimate the gaze pattern 3 seconds before the sonographer freezes at the standard plane. With the monitor refresh rate of $30fps$, we obtain 90 US frames ($3sec * 30fps = 90frames$) which are used as model prediction input and the corresponding 90 saliency map frames used as model ground truth.

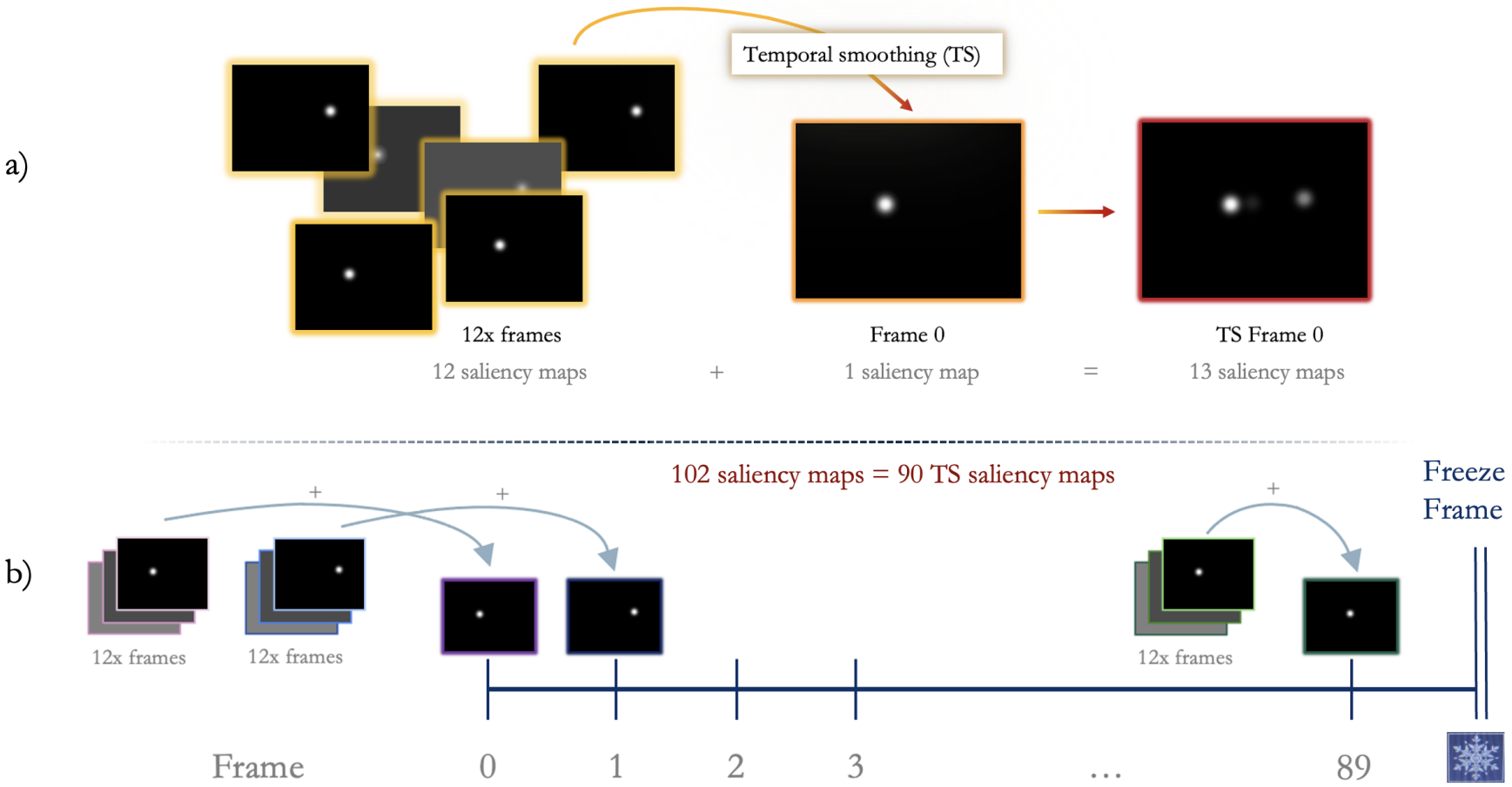


Figure 3.12: Illustration of temporally smoothed saliency maps. a) Temporal smoothing (TS) of a single frame (i.e $Frame_0$): 12 consecutively ordered frames that appear before Frame 0 are superimposed on $Frame_0$ (1 saliency map per frame). After TS, Frame 0 contains 13 saliency maps. b) An overview of TS on a batch of frames (3 sec before the freeze-frame): 90 saliency map frames are each temporally smoothed with 12 preceding frames. Hence, 102 frames are required to temporally smooth a batch of 90 frames to account for $Frame_0$

To prevent neural network from memorizing the data, all the US input frames with the corresponding saliency maps are shuffled. As the shuffling forces frames to be independent of each other, it becomes challenging to predict the direction of the forthcoming gaze point. In a space of a second, the content of US video frames does not dramatically change whilst the trajectory of gaze may. Hence, by superimposing gaze points from the previous frames onto the frame in question, we represent the gaze distribution that a sonographer might have over the similar frames. Hence, even after the frame shuffling a pattern of the eye gaze is stored on each frame and can be analysed. A decision to choose a specific number of frames for an overlay comes from a concept called *persistence of vision* (illustrated in Fig. 3.13 and discussed below).

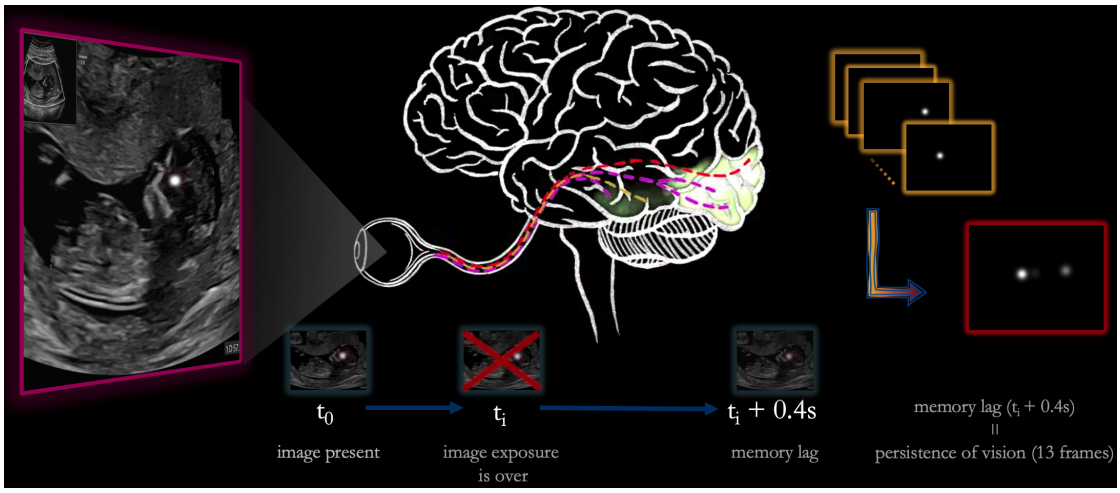


Figure 3.13: Persistence of vision concept. Sonographer's eye perceives an US image at time t_0 where each image has a corresponding saliency map. When the exposure to an image is over (t_i), the brain continues to perceive an image up to 0.4 seconds ($t_i + 0.4s$). New images are viewed and accumulated in retina for 0.4 seconds (a total of 13 are accumulated at time $t_i + 0.4s$ in a $30Hz$ ultrasound video). The concept is applied to saliency maps created from eye gaze (on the right, in yellow) and is used to temporally smooth saliency maps (in red). A more detailed illustration can be found in Fig. 3.12).

The concept of persistence of vision comes from human physiology [334] and how a human eye perceives its surroundings. A residual image printed in the retina of an eye is an optical phenomenon that causes the human brain to continue to perceive an image for 1/25 second, even when its exposure to the eye is over. If a successive image replaces the prior within this time-frame, it creates an illusion of continuity. As a result, the frame persists in a human eye for $0.1 - 0.4$ seconds which translates

to $3-12$ frames of $30Hz$ ultrasound video. As the human eye can retain the image on the retina, the neural network can also be taught that. With the notion of visual persistence, we perform temporal smoothing using 12 past frames to store the maximum amount of temporal information in a single frame (shown in Fig. 3.12a). Specifically, 12 past frames are overlayed on the successive frame, to give $frame_{13}$. That is to say, the $frame_{13}$ is temporally smoothed using a total of 13 saliency maps.

To conclude, a single ground truth frame is temporally smoothed when it contains 13 saliency maps. As we train a model with the data 3 seconds (90 frames) before the freeze-frame, we use a total of 102 saliency map frames to account for $frame_0$ and perform temporal smoothing on all 90 frames. Within each batch, the binary gaze maps are smoothed with a Gaussian kernel and the generated saliency map frames are temporally smoothed. The final dataset for training consists of 90 temporally smoothed saliency maps and 90 corresponding US images (see Fig. 3.8).

3.3.3 Data Transformation prior to Model Training

All ultrasound and ground truth saliency map frames were randomly shuffled with 1-1 correspondence. The US frames were normalized with Min-Max scaling to fit pixel intensities within the range of [0, 1]. For the calculation of the loss, the ground truth saliency maps are scaled to unit length, where components of a feature vector are divided by the Euclidean length of the vector such that the vector has length of 1. All images are resized to 288×244 pixels for data augmentation. Baseline data augmentation is performed by random rotation with an angle uniformly sampled from [- 25, 25] degrees and random horizontal flipping.

3.4 Chapter 5: Spatio-Temporal Analysis Dataset

This section presents the subset of PULSE data used in Chapter 5. The work in Chapter 5 extends into the temporal dimension by exploring spatio-temporal feature representations of US videos. For this study, the data is processed in the same way as described in section 3.3 with the addition of pre-processing steps to accommodate training in a temporal domain. Our task to estimate the gaze

pattern 3 seconds before a sonographer freezes at the standard plane stays the same, except now the architecture takes 2 inputs. One input is a single US frame and the second input is a sequence of frames preceding and including the US frame in question (refer to Fig. 3.14).

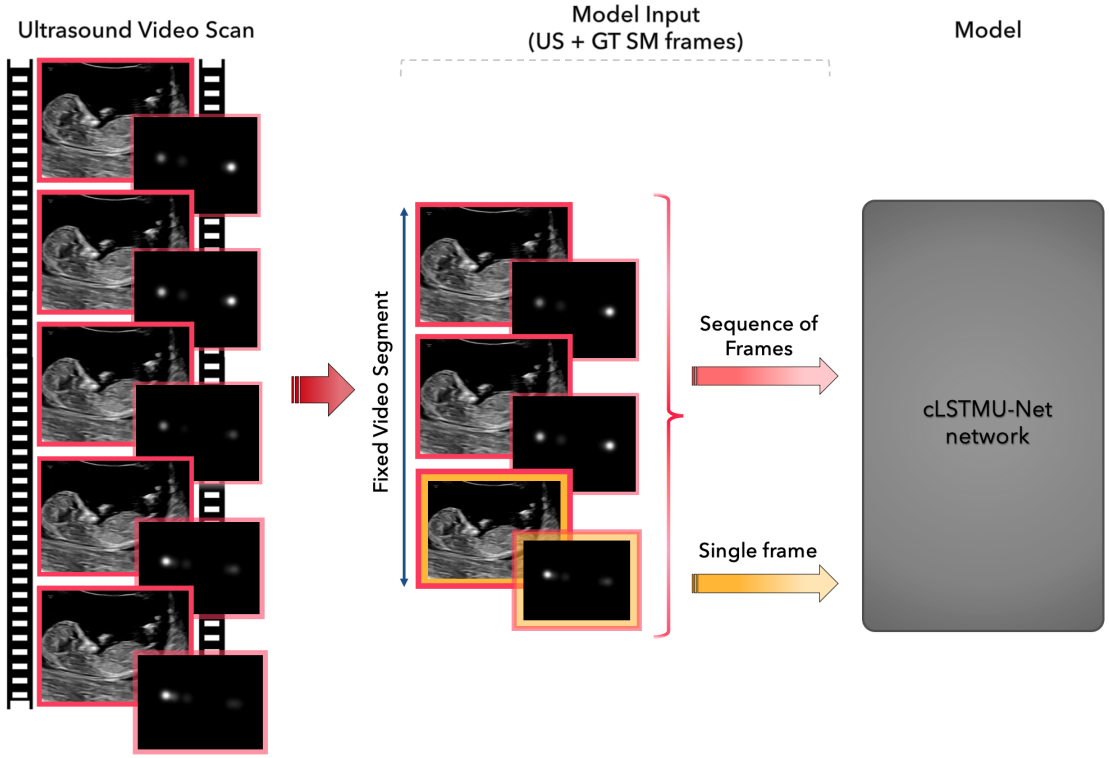


Figure 3.14: Input into a cLSTMU-Net network. The model takes two inputs, a fixed video segment (sequence of frames) and a single frame. The video segment is highlighted in red and its length is predefined before training. A single frame is drawn from the same video clip and fed into a network separately (highlighted in yellow). The input consists of US frames and GT saliency maps.

3.4.1 Data Sampling

Recall from section 3.3.2, that the final dataset for training consists of 90 temporally smoothed saliency maps and 90 corresponding ultrasound frames where each pair is fed independently into a spatial network (see Fig. 3.8). As the work in Chapter 5 explores continuity in a temporal domain, the sampling of an input into a spatio-temporal network changes.

Figure 3.15 illustrates how the data is sampled prior to model training. The estimation of a gaze pattern 3 seconds (90 frames) before a freeze frame stays the

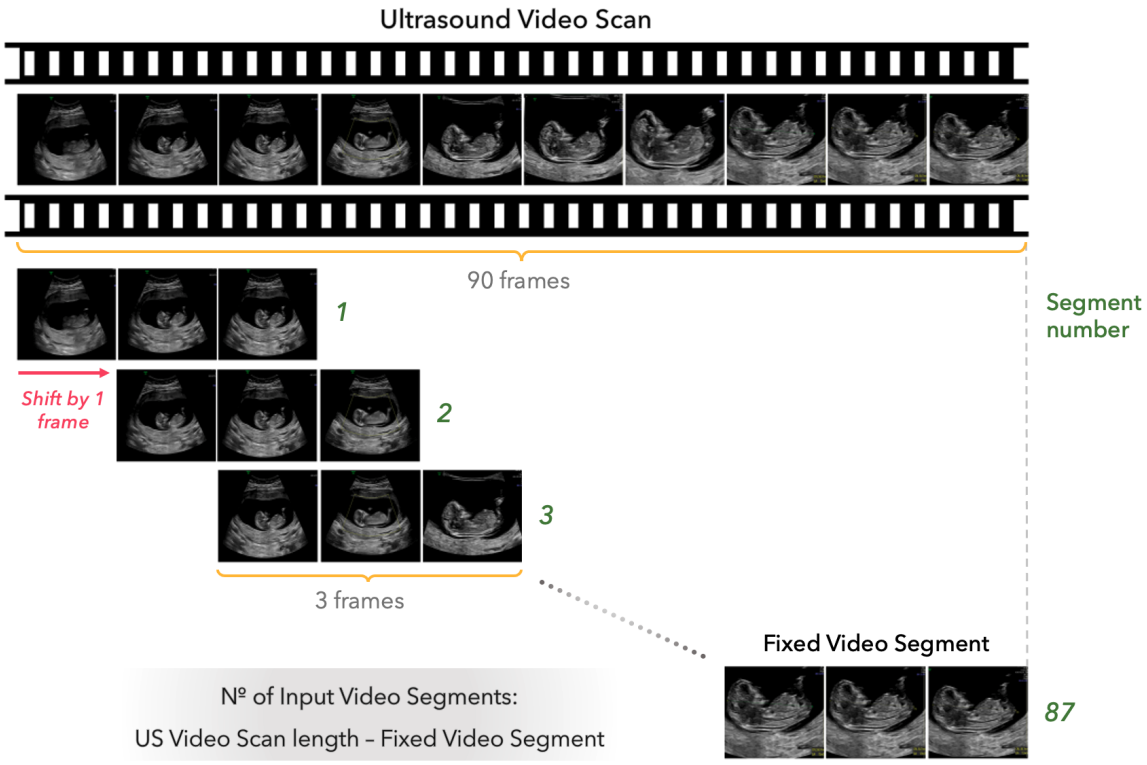


Figure 3.15: Input data sampling for spatio-temporal network (for simplicity, GT saliency maps were excluded from the diagram). The input is sampled from a sequence of 90 frames using a shifting window of a fixed-length video segment (illustrated in this example as 3 frames). Using an interval of 1 frame, 87 video clips (each 3 frames in length) can be sampled from a 3-second video and fed into a spatio-temporal network.

same, except now the sequence of 90 frames is sampled using a shifting window of a fixed-length video segment. A fixed-length video segment represents an input into a spatio-temporal model (described in 3.14). Each video clip is sampled with an interval of 1 frame which allows to capture all the temporal variation without a loss of temporal resolution.

The number of fixed video segments that can be sampled from an original 3-second video can be calculated as: *Number of Video Segments = 90 frames - Fixed-length Video Segment*. Figure 3.15 illustrates an example of US scan data sampling using a shifting window of size 3; where 87 video clips (each 3 frames in length) can be sampled from a 3-second US video. The data is sampled for all stages of neural network analysis, i.e. for training, validation and testing phases.

3.4.2 Data Shuffling

Prior to model training, video clips sampled from a 3-second video are shuffled at random to reduce variance (variability in the model prediction) and increase model generalisation. Whilst the order of fixed-length video segments (in Fig. 3.15) is shuffled, the temporal connectivity between frames of each video clip is retained unchanged. Figure 3.16 illustrates the shuffling process.

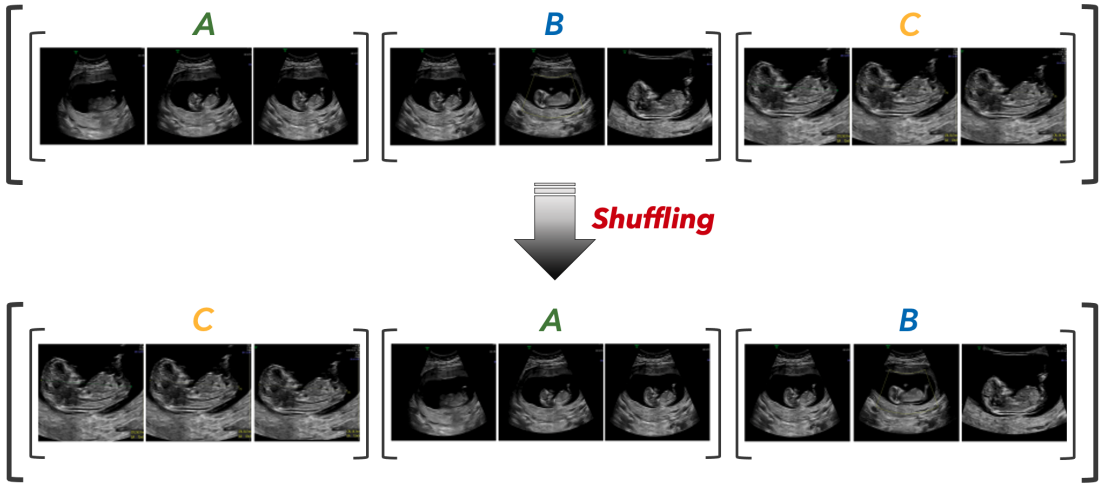


Figure 3.16: Spatio-temporal data shuffling. The order of fixed-length video clips is shuffled (from A, B, C to C, A, B), whereas the order of frames within each video clip remains unchanged. Only the data prior to training phase is shuffled.

The rest of the data pre-processing for Chapter 5 mimics section 3.3.3, which includes US and GT saliency map image normalization, resizing and baseline data augmentations.

3.5 Chapter 6: Visual-Assisted Probe Motion Dataset

This section presents the subset of PULSE data used in Chapter 6. For this study, we used modalities which include first trimester ultrasound scan videos and real-time probe motion data tracked with an Inertial Measurement Unit (IMU). Ultrasound frames were manually annotated by expert sonographers and engineers (see section 3.5.3) for the purpose of classifying the anatomies. Each ultrasound frame contains a corresponding biometry label and a motion data point.

3.5.1 Overview of a Freehand 3D Ultrasound System Setup

This section describes how different tools and components are used to create a freehand 3D ultrasound system. The system setup is comprised of 3 parts: a 2D ultrasound imaging system, a 9 DoF tracking system, and 3D image reconstruction software.

In our prototype system, a Personal Computer (PC) was used as a host for the reconstruction software (described in Chapter 6). Ultrasound images are generated with a General Electric Voluson E8 ultrasound machine. From section 3.2, the video signal of the ultrasound machine monitor is recorded at 30 Hz. The ultrasound machine was equipped with a curvilinear array probes (C2-9-D, C1-5-D) and 3D/4D (RAB6-D) probes with the signal recorded at 400 Hz. Curvilinear or curved linear arrays, also known as convex (sequential) array ultrasound transducers have a frequency range of 2 - 9 MHz. The probes are similar to linear arrays but with piezoelectric elements arranged along with a curved transducer head (refer to Figure 2.11 for visualisation). Ultrasound beams are emitted at 90 degrees to the transducer head. This arrangement results in a trapezoidal field of view due to the divergence of the ultrasound beam with increasing depth. This allows for a wider field of view but with decreased line density at depth and reduced lateral resolution.

The tracking system consists of an inertial measurement unit rigidly attached to the probe to record the probe motion data. IMU-assisted probe tracking processes the raw sensor output to generate estimates of the transducer position and orientation. The IMU sensor consists of an accelerometer, gyroscope, magnetometer and barometer. In this work, only the first three sensors are used (refer to a detailed description of IMU and its components in sections 2.3.1 and 6.3). The selection of suitable motion tracking methods was limited. Electromagnetic trackers or optical trackers could not be utilized for patient safety reasons in the scan room and not wide enough field of view, respectively. Hand motion video recording was also not possible due to patient privacy and no video camera recording allowed in the scan room.

The IMU sensor was thoroughly characterized to quantify performance in terms of accuracy and drift. Section 2.3.2 describes the possible sensor errors and section

6.3 attempts to resolve them. Madgwick's sensor fusion algorithm (described in section 6.3.2 and depicted in Figure 6.2) was used to correct the flaws (accelerometer and gyroscope drift) of raw probe motion data. Simultaneously, sections 6.4 and 3.5.7 describe how 2D ultrasound frames were prepared to be merged with probe motion data (section 6.5). Combined modalities provided a basis for estimating a 3D reconstruction of a fetus (refer to section 6.6.3).

3.5.2 Data Filtering

Recall from Figure 3.2, that the PULSE dataset contains 250 fetal first trimester scan videos and 78 scans of real-time motion data. In the PULSE dataset, the quantity of US frames recorded always exceeds the number of motion data and/or eye-gaze data points. Despite the continuous recording of an ultrasound screen and tracking of probe motion with IMU sensor, there are cases when a sonographer takes the hand away from the patient to perform other tasks; such action creates a gap in motion data which can be detected in the system. For this reason, ultrasound frames which contain no corresponding motion data points were excluded in our analysis. Consequently, from Fig. 3.2 only 33 scans of real-time motion data and ultrasound frames were analysed in Chapter 6.

3.5.3 Human Annotations and US Frame Synchronisation

Ultrasound frames were manually annotated by expert sonographers and engineers for the purpose of classifying the anatomies. Each ultrasound frame contains a single corresponding anatomy label. Human annotations were synchronised with 250 video clips to produce a one-to-one correspondence between each frame and biometry label annotation. Human annotations are an integral part of this chapter as the prime focus of work in Chapter 6 is on 3D reconstruction of a fetal surface. Hence, only ultrasound frames with a crown-rump length (CRL) anatomical label were used to perform fetal segmentation and to extract the fetal masks (refer to section 6.4), followed by further processing to perform reconstruction in 3D.

In total, there are 11 labels which include 7 human annotations and 4 technical annotations which are input by sonographers during the scan. Human annotations consist of 6 anatomical labels which describe the anatomy seen on the screen and one label which describes the 3D mode of ultrasound scanning. Recall from section 3.2 and Figure 3.3, that six labels are defined: abdomen, brain, CRL, heart, nuchal translucency and 'other' anatomy structures (i.e. placenta and femur).

The 4 technical annotations include system annotations or non-anatomical frames and are: screensaver, no probe contact (black frame, no structure detected), anonymised (black screen to conceal input of patients data) and background (searching or transitioning from one structure to another). The labels required to distinguish anatomical and non-anatomical planes are read from the system, populated into the table and further combined with human annotations (see an example in Figure 3.17 below).

labeledclips_2019-10-02T112653_2019-10-02

frame_number	freeze_frame	no_signal	anonymised	doppler	pulse_wave	3d_static	4d_real_time	2d_b_mode	screen_saver	3d_static_prep	xl_mode	anatomical_label
1	NA	0	0	0	0	0	0	0	1	0	NA	Screensaver
2	NA	0	0	0	0	0	0	0	1	0	NA	Screensaver
3	NA	0	0	0	0	0	0	0	1	0	NA	Screensaver
4	NA	0	0	0	0	0	0	0	1	0	NA	Screensaver
4984	FALSE	0	0	0	0	0	0	1	0	0	1	CRL
4985	FALSE	0	0	0	0	0	0	1	0	0	1	OtherStructures
4986	FALSE	0	0	0	0	0	0	1	0	0	1	CRL
4987	FALSE	0	0	0	0	0	0	1	0	0	1	CRL
4988	FALSE	0	0	0	0	0	0	1	0	0	1	OtherStructures
4989	FALSE	0	0	0	0	0	0	1	0	0	1	OtherStructures
4990	FALSE	0	0	0	0	0	0	1	0	0	1	Brain
4991	FALSE	0	0	0	0	0	0	1	0	0	1	Brain
4992	FALSE	0	0	0	0	0	0	1	0	0	1	Brain
4993	FALSE	0	0	0	0	0	0	1	0	0	1	Brain
4994	FALSE	0	0	0	0	0	0	1	0	0	1	OtherStructures

Figure 3.17: An example of human and technical annotations combined to determine ultrasound frame labels.

To perform fetal reconstruction the US frames which contained anything else but the crown-rump length detected on the frame were not considered. An important technical annotation to note is when an ultrasound probe has no contact with the skin of a patient, hence, records a black frame. Whereas, the probe tracked by IMU sensor continues to record motion data.

From Figure 3.5, the percentage of CRL planes appearing throughout the scan and the time spent on the analysis prevails over all other anatomies and planes by taking up approximately 42% of a full-length 12.4 minute scan (on average) (refer to section 3.2). Hence, it is hypothesised that there should be enough continuous CRL data to reconstruct the fetal surface.

3.5.4 IMU-Assisted Probe Motion Data and US Frame Synchronisation

For each scan a local copy of XIO motion data was identified and prepared for further processing. Only continuous XIO motion data was considered (section 6.3.5 describes the process). Recall, only CRL frames were considered for analysis. Hence, once continuous probe motion data was selected and US frames with CRL anatomy were extracted, the total number of ultrasound video scans reduced to 22 synced scans which could be used for 3D reconstruction.

3.5.5 Raw Data Preparation

This section provides a summary of problems faced during the preparation process of XIO probe motion data and ultrasound frames.

3.5.5.1 Motion Data Buffering Problem

One of the data preparation steps included rectifying inaccurately loaded timestamps of an ultrasound probe, i.e. a motion data buffering issue. Ideally, the probe motion data and the timestamps associated with it should have an equal time stamp between each data point to correctly estimate the movement of the probe in space. PULSE motion data contained unequal timestamps with variable gaps between data points at the start of the scan and little to no motion change recorded in the end. The buffering problem was resolved and is described in section 3.5.6 below.

3.5.5.2 Sampling Rate Correlation

The sensor data of both modalities is stored each with a unique timestamp. The inertial measurement unit which tracks the probe motion in space is sampled at 400 Hz (see section 3.2), whereas ultrasound frames are stored at 30 frames per second. XIO probe motion data contains 370 data points per second more compared to US frames recorded in the same scan.

With the final objective of fetal reconstruction in mind, both sensors should match and contain the same sampling rate where each ultrasound frame has a corresponding motion data point. Hence, to reduce the number of IMU data points while preserving useful information, the probe motion data was downsampled to 30 Hz to match the ultrasound video sampling rate. To accurately synchronise four elements, both sensors and their corresponding timestamps, interpolation technique discussed in 3.5.6 was employed.

3.5.5.3 Missing Data

There is missing data in both modalities. There is a way to interpolate US frame timestamps and produce a single ultrasound frame using interpolation and *mixup* (discussed in Chapter 4). However, there is no way to justify motion interpolation as each rotation and change in displacement are unique for each timestamp. In addition, the motion drift is another challenge due to which the interpolation with the aim to fill the missing gaps in data is not considered. In Chapter 6, data point interpolation is solely used for the purpose of final visualisation improvement.

3.5.6 Multi-Sensor Synchronization

This section focuses on solving the issues raised in sections 3.5.5.13.5.5.2/refMissing Data to produce a final synchronisation of probe motion and ultrasound data with their corresponding timestamps.

Prior to synchronisation if an US frame did not have a corresponding probe annotation (i.e. no IMU sensor data), the scan was not considered. To synchronise

data, probe motion data and US frames with their corresponding timestamps required correspondence.

First, the algorithm checks for missing probe data. The sampling rate of a probe is 400 data points per second and therefore the step size is $\frac{1}{400}$. If the total number of data points is not a multiple of 400, the data is missing. To account for that, the probe data is sampled based on the true length of the existing probe data:

$$True\_probe\_step = \frac{(t_{last} - t_0)}{N\_timestamps} \tag{3.3}$$

where t_0 and t_{last} are the first and last probe motion timestamps, respectively. A true probe step is applied to all probe timestamps to resolve the buffering issue and arrange the data points in order with an equal time step. As an example, for a scan X there are $328,463$ probe data points and $24,583$ US frames recorded, with some mismatch in probe motion data. To resolve this, the sampling rates are correlated and buffering problems resolved.

Nearest Neighbour Interpolation Next, based on frame timestamps the nearest probe data and associated probe timestamps are corresponded using nearest neighbour (NN) interpolation. Different types of interpolation were considered and tested, including linear interpolation, NN extrapolation and others. The nearest neighbour method provided the closest fit of motion data points to raw US frame data.

The SciPy library is used to perform 1D interpolation using re-sampled probe timestamps. An interpolation function maps probe timestamps to corresponding indices. Next, the function is applied to US image timestamps to find the indices of probe data that best align with image timestamps, using the nearest neighbor method. In addition, the method allows to estimate values for timestamps outside the range of known probe timestamps by extrapolating from nearby data points. That way, the motion data from the probe has a better match with specific frames of ultrasound images.

Linear Interpolation of Missing Data Both modalities are checked for continuity of data. In case of US frames and possible gaps found between the frame numbers or frame timestamps, linear interpolation is used to estimate the value of only one new data point.

Final Sensor Synchronisation Finally, after a number of interpolation functions are applied to the probe and image dataset, US frames and frame timestamps, IMU motion data with corresponding timestamps are all synchronised (see Figure 3.18).

Time(s)	gyro X (deg/s)	gyro Y (deg/s)	gyro Z (deg/s)	accel X (g)	accel Y (g)	accel Z (g)	magn X (uT)	magn Y (uT)	magn Z (uT)
12.03333333	-0.008526161	-0.154672146	0.010938355	0.976103544	0.038051628	0.29345566	1.14721489	60.7525177	-16.59540939
12.06666667	-0.132556483	-0.038157877	0.010878094	0.981976986	0.03921273	0.282881349	0.772768021	60.75195313	-16.59510803
12.1	-0.008301109	-0.038008396	0.132889941	0.977107823	0.040601425	0.280960143	1.148061752	60.74860382	-16.20709419
12.13333333	-0.070376366	0.020231422	-0.050530329	0.980941951	0.039601896	0.292533934	1.148061752	60.74860382	-16.20709419
12.16666667	0.178385884	-0.154816404	-0.110664502	0.992149115	0.040844295	0.283425778	1.531108856	60.36968231	-15.75835228
12.2	-0.690055013	0.368844539	0.071799532	0.974178195	0.038539432	0.289115906	1.161596298	60.0065918	-16.99170494
12.23333333	-0.070179299	-0.271098197	0.255189449	0.972222567	0.036498383	0.296792775	1.161596298	60.0065918	-16.99170494
12.26666667	0.42662847	-0.038566574	-0.23364754	0.984943688	0.032875903	0.291463315	0.771921158	60.755867	-16.98342133
12.3	-0.131851315	0.252373159	-0.111271895	0.98288548	0.037648596	0.29781276	1.155675888	60.37368774	-16.21108437
12.33333333	0.054059356	-0.039006066	0.011068902	0.983937263	0.040266156	0.277116895	1.155675888	60.37368774	-16.21108437
12.36666667	-0.131573915	0.193619102	-0.111552775	0.980902791	0.042528905	0.291138083	1.528429031	60.38207245	-16.98801613
12.4	-0.007614464	0.077293195	-0.050110124	0.97560215	0.039014831	0.293471009	1.529275894	60.37815857	-16.59970093
12.43333333	0.178391218	0.077166371	0.011261418	0.985839844	0.039248664	0.286754221	1.529275894	60.37815857	-16.59970093
12.46666667	-0.007437035	-0.329869121	0.010927997	0.979100883	0.034741458	0.288582444	0.788843155	59.99819946	-16.21477509

Figure 3.18: Synchronised probe motion and ultrasound data and their corresponding timestamps.

Now, US frames and probe coordinates are aligned and can be merged to create a 3D reconstruction. Figure 3.18 is read by the algorithm in conjunction with the example of data in Fig. 3.17. Recall, from section 3.5.3 only ultrasound frames with recorded human annotation of crown-rump length are considered in this chapter. Figures 3.18 and 3.17 reflect the data of the same ultrasound scan. Due to a non-anatomical label displayed in Figure 3.17 (i.e. *Screensaver*), the first timestamp in Fig. 3.18 s initialised at 12.03 seconds and not earlier.

Further IMU sensor fusion and calibration are discussed in Chapter 6, section 6.3.

3.5.7 Preparation of 2D Ultrasound Frames

This section describes how 2D ultrasound frames were prepared to be merged with probe motion data (in section 6.5) and later used for 3D reconstruction.

First, the fetus is segmented from 2D ultrasound images (the method used is described in 6.4). It was then necessary to refine the fetal masks to form the basis of surface reconstruction which is described in 3.5.7.7.

3.5.7.1 Binary Map Generation

The resulting segmentation maps need to be converted into a binary format for further analysis. The generated fetal masks are represented in RGBA format with red (R), green (G), blue (B) light sources and an alpha (A) channel which specifies the opacity of a color. The alpha parameter ranges between 0 (fully transparent) and 1 (fully opaque). The segmented mask contains pixel values ranging from 0 to 255 with the opacity value of 0.5. Whereas, the background is white (255) with an opacity of 0.

To generate a binary mask, each fetal mask is filtered for the alpha channel values and converted to RGB format. The process is depicted in Figure 3.19.

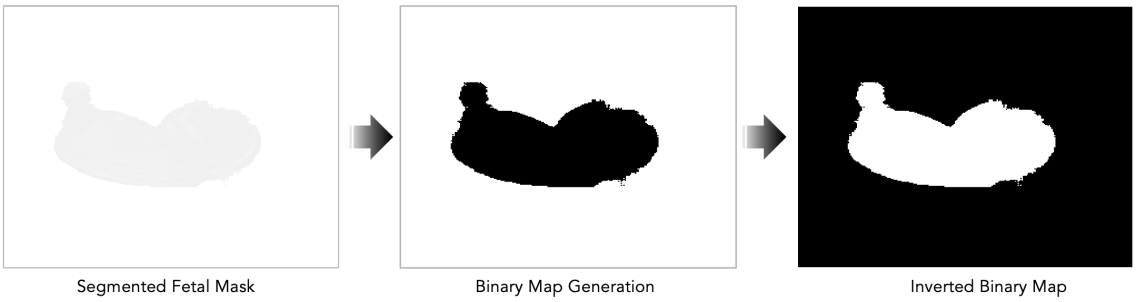


Figure 3.19: Conversion of segmented fetal mask to an inverted binary map. The changes were made to opacity color values and image format followed by an inverse of the map to depict the mask as foreground (white color).

First, the fully transparent background pixels are made opaque (255) and pixel values which satisfy the alpha = 0.5 threshold are made black (see Fig. 3.19 for binary map generation). Next, a binary map is inverted for the fetal mask to appear as a foreground object represented in white. For simplicity, the inverted binary mask is called a *binary fetal mask*.

3.5.7.2 Boundary Detection

This section focuses on the extraction of a fetal edge which later serves as an outline for 3D reconstruction. The binary map is not suitable for direct use due to the need for refinement, which necessitates the implementation of boundary detection.

To identify the end points of a fetal mask and extract them from a binary fetal mask, the OpenCV edge detector algorithm *Canny* is used (see Fig. 3.20). The function discovers edges in the input image (8-bit input picture) and marks them in the output map edges. In addition, the Canny algorithm takes two arguments which include minimum and maximum threshold values for edge linking. Any edges with gradient values below *threshold1* are suppressed, while edges with gradient values above *threshold2* are considered strong edges.

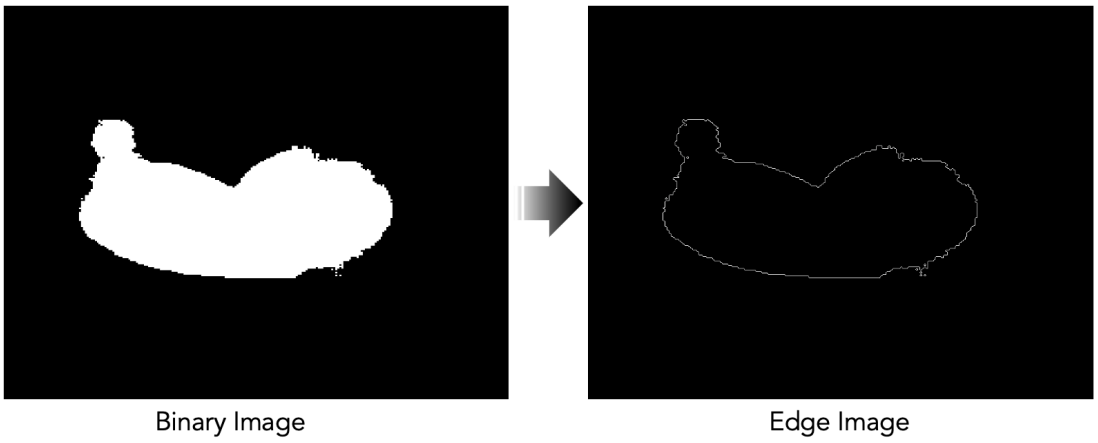


Figure 3.20: Fetal mask edge detection.

The Canny algorithm works by considering gradients in intensity values. It looks for areas of rapid change in intensity (typically correspond to edges). As our mask has already been converted to a binary mask, we set *threshold1* to 245 (values below are suppressed) and *threshold2* to 255 (strong edge). The highest value (255) is used to locate the beginnings of strong edge segments. The transition between the white foreground and black background represents a change in intensity which is detected as a final edge (depicted in Figure 3.20).

3.5.7.3 Edge Refinement

This section provides a brief overview on the different approaches used to refine the extracted edges of fetal masks.

The aim is to analyse the edges and identify regions that may result in an incomplete fetal reconstruction. The root problem of a poor reconstruction are the false edges, where the boundary of an object detected does not mimic the real shape of the fetus due to an incorrect initial segmentation (see Fig. 3.20, *Edge Image* example). The ideal course of action would be to retrain the segmentation algorithm entirely and focus on producing clear edges of CRL masks. Our method focuses on post-processing of already generated binary masks where the mask edges are extracted and manipulated to reduce the occurrence or eliminate false edges from being included as part of a final 3D reconstruction.

A number of image processing techniques specifically related to refinement of mask edges were considered in this chapter. The approaches include a variety of morphological transformations (described in section 3.5.7.4) and edge refinement algorithms (sections 3.5.7.5, 3.5.7.6, 3.5.7.7). The application of each approach and its methods are described next followed by their use as part of our method.

3.5.7.4 Morphological Transformations

Morphological operations are classical image processing operations used to reduce noise or smooth mask edges. Whilst normally the transformations are performed on binary images with existing foreground, we use them to refine edges already extracted from fetal masks.

There are different types of morphological operations which achieve different results, the transformations used to prepare ultrasound images for 3D reconstruction include edge dilation and erosion, opening and top hat morphology (discussed below). The order in which each morphological operation is applied to the original image matters and if any of the operations are swapped, a different transformed masked edge is generated.

All operations take two inputs, an original image in need of noise or outlier removal and a structuring element or kernel. Two basic morphological operators are erosion and dilation, followed by the variant forms such as opening, closing, top hat morphological operations and others [335].

Kernels and Custom Structuring Elements To perform any morphological transformation, a kernel or a structuring element needs to be selected. In theory, morphological operations can be applied using various structuring elements of different sizes and shapes, elements include squares, circles, or custom-shaped kernels. By adjusting the size and shape of the structuring element, the scale and type of structures that are enhanced can be controlled. Smaller structuring elements focus on smaller edge details, while larger kernels capture larger structures.

The common structuring elements are kernels with a rectangular shape. Depending on the image edges in need of transformation its is important to find balance between the noise removal and edge preservation.

Dilation Morphology Dilation is a morphological operation which adds pixels to the boundaries of objects in an image. In Figure 3.21 it can be seen that the dilation operator aims to expand the shapes of binary image edge.

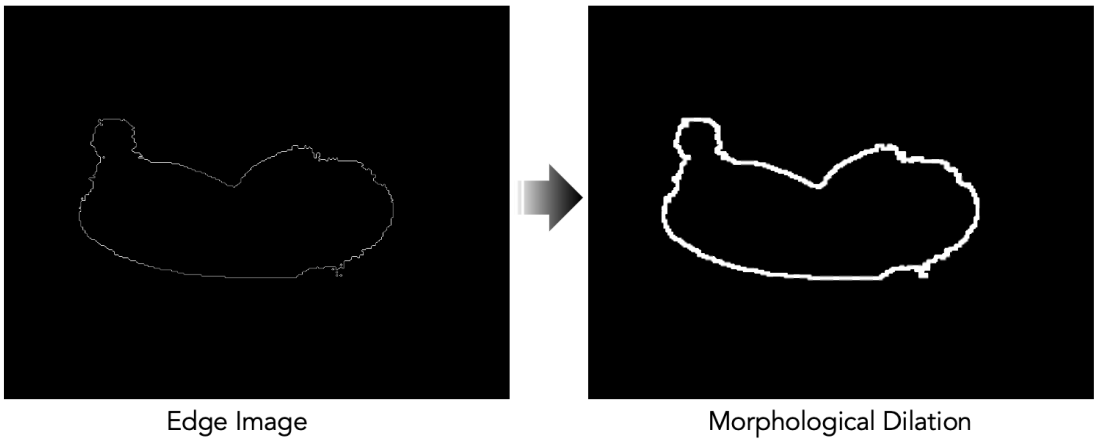


Figure 3.21: Morphological Dilation transformation.

As illustrated in Figure 3.22, the dilation operator applied to a binary image A where B_1 is the kernel defined as $A \oplus B_1 = \{x, y | (B_1)_{x,y} \cap A \neq \emptyset\}$.

The number of pixels added to the edges in an image depends on the size and shape of the structuring element used to process the image. During dilation, each pixel in the image is compared to its neighboring pixels within a defined neighborhood. The neighborhood is defined by a structuring element, which specifies the shape and size of the neighborhood. The most common structuring element is a square or rectangular window (i.e. 3x3 kernel), but other shapes like circles or crosses can also be used.

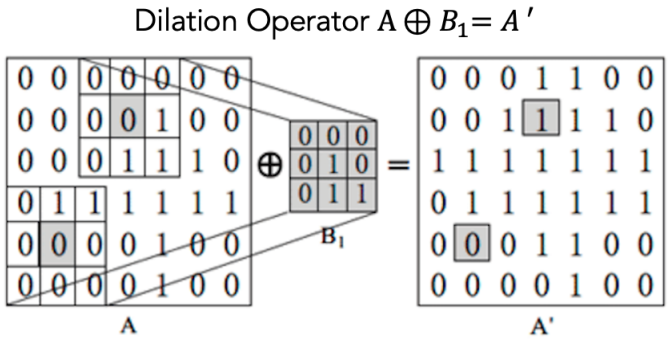


Figure 3.22: Morphological dilation operator of binary image A by structuring element B_1 [336].

Specifically, when performing dilation with a 3x3 kernel, the kernel is sequentially placed on each pixel of the image. The centre of the kernel aligns with the current pixel being processed in image A. If there is at least one foreground pixel (pixel with value 1) anywhere within the 3x3 kernel, the central pixel under the kernel will be set to 1 and no other pixels are affected.

The main purpose of applying a dilation operation first is to remove noise spots (gaps) surrounding the edges of masks into one thick outline. It is important to find the kernel size that best suits the task at hand. An 8x8 kernel size was used to fill the gaps between edges nearby in our work.

In addition to kernel selection, dilation can be applied to an image multiple times. That way, dilated regions are expanded further outward from the object boundaries resulting in a more aggressive dilation. Each iteration applies the operation using the previous result as the input. In this work, the dilation operation was not repeated to avoid over-inflating the fetal mask edge and loosing the important boundary surface details.

Erosion Morphology An erosion operation removes pixels from the boundaries of a foreground object. Figure 3.23 displays how morphological erosion disregards pixels near the boundary edge. The operator removes floating pixels and thins lines so that only substantive objects remain.

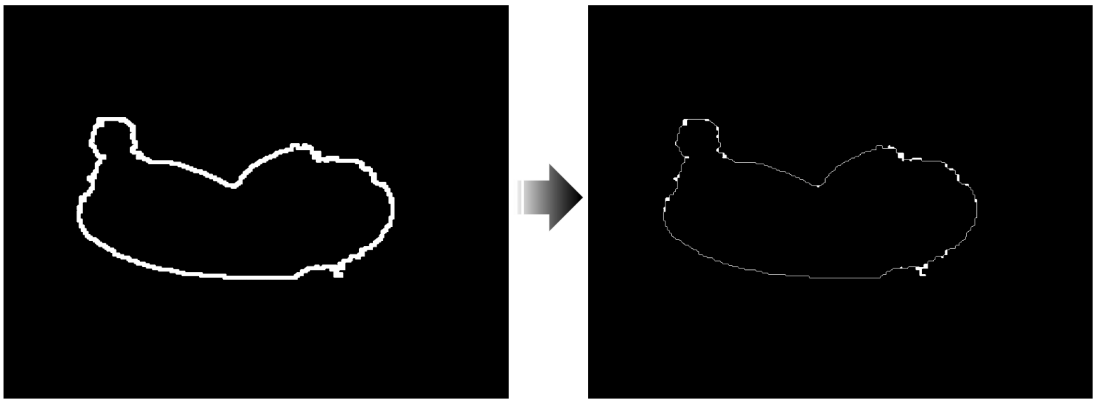


Figure 3.23: Morphological Erosion transformation. *Note:* Dilation operation followed by an image Erosion is called a *Close* operator described later in this section.

Figure 3.24 shows how the erosion operator of binary image A by structuring element B_2 is defined as $A \ominus B_2 = \{x, y | (B_1)_{x,y} \subseteq A\}$.

Erosion follows the same logic where a kernel slides through the image and erodes only the central foreground pixel. The distinct difference is that during erosion, the central pixel is eroded (set to zero) if any of the pixels under the kernel have a value of 0.

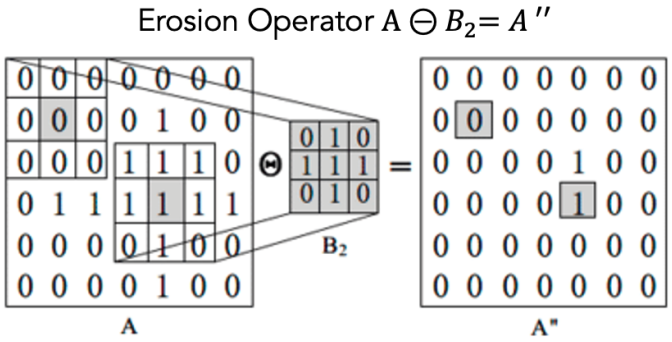


Figure 3.24: Morphological erosion operator of binary image A by structuring element B_2 [336].

With additional iterations, the eroded regions expand further into the object

boundaries, effectively shrinking the objects. Increasing the number of iterations will result in more aggressive erosion, causing objects to become smaller and thinner.

The order in which morphological operations are applied is important. If morphological erosion is applied before dilation (called *Open* operation morphology), the fetal mask (width = 1 pixel) would be erased entirely or in patches and important edge information would be lost.

To refine fetal mask edges, the key is to apply morphological erosion to a dilated mask edge using the same 8x8 size kernel. Such operation allows to thin the fetal mask edges back to the original width whilst retaining the outlier blobs (accumulated pixel noise). Such morphological order of operations applied to the original mask edge is defined as Closing morphology (described in more detail below).

Opening Morphology Opening is a combination of Erosion followed by Dilation. It is often used to remove small isolated noise and small unwanted details while preserving the overall shape and size of the objects. Morphological opening is performed using the same structuring element for both operations.

In essence, the morphological opening aims to remove everything but large contiguous patches of foreground pixels which is of essence to the next steps of the method used for edge refinement. The structuring element should be large enough to remove the lines during image erosion, whilst fully preserving the rectangles or blobs of connected edge outliers.

The number of iterations in opening determines how many times the erosion and dilation operations are applied consecutively. Increasing the iterations will enhance the noise removal effect but may also lead to the loss of finer details in the objects (as previously discussed). The operator is next used as part of a Top Hat morphology described.

Closing Morphology Closing is a combination of Dilation followed by Erosion. It is used to fill in gaps and small holes in objects, preserving isolated pixels that have a binary value of 1. The structural element is shared for both operations and the number of iterations follows the same rule as described in Opening morphology.

Increasing the iterations will fill larger gaps and holes but may also cause objects to become thicker and less defined.

Top Hat Transformation The Top Hat operation is a morphological operation that calculates the difference between the input image and Opening of the image. Recall, the opening of an image is obtained by applying an erosion operation followed by a dilation operation which results in isolated patches of foreground pixels.

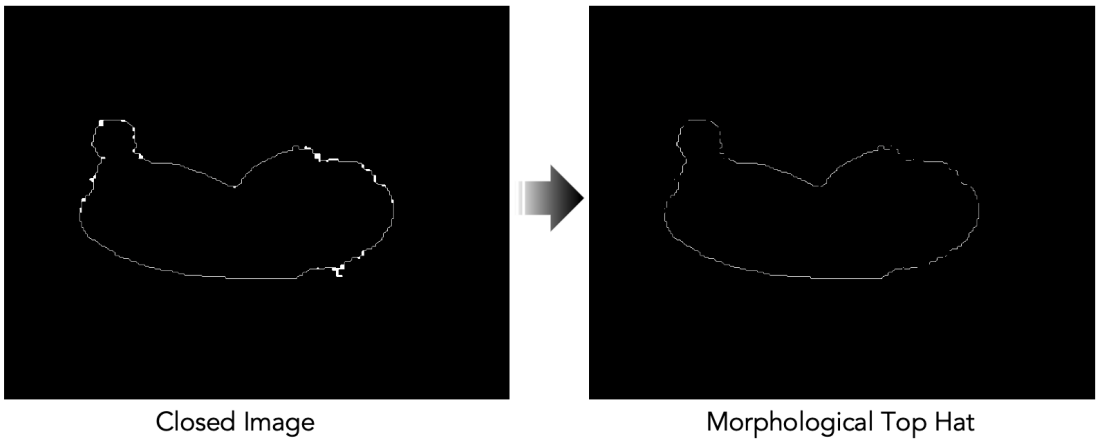


Figure 3.25: Morphological Top Hat transformation.

It is important to note that a 3x3 kernel size was used for the Top Hat transformation. From our experiments, when using the same size kernel as in Closing morphology (8x8 kernel size), the opening operator erases not only the edges of a fetal mask but also the large outlier blobs needed for further processing leaving the image blank. Figure 3.25 provides visualisation on how the Top Hat transforms an open image edge with a 3x3 kernel and subtracts the opened edge from the closed image (eroded image edge). In the Top Hat transformation, smaller structuring elements will highlight smaller details, while larger ones will capture larger structures.

Essentially, the Opening of a Top Hat operation removes the thin outline of a mask and downsizes the foreground patches, next only the patches are dilated back to the original shape and size and are subtracted from the closed image. The Top Hat resulting image contains the outline of a fetal edge minus the outlier blobs. Overall, the Top Hat operation is a useful tool in image processing for

retaining specific features or details whilst removing noise and outliers (considering the same morphology order is followed).

3.5.7.5 Contour Analysis

The purpose of contour analysis is to identify contours and extreme points to help eliminate spurious or disconnected edge segments, and later fill the gaps between the edges of a fetal mask. Contours (or mask edges) can be defined as the split lines that form a fetal mask outline.

Retrieve Contours For the contours to be detected correctly the binary fetal edges are to be assigned with the maximum value (white). That way, the contours of the fetal mask are retrieved instead of the background and image edge. To identify and retrieve contours, an OpenCV function *findContours* is used. If simplified, the function detects the change in image color (from a white fetal edge to a black background) and marks it as contour. A number of parameters determine how the contours are retrieved and approximated.

Contour Retrieval Modes The OpenCV *findContours* function retrieves all contours that are present in an image, and these contours can be analysed and stored in a variety of ways. Depending on the task at hand, contour retrieval modes can provide information about the existing hierarchy which shows how various contours are linked to each other, as well as their relationship with other contours. Some contours can be nested inside other contours, i.e. Parent-Child relation where the outer shape is the *Parent* and inner shape as a *Child*. On the other hand, other contours can have a single edge outline. Therefore, it is important to analyse the existing fetal mask edges as well as select an appropriate mode of contour retrieval.

There are a variety of ways to retrieve contours, function arguments that were considered in this work are depicted in Table 3.1.

Independent of the contour retrieval type, the hierarchy of edges are internally labelled from bottom right corner outwards, i.e. OpenCV labels the hierarchy from the bottom of fetal head outwards.

Arguments	Retrieval Mode	
	List of all Contours	Hierarchical Relationship
RETR_TREE	Yes	Yes
RETR_LIST	Yes	No
RETR_EXTERNAL	No	No

Table 3.1: Contour retrieval modes vary from functions which provide a full contour list and take into account the full hierarchy (Parent-Child) of nested contours, to the functions which only extract extreme outer contours and no hierarchical relationship.

For our work RETR_EXTERNAL arguments is the most appropriate to use as ultrasound fetal mask edges are already extracted, therefore only the outer contours are retrieved and stored in a memory buffer for further analysis.

Contour Approximation Methods A contour approximation method specifies how many points should be stored to represent the contour for further manipulations. There are 2 ways to store contours, either all the contour points are stored in an array (CHAIN_APPROX_NONE) or the mask edges are approximated and only the end points are stored (CHAIN_APPROX_SIMPLE).

From our experiments, storing all contour points at once yielded more information on how to later connect the edges of a fetal mask together. With the CHAIN_APPROX_SIMPLE approximation method, a number of contour points were missed causing a fetal boundary to be patchy (contain gaps) and incomplete.

Contour Noise Removal To remove the leftover pixels (or 'breadcrumbs') around the fetal mask contours (a byproduct of morphological operations) and prevent it from being used to reconstruct the fetal edge later, contours with less than 3 pixels were removed and a new fetal outline generated. Figure 3.26 shows how the retrieved fetal mask contours are transformed with outliers removed (see the zoomed-in pink frame image comparison) and extreme points (in different color dots) identified. The fetal contour transformation was zoomed in to show

how the outliers (extra pixels outside of the contour) were removed (pink frame transformation from left to right shows removed noise).

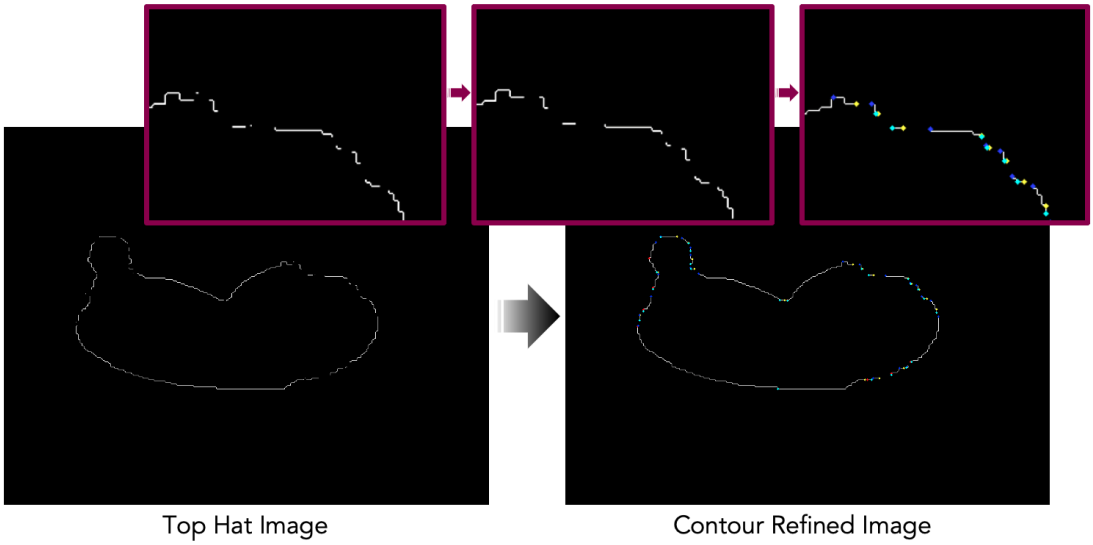


Figure 3.26: Fetal mask contour refinement. Top Hat image is transformed to remove noise and outliers (the transition is depicted in zoomed in pink frame images) and extreme contour points are identified with color dots.

3.5.7.6 Edge Linking

Now that the extreme contour points are identified, the aim is to find the closest distances between the neighbouring contours.

All the contour edges showed in Figure 3.26 are stored in a memory buffer and can now be used to compare the distance between the contours themselves. Specifically, there are multiple contours which if combined represent a fetal outline, the distance between every extreme point of a contour with those of every other contour is compared. Next, a line is drawn between points with least distance (shown in Figure 3.27).

Prior to connecting the edge points, it is important to set a *maximum distance threshold* between the contour points to limit the exposure of a current edge to a contour on the opposite side of a fetal mask. If the threshold between edges is too high, connecting edges with shorter distances would be missed entirely.

The distance between points is calculated using linear norm which measures the length between two vectors. At first iteration, the maximum distance between

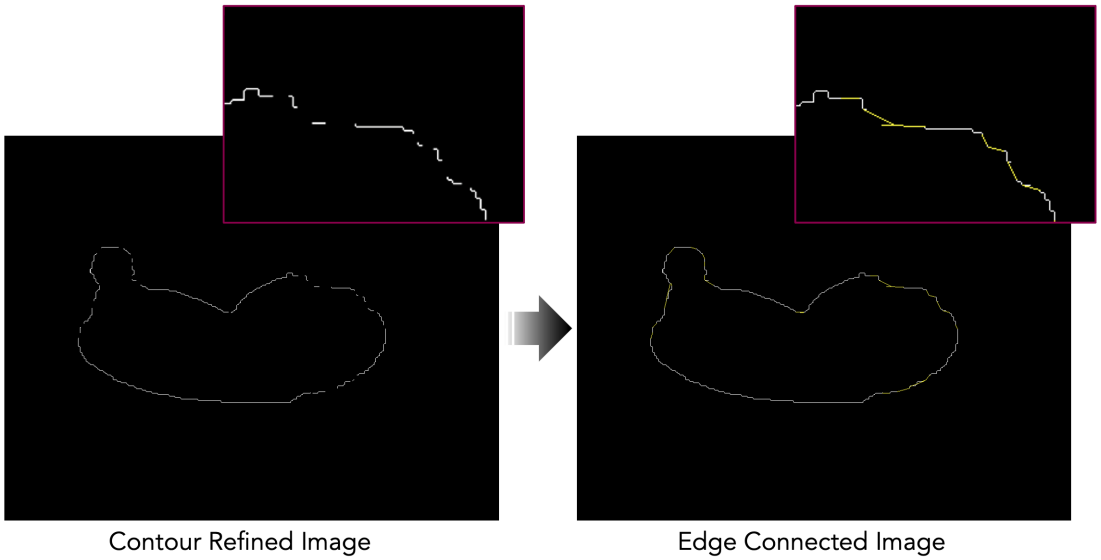


Figure 3.27: Fetal mask edge linking. The fetal mask contour is transformed by linking the extreme contour points to its neighbouring contours (shown in yellow).

contour edges is set to a length of 40. Next, the contour edges with the shortest distances are prioritised and stored in a memory buffer to be filled first. This encourages the edges to connect to the nearby contours. Each contour point can only be paired once.

Finally, once all the extreme contour points are paired, the OpenCV function *drawContours* is used to connect the edge points by drawing a line between the paired edges. The thickness of lines connecting empty edges can be adapted to an existing contour thickness.

3.5.7.7 Complete Edge Refinement Method

This section discusses the final post-processing method used to refine the edges of segmented fetal mask contours.

As the final image with connected fetal edges (discussed in section 3.5.7.6) still contains some noise with rough edges, it requires additional post-processing in a form of edge refinement to create a smoother contour boundary. The process involves the same operators used to perform transformations described above. The change is made to kernels used to perform morphology operations and to a *maximum distance threshold* applied when dealing with gaps between edges.

The full order of transformations include four iterations displayed as green blocks in Figure 3.28. The same transformations can be visualised in Figure 3.29 where major changes to the fetal edge are displayed at each iteration.

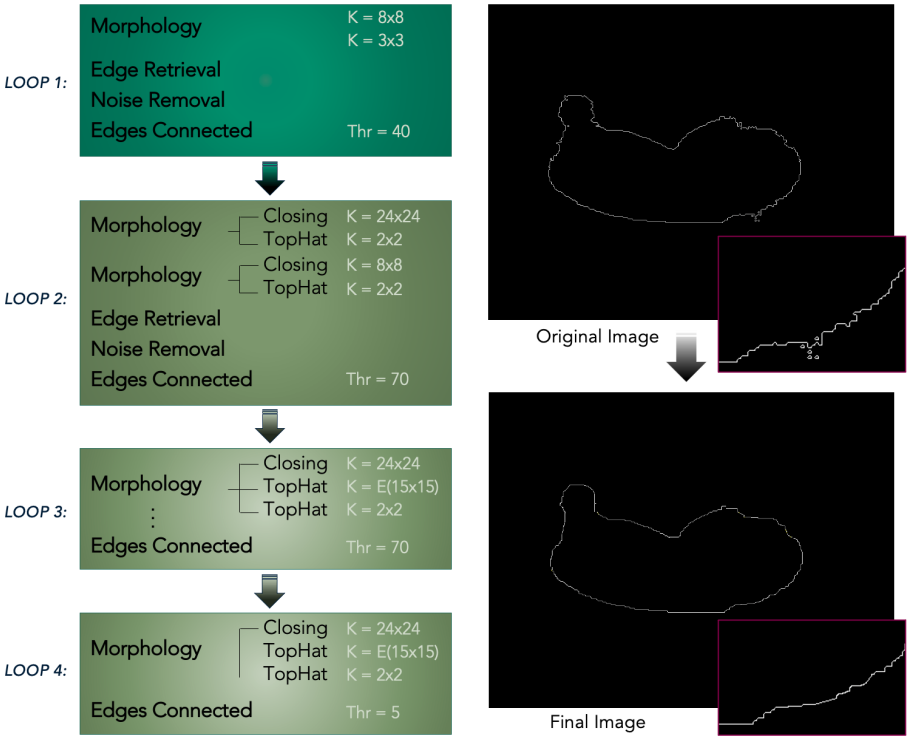


Figure 3.28: The order in which an original image edge is transformed to a final post-processed noise-free fetal outline. On the right, fetal contour before the edge refinement and after are presented as original and final images, respectively.

The first iteration includes morphological transformations such as closing (dilation followed by erosion) with an 8x8 kernel size and tophat with 3x3 kernel (a smaller size kernel allows for a more refined noise removal). The transformed top hat edge is used for contour analysis where the extreme edge points are identified. The noise is removed from the edges of a fetal mask. Next, refined edges are connected using a maximum distance threshold of length 40.

The second iteration involves, the same order of transformations, except now the kernel size used to perform image closing is 24x24 and the top hat operator has a 2x2 kernel (the transformation can be visualised in Fig. 3.29, L2: *Top Hat Image 1*). The morphology combination is performed twice to refine and smooth the edges and generate one single contour. Edge detection and noise removal are

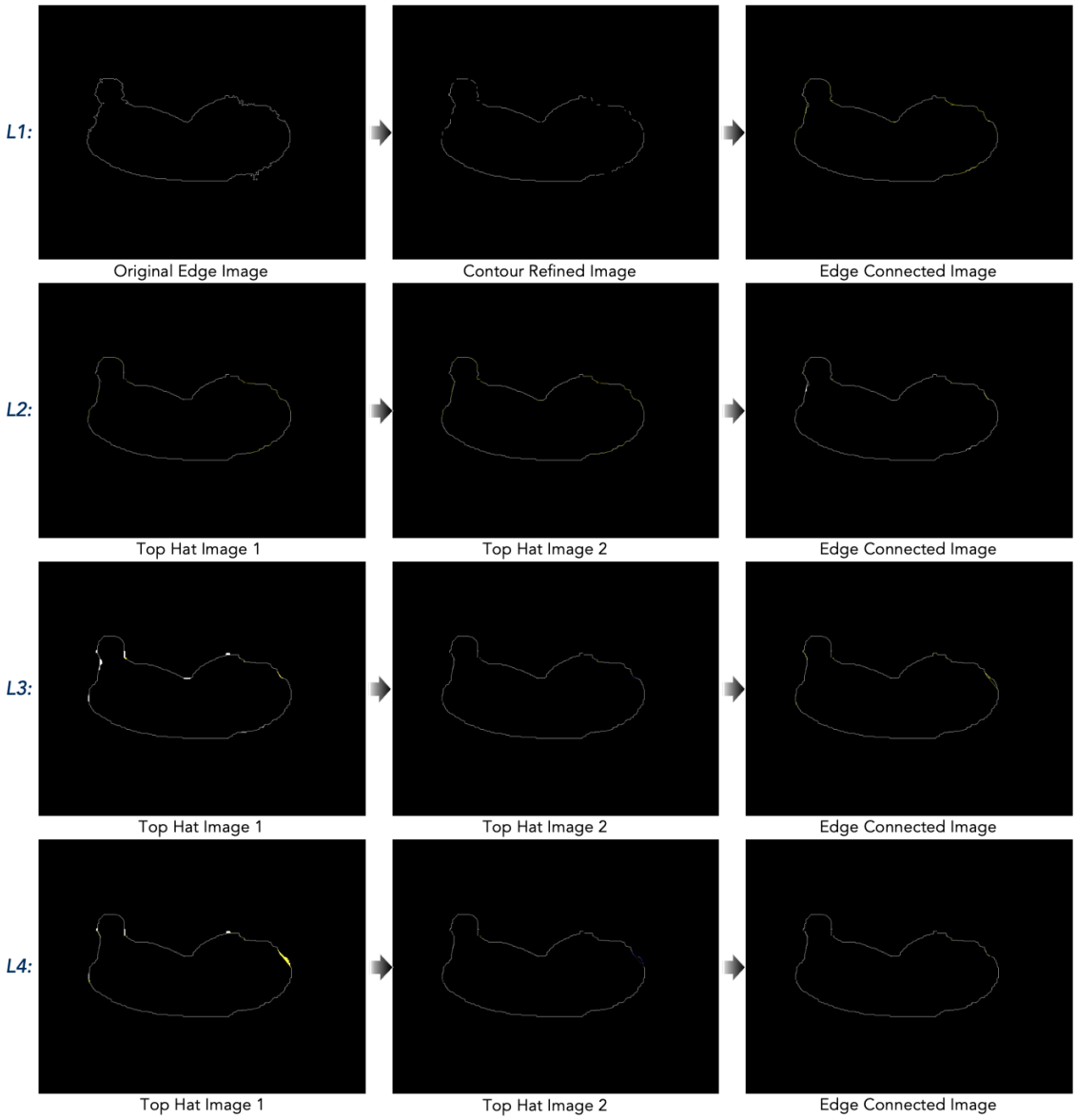


Figure 3.29: A display of major transformations made to a fetal mask edge. Rows represent four iterations (following the order described in Fig. 3.28) and columns are the major transformations made to the fetal outline.

performed in all iterations. Finally, the maximum distance threshold is increased to 70. In some fetal mask edges, the noise would prevail more which leads to a higher number of outliers removed causing larger gaps between the edges to appear. Hence, as a solution a higher maximum threshold is implemented during the second iteration, without affecting the neighbouring contour edges which are connected during the first iteration.

The third and fourth iterations contain the same transformations. In the last

iteration the final threshold value is changed to a length of 5 to account for small gaps. The main change has been made to a kernel of the first tophat operator which is represented by a 15x15 ellipse-shaped structuring element. Instead of the common square-shaped kernel, an ellipse-shaped structuring element better erases the blobs created by morphological erosion.

Section 6.5 applies ultrasound probe to corresponding 2D fetal mask edges with final fetal reconstruction in 3D space presented in section 6.6.3).

4

Stochastic Augmentation Policy Search: Single Frame Saliency Prediction

Contents

The lack of curated data in the medical imaging field is a common obstacle to development and clinical implementation of effective and efficient models. Datasets of any size may present high class imbalance as well as general shortage of data for a particular class which can be difficult to tackle with existing methods. This chapter presents a stochastic augmentation policy search strategy applied to a multi-modal first-trimester fetal ultrasound dataset. The augmentation strategy

aids in prediction of sonographer eye gaze which can be used as a tool to guide clinicians to important anatomical structures. The research presented in this chapter is based on a manuscript that has been published in:

Elizaveta Savochkina, Lok Hin Lee, Lior Drukker, Aris T. Papageorghiou and J. Alison Noble. First Trimester Gaze Pattern Estimation Using Stochastic Augmentation Policy Search for Single Frame Saliency Prediction. In an Annual Conference on Medical Image Understanding and Analysis (MIUA) (pp. 361-374), 2021 - Oral presentation.

4.1 Introduction

The work in this chapter focuses on the imaging guidance task for a first-trimester fetal ultrasound scan, specifically automating the task of predicting where a sonographer should look next. Two modalities were used which include first trimester ultrasound scan videos and real-time gaze tracking data. Recall from section 3.3, a task of predicting saliency in a spatial domain refers to sonographer's visual attention and is modelled on static ultrasound video frames. The model learns static visual attention by treating each US video frame as an independent image.

To train a deep learning network and predict sonographer eye gaze, the dataset needs to be sufficiently large and curated, include expert labeling (i.e. annotations) and be representative of all the US anatomical views (to avoid bias in a network). Datasets of any size may present with high class imbalance and, in the field of fetal ultrasound, it can be presented by little variation of US anatomical views. Class imbalance may lead to a model bias towards the majority class. Medical data can also present with a general data shortage toward a certain class, i.e more examples of CRL anatomy than the brain, causing an unfair and poor performance of a neural net. Poor performance may be seen when a network makes unsatisfactory predictions of sonographer gaze on one type of a fetal anatomy and fails to do so on the minority class. The fetal dataset used in this work was of average size (45,630 US frames, refer to Fig. 3.1). Whilst based on the focus of the first trimester scan - to confirm the gestational age and check for Down's syndrome

(refer to section 1.1), the main biometry plane views are CRL and NT (see Fig. 3.6). Hence, a shortage in US data was expected.

To address challenges associated with data limitations, including both imbalance and shortages in specific classes, re-sampling and cross-validation are common strategies. However, it can be challenging to apply these approaches effectively with limited computational resources and without human annotations. Medical image human annotations are costly to acquire and require resources such as human expertise making it difficult to have access to diverse large medical data.

At the time of conducting work on this chapter, we had no human annotations of biometry planes or anatomical labels (except for the technical annotations input by sonographers into the US machine as part of the scan). The availability of fetal frame annotations was limited as only ultrasound freeze-frames (refer to Fig. 3.7) were labelled for the NHS record and not the rest of a scan. To avoid the labor-intensive task of manually annotating US frames or implementing an unsupervised clustering model, we adopted a stochastic data augmentation policy search method. The aim of work in this chapter was to find a method that could address the shortage of data in specific classes in an automated manner which was hypothesised to be the cause of over-fitting and limiting the model's generalization.

Data augmentation is widely used to increase the size and the diversity of the training set. We have investigated the benefits of data augmentation in medical imaging to combat class imbalance [337], increase model generalization [338–341], alleviate over-fitting [342] and expand training data [343, 344]. It is a common approach to use small augmentation transformations such as scaling, rotation and translation [47] that make training images look natural and realistic. However, recent works [46, 48–50] have found that different types of transformations that cause images to look less realistic lead to improved generalization. We hypothesised that the addition of distorted images in training can improve model generalization performance, where the augmentation process generates artificial images with additional fetal appearance and position characteristics as well as ultrasound artifacts/noise that inevitably appear during the scan.

Our work included the application of US video frames and real-time gaze tracking data to guide clinicians to important anatomies. Hence, the chapter is also inspired by cognitive science applications where gaze-tracking data can assist the analysis of a fetal ultrasound scan [28, 142, 332, 345]. We utilize sonographer gaze, in the form of gaze-tracking data, in a multi-modal imaging deep learning framework to assist the analysis of a first trimester fetal ultrasound scan. Specifically, predicting sonographer gaze can be useful for identification of spatio-temporal patterns that are important for US scanning. Hence, we combined the information from sonographer eye gaze data and corresponding US video frames for prediction of saliency maps (refer to section 3.9). The fetal ultrasound and gaze data used in this chapter are described in Chapter 3.

The primary contribution of this chapter is methodology of data augmentation strategies, therefore, the main focus was a transformation of training examples to improve saliency prediction. Specifically, the aim was to alleviate over-fitting and improve the segmentation performance when predicting saliency of sonographers. Prior to adoption of the augmentation policy search method, we trained a standard encoder-decoder model with common augmentation transformations mentioned in section 3.3.3. The network started to over-fit after the 5th epoch generating very poor predictions of saliency maps.

From section 2.1.2, data augmentation is not the only technique that can reduce over-fitting, many other strategies for increasing generalization performance focus on model's architecture itself. The techniques include model pre-training, transfer learning and regularisations such as dropout and batch normalization. The latter two techniques have been implemented (see Fig. 4.5) prior to data augmentation and provided no improvement to performance of the model. In contrast to the techniques mentioned above, data augmentation approaches over-fitting from the root of the problem, the training dataset.

We show that the dataset of our size benefits from automated data augmentation where the proposed augmentation strategy reduced model over-fitting, improved generalisation and outperformed all conventional ultrasound augmentation strategies

on all saliency metrics. Using the learnt policies, our models outperform the baseline: KLD, SIM, NSS and CC (2.16, 0.27, 4.34 and 0.39 versus 3.17, 0.21, 2.92 and 0.28).

4.2 Originality and Individual Role

The code for stochastic augmentation policy search method was initially created for classification of second-trimester standard biometry planes by Lee Lok Hin, a member of the University of Oxford Institute of Biomedical Engineering. The code mimicked the original data augmentations by Cubuk et al. [46]. I adapted the code which only dealt with image manipulation to support first-trimester saliency map segmentation and incorporate two distinct modalities: fetal ultrasound videos and gaze-tracking data. Specifically, I developed the entire architecture code from scratch and extended the existing code of a fellow colleague to enable processing and segmentation of gaze-tracking saliency maps. I processed all frames from the PULSE project first trimester ultrasound dataset and all the eye-tracking data recorded by an eye-tracker. Freeze frames were annotated by 8 experienced sonographers.

4.3 Stochastic Augmentation Policy Search

We looked to incorporate two main augmentation strategies that would best fit our main purpose of predicting saliency in the first trimester ultrasound scans. Previous works on automated augmentation strategies [49, 54–56], at the time, demonstrated top results in image classification and object detection. However, these methods require a separate and expensive search phase and make it challenging to adopt a policy on a large scale. In this work, a grid search approach is used as part of the policy search algorithm which explores different combinations of augmentation transformations applied to the data, aiming to find an optimal set of augmentations to improve model performance.

Recall from section 2.1.2, the RandomAugment (RA) paper [46] introduces a search space with the regularization strength that can be tailored based on model and dataset size. RA method was found to significantly reduce a search

space and hence, reduce computational overload. RA is one of the augmentation strategies that is used in this chapter.

Our dataset has a high anatomical view variation amongst several types of fetal planes (i.e CRL and NT). However, not all the views are represented well (refer to section 4.1) which poses a problem of data imbalance and data shortage where a network is biased towards a better represented class. Dataset shortage for a specific anatomy can also appear due to ultrasound artifacts and certain underrepresented fetal characteristics that are a result of a fetal position, variable maternal gastric wall thickness, and sonographer expertise. As the dataset used in this chapter is medium in size, there is a higher demand for manual data cleaning and a higher chance that distorted and noisy images occur.

'Mixup' [346] is a data augmentation technique that generates a synthetic training example using a weighted combination of random image pairs from the training data. It is a technique that was claimed to alleviate sensitivity to distorted examples, improve robustness and generalisation of a trained model [346, 347]. As a byproduct, Mixup reduces over-fitting. As our task is specific to medical imaging and it is common to see US image distortion (due to a small size but frequent movement of a fetus, it can be difficult to capture a clear image of anatomy); the Random Augmentation policy search can be extended from a single-image transformation strategy such as RA and include a mixed-example image strategy adopted from [346]. Lee et al. [48] combined RA transformations and included non-linear mixed examples for training a standard plane classifier. The data augmentation strategy was named Mix.RA which is later used in our research.

The methods mentioned above adopt automatic augmentation strategies for a single modality. We employ Random Augmentation and Mixup strategies for a multi-modal dataset where fetal US images and their corresponding saliency maps undergo the augmentation transformations. Not all transformations can be applied to gaze-tracking data which is further discussed and analysed in section 4.3.2.

Figure 4.1 presents an overview of the proposed method which consists of data generation, data augmentation, grid search and an encoder-decoder network for

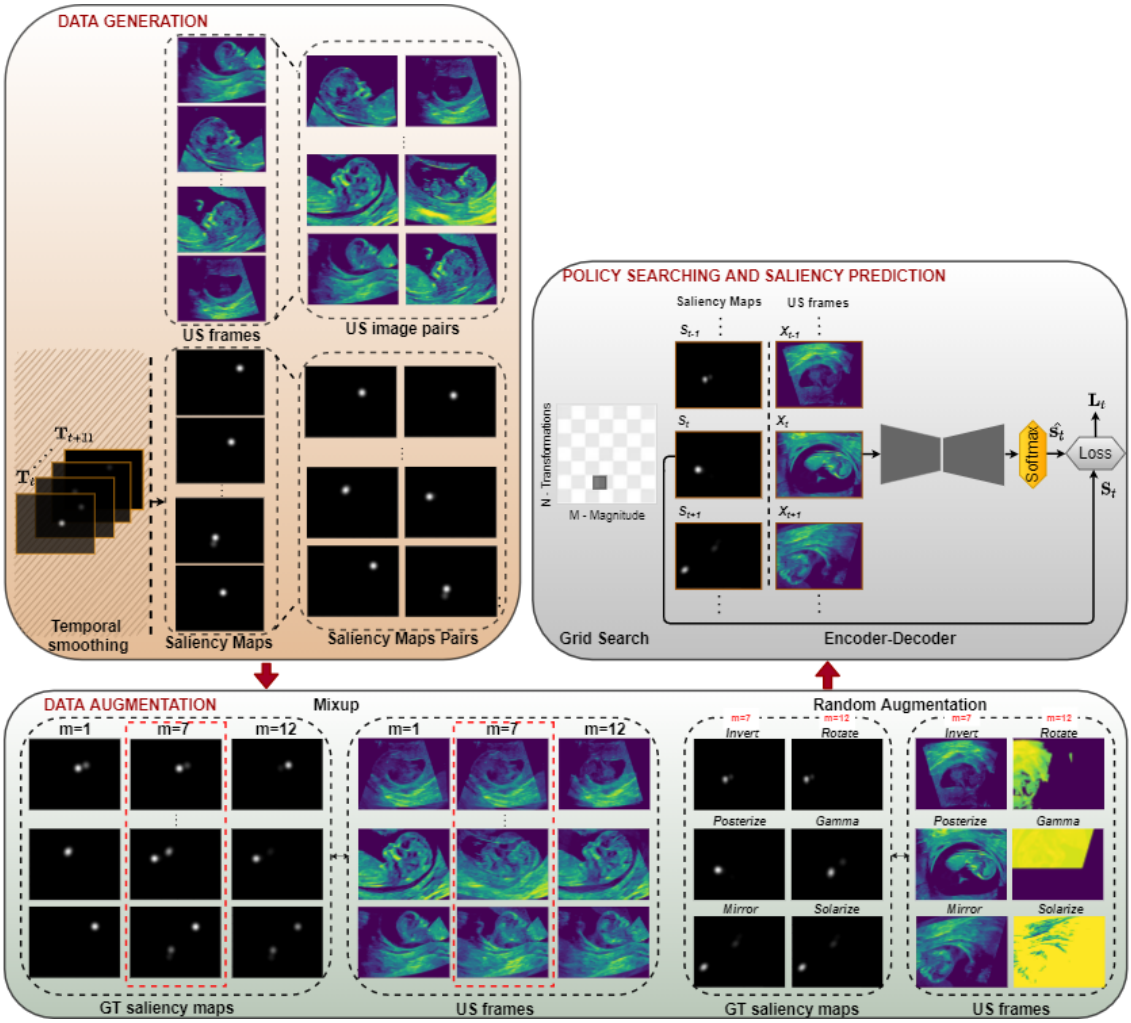


Figure 4.1: Overview of our proposed architecture. The method is divided into blocks for better visualization starting from data generation, followed by data augmentation, policy searching and saliency prediction.

saliency prediction. Data generation details are described in Chapter 3, whereas data augmentation strategy and policy searching with saliency prediction are discussed below. As part of the encode-decoder network, a single US image is fed as an input into a network, whilst the saliency maps generated from sonographer eye gaze are used as ground truth for the model.

Inspired by Cubuk et al. [46], we adopt a simple grid search with fixed magnitude schedule and a total of $K = 19$ transformations which include additional non-linear transformations by Lee et al. [48]. Each augmentation policy is defined by n, which is the number of transformations from K an image undergoes, and m, which is the

magnitude distortion of each transformation. These transformations are then applied to the mixed-example images [346] with which we share the m hyperparameter. Henceforth, we refer to RandomAugment with non-linear transformations as Random Augmentation (RA) strategy and RA strategy with Mixup as Mix.RA.

4.3.1 Mixed-Example Data Augmentation

The Mixup augmentation technique is based on taking convex combinations of pairs of examples and their associated labels [346]. Convex pairs refer to linear combination of points where all coefficients are non-negative and sum to 1. The Mixup technique smooths out network performance using linear interpolation of image feature vectors. Our original dataset $D = \{(X_i, Y_i)\}$ consists of a series of i ultrasound frames X and their corresponding ground truth saliency maps Y. In order to employ the Mixup, we generate two sets of US and GT saliency map images at random and pair them into a dataset, D_p.

$$D_p = \{(x_1, x_2)_{\frac{i}{2}}, (y_1, y_2)_{\frac{i}{2}}\} \tag{4.1}$$

where, (x_1, x_2) and (y_1, y_2) represent an US image and a GT saliency map image pairs, respectively. The training distribution is extended using element-wise weighted averaging of two random examples, denoted as $\frac{i}{2}$. That way, each image in a pair receives a weighted average score, increasing or decreasing the effect the image has on a final mixed pair (displayed in Fig. 4.2).

Although linear intensity averaging does not produce realistic examples of US images, the artificial image generated has a smooth transition from two images into one. In contrast, alternative non-linear methods [50] slice image features into blocks making the edges of the mixed images to appear which is not a good representation of distorted examples in the test set and defeat the method benefits for our application.

As we pair US images, their corresponding GT set of saliency maps undergo the same transformation. After concatenation, the resulted US image and GT saliency map (SM) are denoted as $\tilde{x}$ and $\tilde{y}$ illustrated in Fig. 4.2.

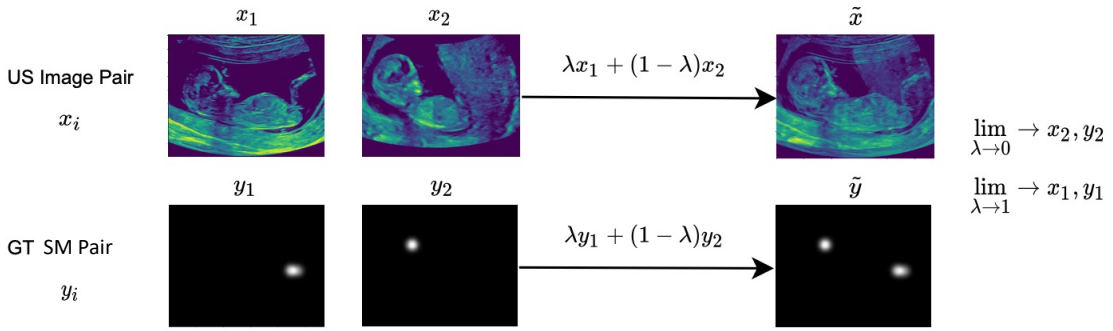


Figure 4.2: The procedure for linear mixed-example data augmentation using US image pair and corresponding GT saliency map image pair and the final artificial mixed-example image pairs.

Formally, given initial US images $x_{1,2}$ and their corresponding GT saliency maps $y_{1,2}$, the generated artificial mixed-example images $\tilde{x}$ and $\tilde{y}$ are:

$$\begin{aligned} \tilde{x} &= \lambda x_1 + (1-\lambda)x_2 \\ \tilde{y} &= \lambda y_1 + (1-\lambda)y_2 \end{aligned} \tag{4.2}$$

where $\lambda \sim Beta(\frac{m}{10}, \frac{m}{10})$ for each pair of examples and m is a learned hyperparameter varied from $0-12$ (explained below). The Beta distribution is chosen over other distributions (i.e. Normal distribution) as it represents the random behaviour of proportions and percentages, i.e. the probability is a random variable and not a parameter in a distribution [51]. Using a Beta probability distribution function (PDF), a random variable λ is approximated as:

$$f(\lambda; \alpha, \beta) = \frac{1}{\mathcal{B}(\alpha, \beta)} \lambda^{\alpha-1}(1-\lambda)^{\beta-1} \tag{4.3}$$

where the beta function, $\mathcal{B}$, is a normalization constant that ensures the total probability is 1. The distribution has two shape parameters $\alpha, \beta > 0$, that appear as exponents of the random variable and manage the shape of the distribution. The Beta distribution is continuous and is set on the interval $\lambda \in [0, 1]$.

In an augmentation policy, m represents the magnitude distortion of an image transformation. In this Chapter, the range of a learnt hyperparameter is $0 \leq m \leq 12$. To better explain the mathematics behind the mixed-example data augmentation, we start from $m = 100$.

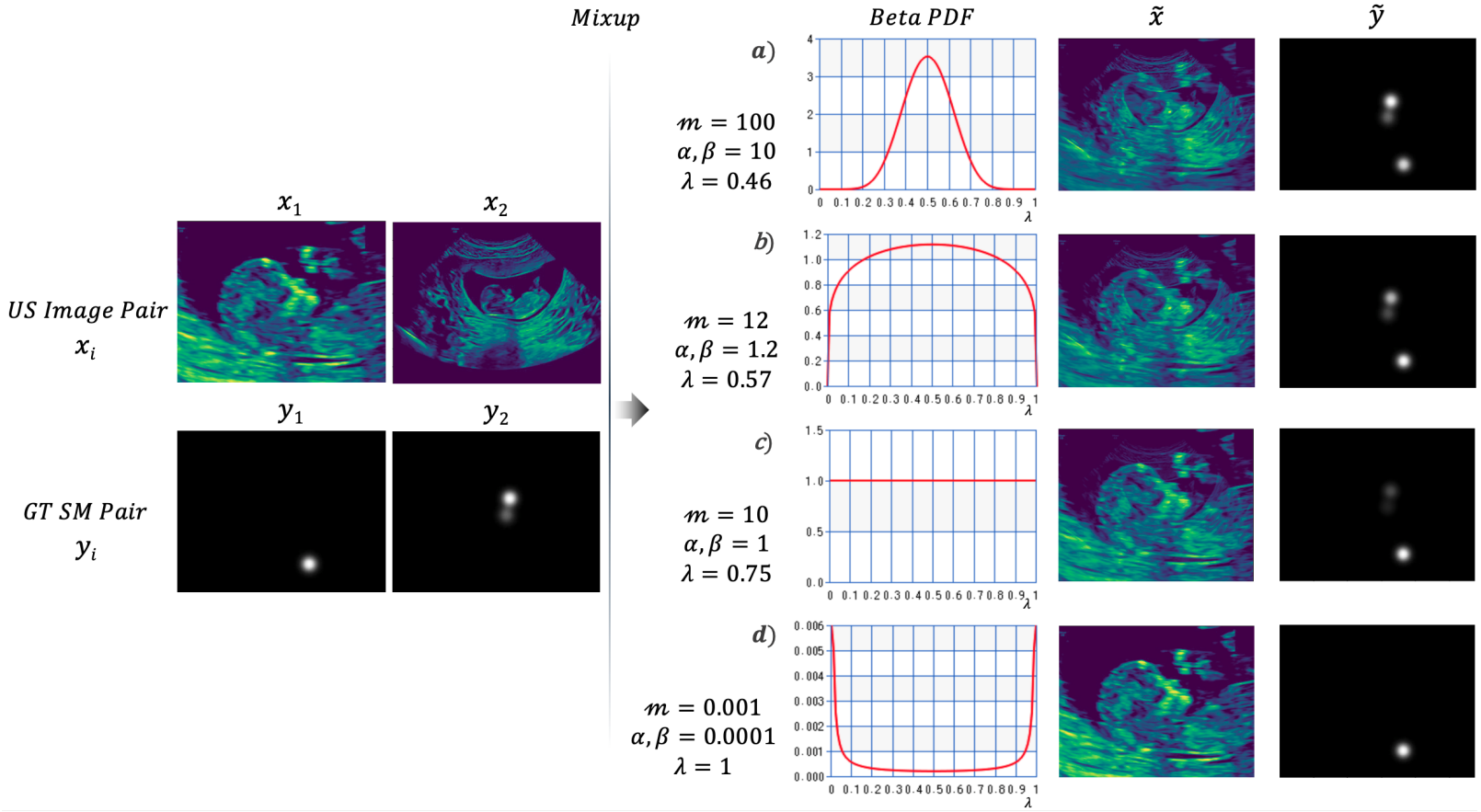


Figure 4.3: A more detailed illustration of how mixed-example images, $\tilde{x}$ and $\tilde{y}$, are generated using Beta probability distribution function (PDF). Based on the image distortion intensity value m, the Beta distribution parameters (α, β) determine the shape of a Beta PDF curve. A variable λ is drawn from the Beta PDF curve at random which defines the degree of mixup each image has in a pair; λ is applied to ultrasound and GT saliency map image pairs, x_i and y_i (see Equation 4.2), to determine the final mixed-example image. Each Mixup scenario is described below.

As $m \to 100$, the value of shape parameters $\alpha, \beta = \frac{m}{10} \approx 10$ (see Fig. 4.3). Higher Beta distribution parameters present a higher probability of image distortion due the Beta PDF curve transforming into a Normal distribution (bell-shaped) curve. With higher values of m and hence, α, β, the shape of a parameterized Beta distribution (in this case bell-shaped curve) becomes narrower and centres around $\lambda = 0.5$ where $0 \leq \lambda \leq 1$ (displayed in Fig. 4.3a). The width of the area under the curve (λ-axis) defines the boundaries from which the Mixup variables can be drawn at random.

As $m \to 100$, the Mixup random variable $\lambda = 0.46$ is drawn from the Beta PDF and applied to US image (x_1, x_2) and corresponding GT saliency map (y_1, y_2) pairs (refer to Equations 4.2). As λ value is close to 0.5 (with slightly less weight given to x_1, y_1), each image weights equally the same, producing the most distorted mixed-example images $\tilde{x}$ and $\tilde{y}$.

When $m = 12$, the Mixup random variable $\lambda = 0.57$. Whereas, when $m = 10$ both Beta distribution parameters $\alpha, \beta = 1$ represent a uniform distribution with random variables (i.e $\lambda = 0.75$ in Fig. 4.3c) drawn from $[0, 1]$ interval. Finally, the lower the magnitude distortion m is, the higher the probability of one image prevailing over the other. When $m = 0.001$ ($\alpha, \beta = 0.0001$) the Beta PDF curve peaks at $\lambda = 0$ or $\lambda = 1$. When a random variable $\lambda = 1$, images x_1, y_1 override x_2, y_2 where the NT measurement plane and a corresponding gaze are displayed as the mixed-example images.

4.3.2 Random Augmentation

We employ a random augmentation strategy, RandAugment, to tackle over-fitting which we believe arose from the limited number of 1st trimester images of abdomen, brain, heart, spine and other as well as from the structural differences in the anatomical views. Structural differences refer to variations in the appearance, orientation, and spatial arrangement of anatomical features across different ultrasound frames. These differences can lead the model to favor certain views over others, especially if some views are more prominent in the training data, causing it to overfit to those more frequent structures.

RandAugment helps address this issue by applying a range of random transformations, such as rotations, translations, and contrast adjustments, which introduce synthetic variations across the dataset. This strategy not only makes the model more robust to diverse anatomical views by simulating multiple ways each structure can appear but also effectively increases the number of examples for underrepresented anatomical views. By applying these transformations to frames representing less common views, the presence underrepresented classes in the dataset are artificially expanded, helping to mitigate the impact of structural differences without explicit class annotations. However, without class annotations for each ultrasound (US) frame, we are unable to directly balance anatomical views within the training and test sets. Thus, RandAugment serves as an alternative approach, adding variability and balancing examples indirectly, which aids the model in generalizing across different structural patterns in ultrasound images.

As mentioned in the overview of our proposed method, we adopt a grid search with fixed magnitude schedule and a total of $K = 19$ available transformations. Each augmentation policy is defined by n, which is the number of transformations from K that an image undergoes, and m, which is the magnitude distortion of each transformation. These transformations are applied to the mixed-example images with which we share the m hyperparameter. Following [48], the inclusion of non-linear transformations such as grid distortion, speckle and elastic transformation are beneficial to use on US images. Fig. 4.4 depicts 5 affine, 10 histogram and 3 non-linear transformations examples where each US frame and corresponding GT saliency map change with respect to the type of transformation. Affine transformations change the shape, position, and orientation of an image without altering its internal pixel intensity values; these include rotation, translation in x and y direction, shear in x and y directions. Histogram transformations adjust intensity values and color distributions in an image by inverting colors, equalizing intensity levels, adjusting color balance, posterizing, solarizing, and enhancing contrast, brightness, sharpness, auto-contrast, and gamma. The Identity transformation leaves the image unchanged. The remaining transformations are non-linear: grid distortion modifies the spatial

appearance of the image, speckle adds intensity noise, and elastic transformation deforms the image in a flexible manner.

All types of transformations are applied to US images, whereas, only affine and and non-linear transformations affect GT saliency maps (excluding the effect of speckle). In Figure 4.4, a star denotes when an augmentation is applied to the GT saliency map, transforming it in conjunction with the US image. To increase the number of training examples and introduce more anatomical variation, US input images can be changed in terms of color, shape and position. In contrast, GT saliency maps can only be adjusted in terms of position and shape (through elastic grid and shear distortions).

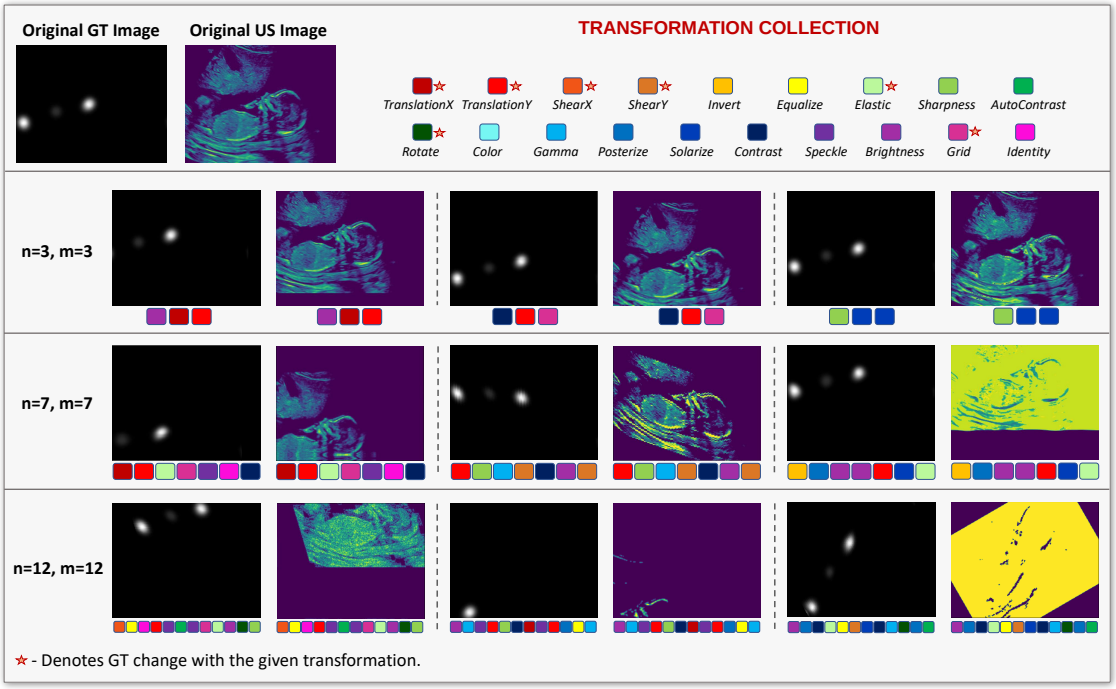


Figure 4.4: Examples of how each transformation and augmentation policy affect US input images and their corresponding GT saliency maps. Each color represents a transformation where original images (US and GT) are transformed using a number of transformations (n) at a magnitude of (m). ***Star***: denotes change of GT saliency maps with given transformations.

During augmentation implementation, US and GT saliency map images undergo n transformations at a magnitude m of augmentation intensity. Sometimes, a high number of transformations in conjunction with highly distorted images, can cause US images to go black (or all pixels to be of the same color) or the saliency

maps to be augmented outside of an image space (i.e all pixel values are 0). In such cases, if the augmentation of one of the images is not plausible, the image pair (US and GT saliency map) is discarded and the augmentation process is repeated on original images.

4.3.3 Model and Training Details

This section describes the structure of the model used for single frame saliency prediction and model's training details.

4.3.3.1 Encoder-Decoder Network

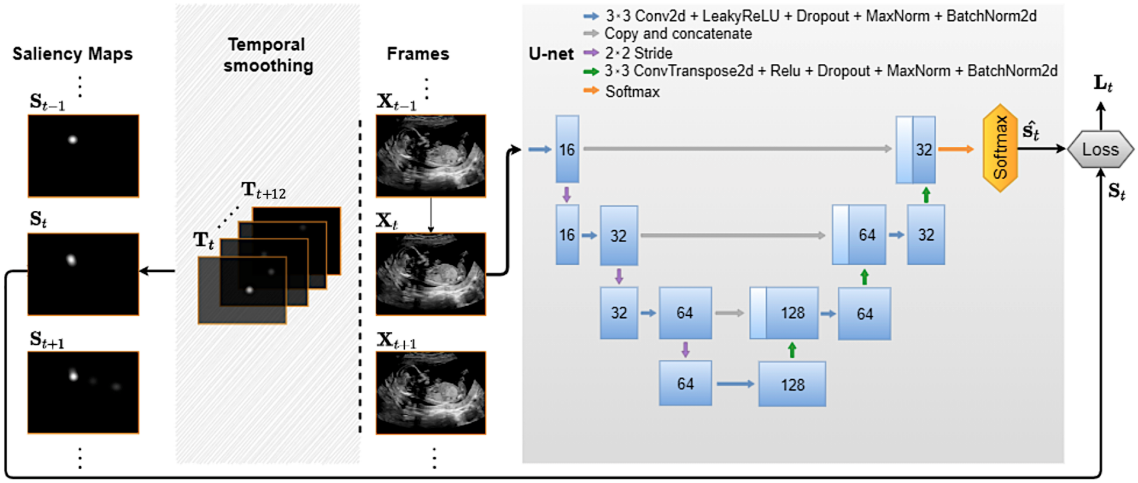


Figure 4.5: An encoder-decoder network for a single frame saliency prediction. The US input frames (X_t) and ground truth saliency maps (S_t) on the left are described in Chapter 3. The loss (L_t) and saliency prediction ($\hat{s}_t$) are described in section 4.3.4.

We utilize an encoder-decoder convolutional neural network with skip connections (illustrated in Fig. 4.5) to predict the visual saliency maps for each frame. The encoder consists of four (Conv → MaxNorm → BatchNorm → Relu → Conv → Dropout) blocks, with channels 16, 32, 64 and 128. The decoder consists of 4 (ConvTranspose → MaxNorm → Dropout → Conv → Conv) blocks, with channels 128, 64, 32 and 16. Each block has a dropout of 0.5, resulting in a model with 218k trainable parameters. As the primary focus of our work is on data augmentation strategies, the skeleton of the encoder-decoder network will not be discussed in detail.

4.3.3.2 Training

The models are implemented in Tensorflow 2.1 and image manipulations are performed with Pillow 7.1.2 and OpenCV 3.4.9 libraries. Each training run varied from 8 hours to 5 days on a single Nvidia GTX 2060 Ti with a total of 40 models run. Networks were trained up to 120 epochs with a batch size of 50 via adaptive moment estimation (Adam) with an initial learning rate of 0.001. For the description of input and ground truth frames used for training, refer to chapter 3.

Random horizontal and vertical flipping were used in all RA policies as a baseline augmentation. We also compare our augmentation policies with the augmentation policy in previous work [20, 27, 348] as an example of a conventional augmentation policy, consisting of random rotation with an angle uniformly sampled from [- 25, 25] degrees and random horizontal flipping. We denote this as the baseline augmentation policy. The performance of networks trained with augmentation was evaluated with values of n, $m = \{1, 3, 5, 7, 9, 12\}$ using a grid search to find optimal n, m values.

4.3.4 Saliency Map Prediction

Given an image and a gaze point set $(\mathbf{X}, G)\epsilon D$, we generate a visual saliency map $\mathbf{S}\epsilon[0,1]^{H_D \times W_D}$, where $S_{i,j}$ is the probability that pixel $X_{i,j}$ is fixated upon [28]. The saliency map is then used as the target for the predicted probability map $\hat{\mathbf{S}}$. We generate $\mathbf{S}$ as a sum of Gaussians around the gaze points in G, normalized such that $\sum_{i,j} S_{i,j} = 1$ (refer to Chapter 3, section 3.3.2.2). The saliency map yields the training target $\mathbf{S}^*\epsilon[0,1]^{H_D \times W_D}$ which is used as part of the augmented mini-batch described in detail in Section 2.4. Finally, the training loss is computed via the Kullback-Leibler divergence (KLD) between the ground truth distribution and predicted distribution:

$$\mathbf{L}_s(\mathbf{S}^*, \hat{\mathbf{S}}) = D_{KL}(\mathbf{S}^* \parallel \hat{\mathbf{S}}) = \sum_{i,j} S^*_{i,j} \cdot (log(S^*_{i,j}) - (log(\hat{S}_{i,j})) \tag{4.4}$$

The hyperparameters n and m that yield the best segmentation performance are used during final model evaluation.

4.3.5 Performance Metrics

We evaluate the models using the metrics of the MIT Saliency Benchmark [116]. Several measures have been proposed for evaluating saliency models which are distribution-based and location-based, i.e. saliency map distribution and fixation point (gaze map) metrics. Pearson's Correlation Coefficient (CC), Kullback-Leibler divergence (KL) and Similarity or histogram intersection (SIM) fall under the first category. Normalized Scanpath Saliency (NSS) falls under the second category.

The CC metric finds the linear correlation coefficient between two different saliency maps. Saliency maps are correlated when the CC $score = 1$ or -1, and is completely uncorrelated when the $score = 0$. KL measures the difference between saliency maps and views them as probability distributions. The metric evaluates the loss of information when a saliency map is used to measure a ground truth fixation map. KL is a non-symmetric dissimilarity metric, with a lower score indicating a better approximation of the ground truth by the saliency map. The SIM metric measures the similarity between two distributions (saliency maps), viewed as histograms. Saliency maps are identical when SIM $score = 1$ and completely different when the $score = 0$. The NSS metric finds the normalized scan-path saliency between two different saliency maps as the mean value of the normalized saliency map at fixation locations. NSS is a discrete approximation of CC that is additionally parameter-free (operates on raw fixation locations) [349].

KL is much more sensitive to false negatives (FN) than SIM, i.e. it heavily penalizes a mis-prediction of sonographer gaze. A similarity metric is penalised due to false positives (FP), while NSS and CC treat both symmetrically, i.e. penalize FN and FP equally. Further analysis of metrics and their comparison in conjunction with saliency predictions are discussed in section 4.5.

RANDOM AUGMENTATION

Baseline

KLD	3.17
SIM	0.21
NSS	2.92
CC	0.28

m → , n ↓

KLD

n \ m	1	3	5	7	9	12
1	-	3.09	2.82	2.64	2.5	2.58
3	3.00	2.72	2.72	2.25	2.27	2.23
5	2.30	2.53	2.33	2.26	2.21	2.21
7	2.75	2.25	2.42	2.21	**2.16**	2.17
9	2.66	2.40	2.27	2.21	2.23	2.20
12	2.60	2.40	2.25	2.22	2.25	2.26

SIM

n \ m	1	3	5	7	9	12
1	-	0.24	0.24	0.24	0.25	0.23
3	0.25	0.24	0.24	0.24	**0.27**	0.25
5	0.23	0.24	0.24	0.25	0.25	0.25
7	0.24	0.26	0.24	0.25	0.25	0.24
9	0.25	0.23	0.25	0.24	0.22	0.22
12	0.24	0.25	0.23	0.23	0.23	0.22

NSS

n \ m	1	3	5	7	9	12
1	-	3.39	3.67	3.85	3.92	3.60
3	3.81	3.66	3.97	4.00	4.21	4.10
5	3.40	3.67	3.92	4.11	4.07	4.15
7	3.76	4.17	3.92	4.09	4.19	4.22
9	3.83	3.61	3.95	4.09	4.01	4.04
12	3.64	3.96	3.89	3.96	3.89	3.94

CC

n \ m	1	3	5	7	9	12
1	-	0.31	0.30	0.34	0.35	0.34
3	0.34	0.33	0.37	0.37	0.38	0.38
5	0.30	0.33	0.36	0.37	0.37	0.38
7	0.34	0.38	0.36	0.38	**0.39**	**0.39**
9	0.35	0.33	0.37	0.38	0.38	0.38
12	0.33	0.36	0.36	0.37	0.37	0.38

MIXED-EXAMPLE RANDOM AUGMENTATION

Baseline

KLD	3.17
SIM	0.21
NSS	2.92
CC	0.28

m → , n ↓

KLD

n \ m	1	3	5	7	9	12
1	-	2.64	2.37	2.35	2.44	2.63
3	2.62	2.49	2.30	2.28	2.33	2.29
5	2.50	2.26	2.33	2.25	2.28	2.21
7	2.51	2.28	2.29	2.26	2.33	2.34
9	2.42	2.31	2.32	2.34	2.31	2.29
12	2.63	2.33	2.37	2.32	2.33	2.34

SIM

n \ m	1	3	5	7	9	12
1	-	0.24	0.21	0.25	0.24	0.18
3	0.24	0.21	0.26	0.26	0.24	0.23
5	0.25	0.23	0.24	0.22	0.21	0.23
7	0.23	0.23	0.22	0.23	0.20	0.24
9	0.24	0.23	0.22	0.20	0.20	0.22
12	0.25	0.22	0.22	0.23	0.20	0.21

NSS

n \ m	1	3	5	7	9	12
1	-	3.47	3.57	4.10	3.72	2.91
3	3.78	3.36	4.26	**4.34**	3.77	3.79
5	3.88	4.00	3.47	3.77	3.70	4.01
7	3.50	3.94	3.67	3.77	3.62	4.01
9	3.81	3.86	3.80	3.55	3.51	3.84
12	3.80	3.51	3.63	3.79	3.44	3.74

CC

n \ m	1	3	5	7	9	12
1	-	0.32	0.33	0.37	0.34	0.28
3	0.34	0.32	0.38	0.38	0.35	0.36
5	0.35	0.37	0.35	0.36	0.36	0.37
7	0.32	0.36	0.34	0.35	0.34	0.36
9	0.35	0.35	0.35	0.34	0.34	0.37
12	0.34	0.33	0.34	0.35	0.34	0.36

Figure 4.6: Results of visual saliency prediction with RA and Mix.RA compared to baseline with 2 augmentations. Next to the training loss (KLD), models are evaluated on the metrics normalized scan-path saliency (NSS), Pearson's correlation coefficient (CC) and histogram intersection (SIM) (for references see [30]). Best performing augmentation strategies are marked in bold.

4.4 Results

4.4.1 Quantitative Results

Figure 4.6 shows the average test scores for the Random Augmentation, Mix.RA and the baseline with 2 augmentations (described in section 3.3.3). Recall, the Random Augmentation strategy includes RandomAugment and non-linear transformations, whilst Mix.RA is the Random Augmentation strategy with linear mixed-example transformations. Both augmentation strategies, RA and Mix.RA, outperform the baseline on all metrics. RA with values n, $m = \{7, 9\}$ performs best on KLD and CC metrics, whilst RA with values n, $m = \{3, 9\}$ scores the highest by the SIM metric. RA with Mixup with values n, $m = \{3, 7\}$ receives the best score by the NSS metric.

For our segmentation task, Mixup complements the RA by producing comparable or improved results for the saliency metrics at lower magnitudes, particularly up to $m = 9$. For instance, with Mixup, NSS increased by 48.6% over the baseline, indicating better alignment with fixation points. Similarly, RA with n, $m = \{7, 9\}$ showed improvements in KLD and CC, with KLD decreasing by 31.9% and CC increasing by 39.3%. However, at higher magnitudes and numbers of augmentations, both RA and Mixup saw a decline in saliency performance, as seen in Fig. 4.6, likely due to over-distortion of the input frames. This indicates that while Mixup adds value, it is best used within a controlled augmentation range to avoid performance degradation. Both strategies are discussed in more detail in section 4.5.

4.4.2 Representative Examples

Figure 4.7 shows examples of the predictions of the Random Augmentation strategy with Mixup strategy and the comparative models on test data. Specifically, the RA strategy is represented by the best tuples of values, n, $m = \{7, 9\}$, and Mix.RA is represented by n, $m = \{3,7\}$. Our strategies are compared to the ground truth gaze distribution. Since the training and validation data were divided scan-wise fulfilling the case for 90 consecutive and temporally smoothed frames, the frames

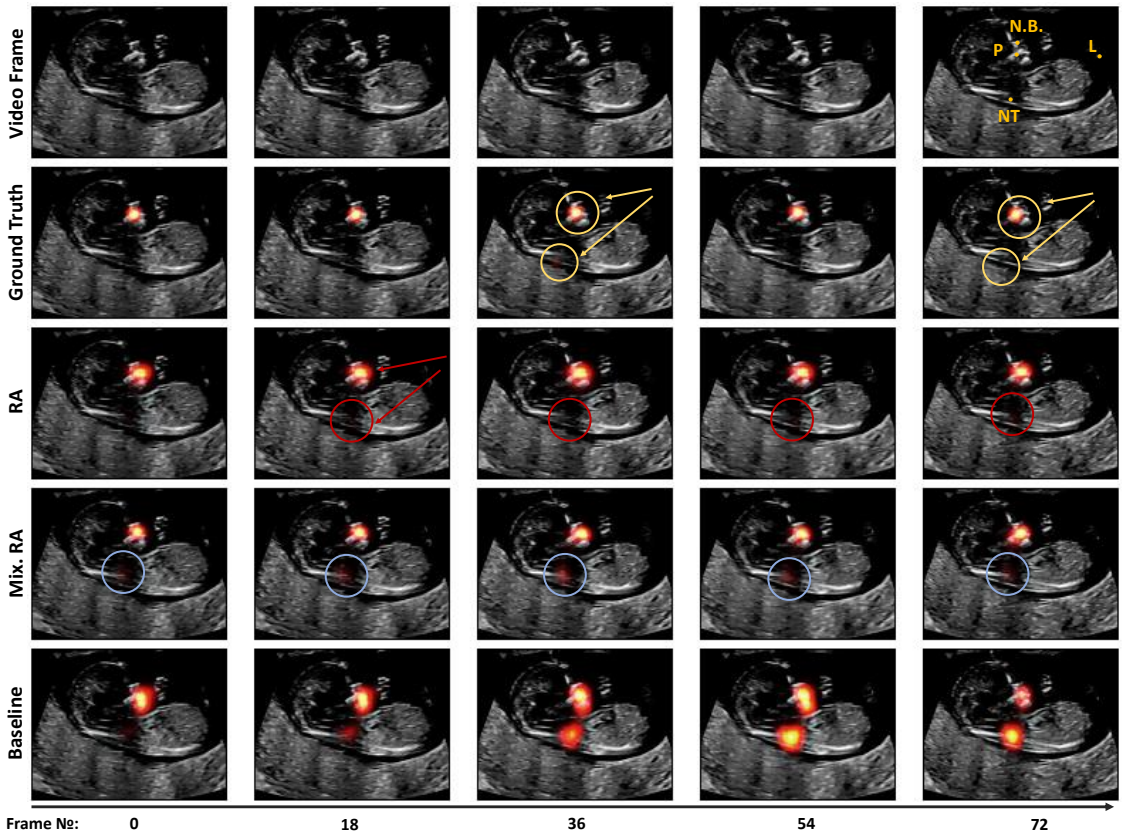


Figure 4.7: Five frames from an exemplary search sequence. The rows show the input frames, the ground truth saliency annotations, saliency predictions of our encoder-decoder network with Mix.RA against RA and baseline models, respectively. The relevant anatomical structures denoted in the last input frame (top right) include palate (P), nasal bone (N.B.), limbs (L) and nuchal translucency (NT). The ground truth is circled in yellow, RA secondary predictions are circled in red and Mix.RA secondary predictions are circled in blue.

are entirely unseen by the network. Moreover, the network is agnostic as to which sonographer is performing the scan.

From the circled ground truth frame (yellow), we see that the sonographer primarily focuses on the palate (P) and checks the nuchal translucency (NT) for guidance during scanning.

At frame zero, RA and Mix.RA predict the sonographer focusing on the end of the palate where Mix.RA also predicts the gaze at the end of the nuchal translucency (circled in blue). The baseline model has more spread-out saliency prediction around the fetal profile - lips, chin and NT.

Over the next 3 exemplary frames, RA and Mix.RA strategies predict temporally

smooth saliency maps maintaining similar positions of the palate. Mix.RA predicts fixations around the NT (circled in blue), whilst, RA assigns low saliency values to NT (circled in red). The ground truth fixations are mainly on the palate with a low probability assigned to the NT. Starting from frame 36, the baseline augmentation prediction heavily over-estimates the gaze of sonographer looking at the palate and NT.

On the last frame, the sonographer fixates on the bottom end of the palate and focuses less on the NT. Both, RA and Mix.RA models show a better prediction of the sonographer fixating on the palate and less on the NT. In contrast, the baseline predicts the saliency maps with maxima around the NT and varying maxima between the nasal bone and the palate.

4.5 Discussion

The results presented demonstrate that both augmentation strategies, RA and Mix.RA, outperform the baseline across all saliency metrics, highlighting the benefit of data augmentation for this task. As suggested in section 4.4, combining augmentation strategies helps improve model generalization at moderate transformation intensities, up to a magnitude of $m = 9$. Further augmentation with values $n,\ m = 12$ reduces the performance of the saliency prediction due to excessive distortion of images. Conversely, for both strategies lower values of $n,\ m$ display low diversity of augmented data making inferior predictions due to limited variability in the augmented data. The optimal augmentation strategy is directly dependent on the size of the model and the dataset. Given our dataset is larger than [48], a broader range of augmentations was necessary to achieve an ideal balance of data diversity and fidelity.

In terms of individual metric performance, Random Augmentation strategy with the tuple combination $n,\ m = \{7, 9\}$ achieved the best results on KLD and CC metrics, with KLD decreasing by 1.01 (approximately 31.9%) and CC increasing by 0.11 (around 39.3%) over the baseline. This indicates that RA is particularly effective in minimizing divergence from the ground truth, as measured by KLD,

and achieving strong linear correlation with actual gaze locations, as indicated by CC. The KLD metric is an integral part of our analysis as it is the only metric that measures the difference between two probability distributions, and is used for segmentation and saliency maps detection [349]. Considering the two of our ground truth fixations, if any ground truth fixation locations are missed, KLD is highly penalized which makes it the most relevant metric for our task. Additionally, RA with $n,\ m = 3, 9$ scored highest on the SIM metric, with a 0.06 increase over the baseline, representing a 28.6% improvement in similarity.

On the other hand, Mix.RA with values $n,\ m = 3, 7$ performed best on the NSS metric, with an increase of 1.42, translating to approximately a 48.6% improvement over the baseline. NSS is particularly sensitive to false positives which can be seen in Fig. 4.7 for the secondary prediction; making Mix.RA a better choice for applications that prioritize fixation accuracy and penalize incorrect predictions in sparse regions.

SIM and CC metrics are similar with the CC score being penalised due to false negatives and SIM penalizing the misalignment of density of predicted saliency maps. Unless the shape of the Gaussian blur of our predicted saliency map exactly corresponds to the GT shape, SIM will see a dramatic decline in its score. High values of similarity appear in RA strategy due to the metric being very sensitive to false positives which is reflected in Fig. 4.7. RA secondary prediction receives either no saliency prediction or a very low probability which is symmetric to the ground truth.

Qualitative analysis further supports these quantitative results. The baseline model tends to overestimate gaze on the nuchal translucency (NT), whereas RA and Mix.RA offer more refined predictions. RA and Mix.RA achieve a focused prediction on the palate across frames, successfully capturing subtle fixations on NT, with less dispersion, aligning more closely with ground truth observations.

4.5.1 Challenging cases

This section reports challenging cases when a model fails to correctly predict sonographer visual attention (see Fig. 4.8). As the saliency map was originally

generated from a gaze point, in this section (for simplicity) we sometimes refer to saliency map prediction as gaze point prediction.

When analysing the input and results of the model we noted that the acquired ultrasound and gaze data are naturally temporal. The current model learns static visual attention by treating each US video frame (and the corresponding saliency map) as an independent image. Due to shuffling prior to model training, the frames fed into a network are completely randomized and, therefore, display no temporal pattern for a network to learn. The model's prediction of sonographer gaze was based off examples of eye gaze spatially located on an image (on a specific anatomy). Whereas, in real world, the video frames have temporal connection that form a pattern which is imperative when describing or predicting sonographer gaze.

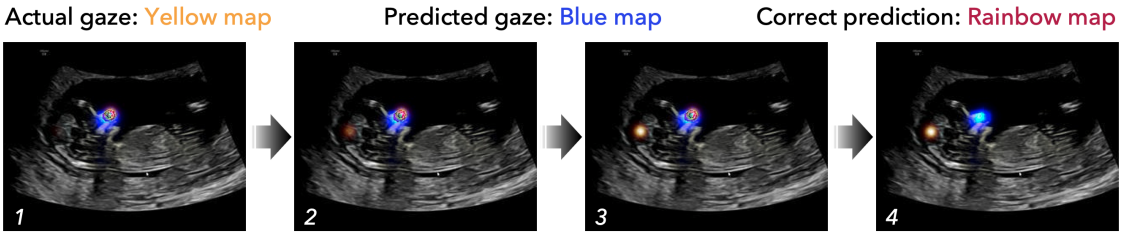


Figure 4.8: Failure cases of sonographer saliency map predictions. Given 2 similar US frames with a different gaze point location on each, the model fails to predict the fast moving gaze point. $Frame1$ displays sonographer actual gaze (yellow map) on the nasal bone and the prediction (blue map) confirms it (correct prediction is displayed as a rainbow map). $Frame2-3$: the actual gaze is moving away from the nasal bone past the diencephalon, whilst prediction stays the same. $Frame4$: sonographer gaze fully migrated to the back of the brain, whereas the prediction stayed on the nasal bone. Note: a change from $frame1$ to $frame2$ is $33.3ms$.

Figure 4.8 provides an example of how the network can receive two very similar ultrasound frames with a different gaze point location on each. As little to no temporal information about the sequence of past frames is provided, the model would occasionally mispredict the saliency map locations as it is not able to differentiate between the fast moving and slow moving segments. A solution to this problem was explored in Chapter 5.

4.6 Summary

We have presented a single frame saliency prediction for first trimester ultrasound images. We were the first to apply a stochastic augmentation policy search to a multi-modal dataset and perform saliency map segmentation. The results support the use of multi-modal data, specifically, fetal ultrasound videos combined with gaze-tracking information; as it provides a more comprehensive representation of anatomical features and sonographer behavior, enhancing the model's ability to learn meaningful patterns. This multi-modal input allows the model to capture not only the spatial relationships in ultrasound images but also the areas of interest that expert sonographers typically focus on, making it better suited to predict anatomically relevant saliency. Using a simple hyper-parameter grid search, we determined an augmentation strategy that outperformed conventional ultrasound augmentation approaches on all saliency metrics, reducing model over-fitting and improving generalization across the dataset

Clinically, accurate saliency predictions can aid real-time guidance for sonographers by highlighting relevant anatomical structures (i.e. palate and nuchal translucency) as the sonographer performs a scan. This focus improves the reliability of fetal anatomical scans by reinforcing consistent attention to critical regions. The improved performance of RA and Mix.RA highlights that augmented data, especially with appropriate transformations, can reduce the risk of overlooking critical structures, thereby supporting more standardized and accurate measurements. For example, RA with values $n,\ m = 7, 9$ achieved a notable 31.9% decrease in KLD-score and a 39.3% increase in the CC metric, indicating a stronger alignment with ground truth fixation distributions and more accurate correlations with sonographer gaze

Overall, the study demonstrates that combining RA and Mixup with the right augmentation policies, particularly up to a magnitude of $m = 9$, provides a significant performance advantage over the baseline. For instance, Mix.RA produced a 48.6% increase in NSS, showing an improved capacity to capture fixation points. Both RA and Mix.RA localized the palate and nuchal translucency well which is critical for clinical measurements. This highlights the potential of multi-modal, augmented

models to enhance clinical workflows by guiding sonographers toward key anatomical structures in fetal ultrasound imaging.

5

Spatio-Temporal Analysis: Video Saliency Prediction with Stochastic Augmentation

Contents

The previous chapter focused on predicting sonographer visual attention (saliency) in the spatial domain using a stochastic augmentation policy search method. Due to the limitations described in section 4.5.1, this chapter now turns to look at video saliency prediction and explores temporal connectivity between frames to aid in spatio-temporal saliency prediction. Specifically, this chapter automates the task of predicting where a sonographer should look next by differentiating between the fast

and slow-moving video segments. The approach is based on a novel spatio-temporal network that learns an intra-dependence of frames within a sequence and enforces a better data representation, predicting saliency for all structures and planes that come into sonographer view. Validation experiments are presented to perform saliency map segmentation using video clips of different length that record enough gaze pattern variation to track changes in sonographer gaze.

The work presented in this chapter has been published as a first-authored paper:

Elizaveta Savochkina, Lok Hin Lee, He Zhao, Lior Drukker, Aris T. Papageorghiou and J. Alison Noble. First Trimester Video Saliency Prediction Using cLSTMU-Net with Stochastic Augmentation. At the IEEE International Symposium on Biomedical Imaging (ISBI), 2022.

5.1 Introduction

So far in this thesis we have considered methods to guide novice and expert sonographers to important structures during fetal ultrasound scanning. Recall from Chapter 3, the main goal in ultrasound scanning is to find standard anatomical planes of the fetus to allow their diagnostic examination. As the total first trimester fetal examination time is limited to approximately 20 minutes per scan, it is known to be a challenging task in sonographer training to navigate around the fetal womb and capture images in an optimal fetal position. Sonographers may also lose the fetus from the field of view due to its small size and frequent movement.

An automatic method to guide sonographers to important anatomical structures is desirable to shorten the time spent navigating around the fetal womb. Inspired by how sonographers view imaging planes during an ultrasound scan, a visual guidance model in the form of predicted sonographer gaze is developed in this chapter. Similar to Chapter 4, our approach uses video and expert knowledge, defined by gaze tracking data, which is acquired during routine first-trimester fetal ultrasound scanning. In particular, we explore a temporal connection between frames and consider video saliency prediction.

The aim of work in this chapter is to utilise naturally temporal ultrasound and gaze data as well as differentiate between the fast and slow moving segments, to inform gaze prediction on unseen US frames. Specifically, we propose a spatio-temporal convolutional LSTMU-Net neural network (cLSTMU-Net) for video saliency prediction with stochastic augmentation. The architecture design consists of a U-Net based encoder-decoder network and a cLSTM that takes into account temporal information (described in section 5.3.3).

The recurrent structure of the cLSTMU-Net architecture is essential as it enables the model to learn temporal patterns and capture the continuity of sonographer gaze across frames. Unlike feedforward models like U-Net, which focus only on spatial features, recurrent layers in the cLSTM component explicitly model temporal dependencies between frames. This is critical in ultrasound, where consecutive frames contain subtle, gradual changes that reflect important anatomical structures and sonographer focus. From a practical standpoint, the recurrent layers improve generalization by providing a richer, temporally-aware representation of gaze transitions.

The LSTMs gated mechanism (input, forget, and output gates) allows the model to selectively retain or discard information at each step, adjusting emphasis on relevant frames. This is particularly useful in ultrasound as sonographers may shift focus based on multiple anatomical cues seen over a few frames. The gates allow the model to adjust the emphasis on certain frames based on their relevance, smoothing out predictions over time and preventing abrupt, disjointed saliency maps.

Spatio-temporal networks for segmentation have been proposed before [128, 350, 351]. Salvador et al. [350] designed an encoder-decoder network for semantic instance segmentation where an encoder is used for classification and a decoder is composed of a series of cLSTM layers merged with the encoder outputs in the form of skip connections. Xu et al. [351] proposed an LSTM multi-modal U-Net for brain tumor segmentation using hyper-dense connectivity that leverages different MRI modalities and temporal information. The authors first use a multi-modal U-Net to produce a pixel-wise segmentation mask which is then fed into the cLSTM.

Unlike [350][351], we use gaze-tracking data as a strong prior to guide the model towards important US structures. Wu et al. [128] constructed a SalSAC network for video saliency prediction which follows a CNN-shuffle attention module-cLSTM pipeline. Similarly, we use an encoder-decoder with cLSTM in the middle (refer to Fig. 5.1). Yet, we process the temporal input outside the spatial U-Net, pass it through the cLSTM and feed into the bottom of the decoder.

Previous work on gaze prediction for fetal ultrasound has been reported for the first trimester in Chapter 4 and for the second trimester in [207]. Recall, in Chapter 4 we investigated the prediction of spatial gaze distribution for the first trimester ultrasound. However, the model would occasionally mispredict the saliency map locations due to having no prior knowledge of the previous frames, i.e. forming no temporal pattern for a network to learn. In addition to the trimester of application, our approach is different to [207] in the use of spatio-temporal gaze patterns together with US video in training. Specifically, we utilise an encoder-decoder network with skip connections and add temporal information to improve the saliency prediction through cLSTM, exploiting the relationship between consecutive US video frames. Different from the above mentioned works that achieve performance gains due to pre-training on large image datasets, our model was trained from scratch. It is not always possible to pre-train a model on a similar dataset as ultrasound medical data is scarce. In addition, training from scratch gives more clarity on what DL tools resulted in a high prediction accuracy.

Aiming to incorporate merits from previous methods, an original spatio-temporal cLSTMU-Net neural network is proposed here for video saliency prediction (VSP) in the first trimester ultrasound scans. The first contributions of this chapter considers VSP using a first trimester multi-modal ultrasound dataset. The model learns a mapping between the US and ground truth (GT) saliency maps, predicting gaze for all structures and planes that come into sonographer view. Second, we propose a new variation of a U-Net [2] with feature sharing between 2 inputs where an additional cLSTM module incorporates temporal information, learns an intra-dependence of frames within a sequence, and enforces a better data representation.

Additional benefits to model performance are provided by applying stochastic augmentations to input of cLSTMU-Net.

5.2 Originality and Individual Role

I adapted and extended the code for cLSTMU-Net from the previous Chapter 4. I also adapted the stochastic augmentation code to be used for a spatio-temporal analysis. I processed all frames from the PULSE project first trimester ultrasound dataset and all the eye-tracking data recorded by an eye-tracker. Freeze frames were annotated by 8 experienced sonographers.

5.3 Video Saliency Prediction

The focus of this chapter is on identification of spatio-temporal eye gaze patterns that are important for US scanning and which can be useful for video saliency prediction.

5.3.1 Data and Data Preparation

For our experiments, sonographer eye gaze data and corresponding US video frames are processed 3 seconds before the freeze frame (refer to Fig. 3.8 for the scan breakdown and the reasoning in section 3.3.1) for prediction of saliency maps (described in Chapter 3). The fetal ultrasound and gaze data partitioning and data preparation (i.e. data cleaning, eye gaze manipulation and saliency map generation) are described in sections 3.3.1 and 3.3.2, respectively.

For spatio-temporal analysis of data in this chapter, there are additional pre-processing steps to be considered prior to model training (identified in Chapter 3.4). The cLSTMU-Net architecture takes 2 inputs which are a sequence of frames (a video clip) and a single frame drawn from the same video clip (illustrated in Fig. 3.14). Whilst both of the inputs are fed into the network at different stages of training, the placement of temporal frames with respect to a single frame (before or after) can be further evaluated. We performed an ablation study, reported in

Table 5.2 to identify which placement of additional temporal frames with respect to a single frame benefits the model prediction performance the most.

The cLSTMU-Net consists of a U-Net and convolutional LSTM module that allows for temporal analysis. A spatial input (single frame) is fed into a U-Net and a fixed-length video clip is an input to the cLSTM. To accommodate the cLSTM module, the data is sampled as described in 3.4.1 producing a fixed number of video clips that can be extracted from an original 3-second US video. The number of clips that can be drawn 3 seconds before the freeze frame depends on a length of a video clip pre-defined before training. We investigated different video clip lengths and select one by its performance, as summarized in Table 5.1*b* and *c*.

Prior to input data augmentation (described below) and general model training, the data is shuffled (described in 3.4.2). The order of video clips drawn from a 3-second video is shuffled at random, whereas the order of frames within each video clip remains unchanged. Only training data is shuffled.

5.3.2 Stochastic Gaze and Image Augmentation

To increase the number of training examples, increase the variation of anatomical views and reduce over-fitting, we employ a Random Augmentation (RA) strategy using stochastic augmentation policy search for segmentation purposes described in section 4.3.

We adopt a grid search with fixed magnitude schedule and a total of K=17 transformations, as in section 4.3.2. Each augmentation policy is defined by n, which is the number of transformations from the list of K an image undergoes, and m, which is the magnitude distortion of each transformation. These transformations are applied to the mixed-example images (Mix.RA), with which we share the m hyperparameter (described in section 4.3.1).

We compared two random augmentation (RA) strategies with tuple values $n,\ m = \{5,\ 5\}$ and $n,\ m = \{7,\ 9\}$, the two best results in [48] and Chapter 4, respectively (summarized in Table 5.1*b* and *c*). We selected a RA strategy with tuple values $n,\ m = \{7,\ 9\}$ as it performed best on 2/4 metrics, including the

most important saliency metric Kullback-Leibler divergence (KL) which is highly penalized if any ground truth eye fixation locations are missed (false negatives); and Pearson's Correlation Coefficient (CC) which is penalized due to both false negatives and false positives. Saliency metrics, CC and the histogram intersection metric (SIM), are similar in nature as they both evaluate the relative importance of different image regions, placing more or less weight (density) on different locations (i.e anatomical structures in our case) [349].

We used RA with values n, $m = \{7, 9\}$ in the results reported subsequently; denoted as $cLSTMU\text{-}Net(7, 9)$. Out of all transformations in section 4.3.2, we removed non-linear transformations which are already stochastic in nature (elastic and grid distortions); these transformations affect both, US images and their corresponding GT saliency maps.

The video clips are augmented at random whereas the type of augmentation transformation is shared between each frame in a clip. Such procedure is crucial to preserve the temporal information of US frames and the sonographer gaze pattern.

Similar to section 4.3.2, due to a high number of transformations in conjunction with highly distorted images, US or GT saliency map images can be transformed to display a matrix filled with the same pixel values (i.e. all pixels are 0s); the video sequence pair (US and GT saliency map sequence of frames) is discarded and the augmentation process is repeated on original video clip.

5.3.3 VSP Network Architecture

We experimented with a video saliency prediction (VSP) architecture that takes two inputs. An overview of the network is shown in Fig. 5.1. The cLSTMU-Net has two modules. U-Net is an encoder-decoder network with skip connections, and cLSTM is a recurrent network that manages a series of data that are chronologically ordered. The input to the network consists of two parts, one is a single US frame and the other is a sequence of frames preceding and including the US frame in question. The first input is fed into a spatial U-Net, and the second input becomes part of the cLSTM module. US and GT saliency map sequence of frames (video

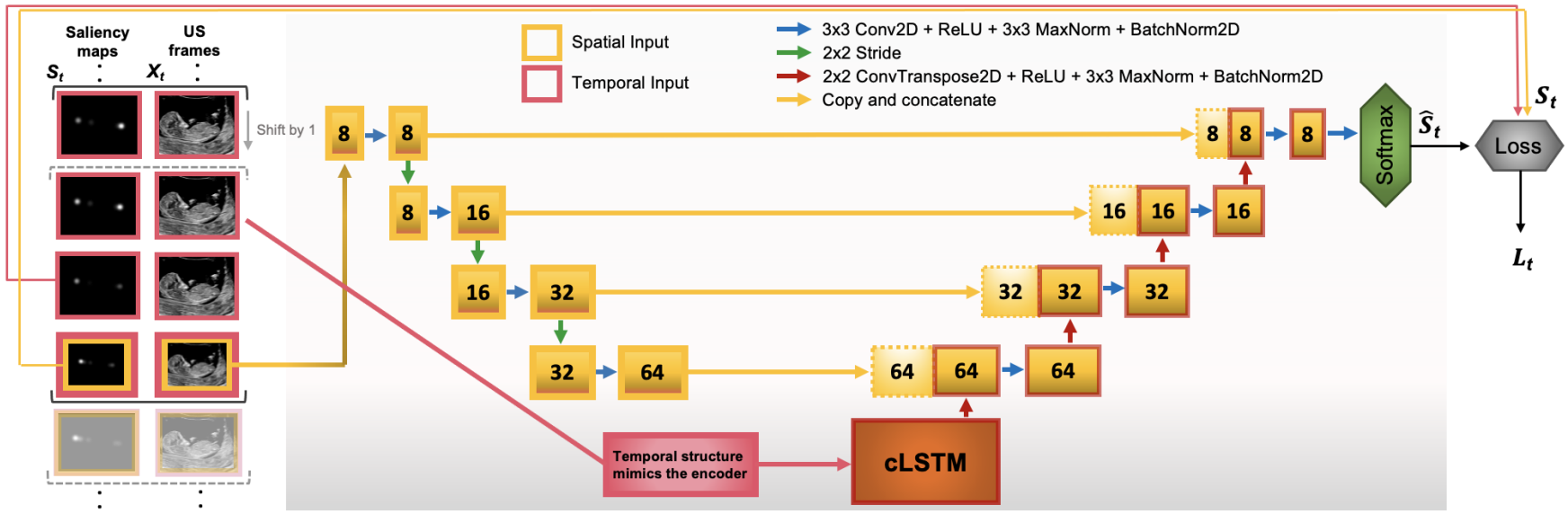


Figure 5.1: Overview of the proposed architecture for video saliency prediction. The US input frames (X_t) and ground truth saliency maps (S_t) on the left are described in Chapter 3. The loss (L_t) and saliency prediction ($\hat{s}_t$) are described in section 5.3.3.1.

clips) are sampled from an original 3-second video, as described in section 5.3.1. Next, the training video clips are shuffled, followed by Random Augmentation applied to video clips (described in 5.3.2).

The encoder layer structure of a spatial U-Net is mimicked by a temporal input. We adopt a *time distributed layer* (keras API) as it shares the weights between the images in a sequence, training them in parallel. The weights are assigned once and distributed to every image under a time distributed layer. This way, we avoid excessive training time each image in a sequence is processed in parallel; and only features that are relevant to a type of detection are used for processing, i.e. features that represent sonographer eye gaze pattern change in a video clip.

Next, we use a cLSTM to detect changes in features from frame to frame in a given time, preserving their chronological order. The output of the cLSTM is fed into a decoder layer. After a transpose convolution, cLSTM output is concatenated with skip connection from the corresponding encoder layer. The rest of the network follows the structure of the decoder with a softmax activation function applied to a final output layer.

Note, the computational order of convolutional layers and the cLSTM module is important. First, we wrapped a sequence of convolutional images into a time distributed layer to undergo the same transformation for each image in a sequence. Second, we added a cLSTM module to process the images in chronological order and detect the movement in time. Each cLSTM module works with an input in the form of convolutional layers, and not the other way round.

5.3.3.1 Saliency Map Prediction

A distribution-based loss function is used for saliency map prediction. Hence, the ground truth and predicted saliency maps are represented by probability distributions where an output is a pixel-wise prediction.

The dataset $D = \{(\mathbf{X}^{(t)}, G^{(t)})\}_{t=1}^{N^x}$ consists of N_x pairs of video frames and gaze point sets. Given an image and a gaze point set $(\mathbf{X}, G)\epsilon D$, we generate a visual saliency map $\mathbf{S}\epsilon[0,1]^{H_D \times W_D}$, where $S_{i,j}$ is the probability that pixel $X_{i,j}$ is fixated

upon. The saliency map is then used as the target for the predicted probability map $\hat{\mathbf{S}}$. Around the gaze points in G, $\mathbf{S}$ is a sum of Gaussians normalized such that $\sum_{i,j} S_{i,j} = 1$. The saliency map yields the training target $\mathbf{S}^* \epsilon [0,1]^{H_D \times W_D}$. Finally, Kullback-Leibler divergence measures the divergence between the true ($\mathbf{S}^*$) and predicted distributions ($\hat{\mathbf{S}}$):

$$\mathbf{L}_s(\mathbf{S}^*, \hat{\mathbf{S}}) = D_{KL}(\mathbf{S}^* \parallel \hat{\mathbf{S}}) = \sum_{i,j} S^*_{i,j} \cdot (log(S^*_{i,j}) - (log(\hat{S}_{i,j}))) \tag{5.1}$$

5.3.4 Network Implementation Details

A VSP architecture was trained from scratch via Adam optimization with a momentum of 0.01 and a learning rate of 0.0001 with early stopping. The batch size was set to 16 across all models. The models were implemented in Tensorflow 2.1 on a Nvidia GTX 2060 Ti. Image manipulations were performed with Pillow 7.1.2 and OpenCV 3.4.9 libraries. Networks were trained up to 40 epochs. For description of input and ground truth frames, refer to Chapter 3.

Random horizontal and vertical flipping were used in all RA policies as a baseline augmentation. The performance of networks trained with augmentation was evaluated using RA strategy with tuple values n, m = 7, 9 and n, m = 5, 5. Random augmentation strategy, $RA(7, 9)$ performed best and was consequently compared against the spatio-temporal models (Table 5.1a) from Chapter 4.

5.4 Results

5.4.1 Quantitative Results

Table 5.1 reports the average test scores for the three best performing spatial models from Chapter 4 and 2 spatio-temporal cLSTMU-Net models trained with different video clip lengths. Specifically, the spatial-only models are based on the U-Net architecture, enhanced with different augmentation strategies (RA and Mix.RA). These spatial U-Net models represent feedforward architectures where each frame is processed independently, without temporal dependencies. This comparison serves

	RA(7,9)	**RA(3,9)**	**Mix.RA(3,7)**	**cLSTMU-Net(7,9)**				**cLSTMU-Net(5,5)**			
Seq. length	1	1	1	3	**6**	9	10	3	6	9	10
KLD	2.16	2.27	2.28	2.22	**2.08**	2.11	2.22	2.31	2.18	2.25	2.17
SIM	0.25	0.27	0.26	0.23	**0.28**	0.26	0.25	0.24	0.24	0.25	0.23
NSS	4.19	4.21	4.34	4.16	**4.53**	4.41	4.06	3.96	4.23	4.14	4.26
CC	0.39	0.38	0.38	0.38	**0.42**	0.40	0.37	0.36	0.39	0.37	0.39
	a			***b***				***c***			

Table 5.1: Quantitative results of visual saliency prediction. Sequence length is displayed at the top. The best performing model is marked in **bold**. *a*: Models with best performance from Chapter 4, *b*: Spatio-Temporal cLSTMU-Net model with RA tuple values n, $m = 7, 9$ vs *c*: RA with tuple values n, $m = 5, 5$.

to evaluate the added value of incorporating recurrent layers in cLSTMU-Net for video saliency prediction.

Models are evaluated using Kullback-Leibler divergence, Normalized Scanpath Saliency, Pearson's Correlation Coefficient and Similarity metric [349]. Three spatial models include 2 models with a RA strategy and one with Mixed Random Augmentation, denoted as $RA(7, 9)$, $RA(3, 9)$ and $Mix.RA(3, 7)$, respectively. The results show that the spatio-temporal cLSTMU-Net model using a RA strategy with tuple values n, m = 7,9, outperforms all models on all saliency metrics using a video clip of length 6 frames (representative example is shown in Fig. 5.2).

The reasons behind a superior performance of $cLSTMU\text{-}Net(7, 9)$ over $cLSTMU\text{-}Net(5, 5)$ and the rest of spatial models are discussed in section 5.5. Further analysis is also made with regards to the number of consecutive frames used to predict saliency, i.e. an effective amount of frames representative of sonographer eye gaze pattern.

5.4.2 Representative Examples

Fig. 5.2 shows exemplary test results of the VSP model and the comparative spatial-only models. The spatio-temporal cLSTMU-Net network with $RA(7, 9)$ using a video clip of 6 frames better localizes the nasal bone and rump than all the other spatial-only models. Models are compared to the GT gaze distribution (yellow). Since the training and validation data were divided scan-wise fulfilling the case for 90 consecutive frames, the frames are unseen by the network.

From the GT frames, the sonographer primarily focuses on the nasal bone, nasal tip and checks the rump for guidance during scanning. In the first three exemplary frames, 3 spatial-only models fail to predict the sonographer gaze. Our cLSTMU-Net model shows an almost identical saliency map prediction of the nasal bone and gives low probability values to rump or limbs (white). The GT fixations are on the nasal bone with extremely low probability assigned to the rump at frame zero.

The latter 3 frames show cLSTMU-Net steady adjustment of the saliency prediction from the maxima around the palate to the nasal bone, which is the correct saliency map location. The less salient rump or limbs are correctly predicted

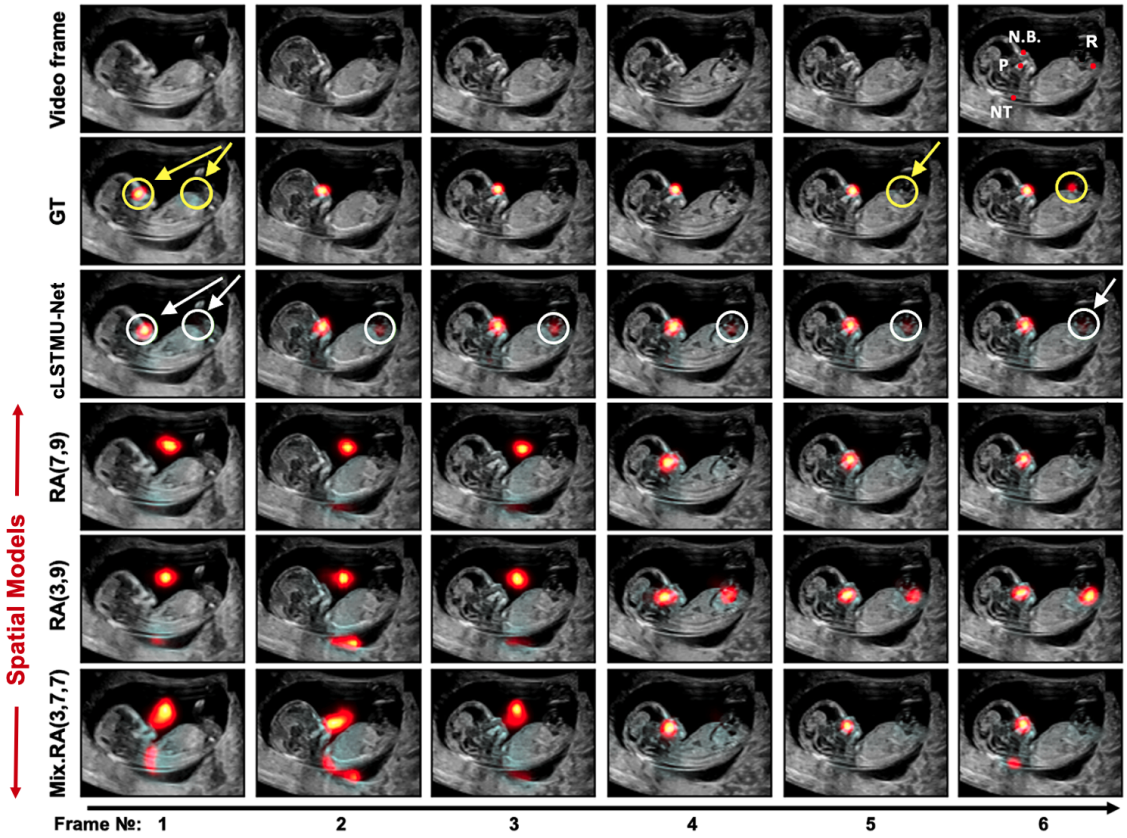


Figure 5.2: Six frames from an exemplary search sequence. The rows show the input frames, the ground truth saliency annotations, 3 spatial-only saliency models with the best metric results from Chapter 4 against video saliency predictions of cLSTMU-Net network with $RA(7, 9)$, respectively. The relevant anatomical structures denoted in the last input frame (top right) include palate (P), nasal bone (N.B.), rump (R) and nuchal translucency (NT). The less visible ground truth is circled in yellow and cLSTMU-Net predictions are circled in white.

in the last 2 frames, with slight misalignment towards the buttocks. The alternative models focus on the bottom end of the palate. Only $RA(3, 9)$ over-estimates the gaze of the sonographer looking at fetal buttocks; the other models fail to localize the structure and instead focus on NT.

The GT fixation on limbs can also be regarded as an intermediate eye movement with a trajectory towards the fetal buttocks (to measure the crown-rump length). The CRL is measured from the top of the fetal head (crown) to the bottom of the buttocks (rump). Hence, the ground truth saliency map can be viewed to be *in transition* to a final anatomy (rump) located further to the right of the GT. Consequently, cLSTMU-Net focuses on most viewed anatomy (averaged amongst

8 sonographers, refer to 5.2).

5.4.3 Ablation Study

	Temporal Information		
	Before	Before & After	After
KLD	**2.08**	2.14	2.15
SIM	**0.28**	0.23	0.26
NSS	**4.53**	4.05	4.33
CC	**0.42**	0.38	0.39

Table 5.2: Ablation study on the placement of temporal information in regard to the single frame. The best placement of the temporal information is marked in **bold**.

We evaluated the impact that placement of temporal information with respect to a single frame has on saliency prediction. We performed an ablation study with results reported in Table 5.2. A video clip of length 6 is used as an example. We observe that the addition of temporal information before the predicted frame adds the most value to saliency prediction.

5.5 Discussion

We presented a video saliency prediction network for the first trimester ultrasound images. The results show that the spatio-temporal cLSTMU-Net architecture with $RA(7, 9)$ using a video clip of 6 frames outperformed all other spatial-only models. Model training in a spatio-temporal domain allows gradients to back-propagate with respect to time and space which helps the model learn change in eye gaze pattern over time. In contrast, the gradients of spatial-only models are solely back-propagated with respect to each frame (i.e. only space). The augmentation transformations were applied to a sequence of frames, preserving the gaze pattern over time. The best performing augmentation strategy helped to populate and increase diversity of input data with new video clip examples.

From the results presented in Table 5.1*b* and *c*, we hypothesise that the superior performance when using a RA strategy with tuple values *n, m* = {7, 9} over *n, m* = {5, 5} is directly dependent on the model and the dataset size. Due to the fact that our dataset is large (a total of 45,630 images in Fig. 3.2) compared to [48] (under 2,500 images), we required a higher number of augmentations to determine the tuple that performs best for the saliency metrics.

The available computer memory could handle training a spatio-temporal model with a maximum temporal input of 10 frames. The final model performed best using 6 consecutive frames. For a first trimester dataset and a task of predicting sonographer gaze 3 seconds before the first freeze frame, we discovered that 6 consecutive frames account for a good eye gaze variation; with extra frames adding little additional useful information with no performance boost to the model. After performing an ablation study we found that adding temporal information before the frame that the saliency map is predicted for gave the best model performance.

The addition of temporal information in gaze prediction offers a more clinically relevant and accurate representation of sonographer focus, as clinical decisions often rely on trends in prior observations. By capturing eye gaze patterns across consecutive frames, the model effectively tracks the trajectory of attention, allowing it to predict regions of likely interest. This temporal approach is particularly valuable for guiding sonographers through complex anatomy, as it enables the model to anticipate areas of focus a few seconds before a critical freeze frame. This predictive guidance can be especially helpful for novice sonographers, as it provides real-time feedback on where to focus, improving scan quality and training efficacy.

Quantitatively, the KLD metric is highly penalized if any GT fixation locations are missed, for instance the nasal bone or rump in Fig. 5.2. In comparison to the three best performed models from Chapter 4, the cLSTMU-Net led to a decrease in KLD score of 0.08 (approximately 3.7%). SIM and CC metrics improved by 0.01 (12%) and 0.03 (7.7%), respectively. CC penalizes false negatives and false positives symmetrically and SIM penalizes predictions that fail to account for all the GT density. NSS is the only location-based metric, it benefits from the temporal

information with score increase of 0.19 (8.1%). The NSS metric is sensitive to false positives which is seen in Fig. 5.2 showing no false saliency prediction on the NT.

5.6 Conclusion

We presented a first trimester video saliency prediction architecture, cLSTMU-Net, combined with stochastic augmentation to enhance training data diversity. The proposed cLSTMU-Net model is able to better track changes in sonographer gaze compared to previous methods from Chapter 4; by exploring temporal connectivity and learning an intra-dependence of frames in a sequence, the sonographer eye gaze was better represented to train a saliency predictor model.

This model may be suitable for automatic guidance mechanism for real-time first trimester US scanning where the saliency predictions direct sonographer gaze to important anatomy. Ongoing studies as part of the PURFECT study are currently investigating its potential in clinical settings to assess its impact on diagnostic accuracy and workflow efficiency.

6

Visual-Assisted Probe Movement Guidance

Fetal ultrasound scanning is considered to be an operator-dependent imaging modality due to the variability in probe manipulation skills, and differences in clinical knowledge and training of sonographers. With little expertise, it is hard to mentally reconstruct and analyse a fetal womb from solely visualizing 2D ultrasound scans. To ease the mental workload and visually-assist sonographers during scanning, this chapter presents an approach to multi-modal probe movement guidance for freehand 3D ultrasound imaging.

Contents

6.1 Introduction

This chapter develops an initial prototype tracking system for freehand 3D ultrasound imaging that is completely contained within the ultrasound probe (i.e. does not need any external hardware) to visually-assist sonographers during an ultrasound examination. Recall from section 2.4, that a significant level of skill is required to obtain a recognizable, clinically useful, high-quality ultrasound image [352, 353]. Depending on the placement of the probe relative to the body and the target anatomical structure, the exact orientation of the acquired image plane relative to the patient is not precisely known. Furthermore, the level of skill varies for each operator where trainees may tend to struggle to navigate around a fetal womb and capture an optimal fetal position. With little scanning experience, it can be challenging to mentally visualize the anatomy and understand how hand movements directly affect the ultrasound image on the screen. Hand-probe movements are crucial as they dictate the 2D view of the fetus, and any rotation or translation relies entirely on the sonographer's skill. Novice sonographers often hesitate to make significant movements, fearing to loose the fetus from the field of view, as they lack the experience to predict how specific probe adjustments, such as rotating it by a certain angle, would change the view to reveal critical structures[1].

[1] comment from personal communication with sonograhers

Since each 2D US image only represents a cross-sectional view of the three-dimensional (3D) scanned structure, it requires significant mental workload to convert 2D images into a 3D mental representation. In the case of the first trimester ultrasound data, as considered in this thesis, the small size of a fetus, combined with its frequent movement, causes sonographers to lose the fetus from the field of view. Hence, this chapter investigates methods potentially suitable to guide trainee and expert sonographers to important structures during fetal ultrasound scanning. The purpose of a such system would be to ease the mental workload of sonographers.

Conventional 2D ultrasound imaging utilizes a hand-held probe that the physician moves over the patient's skin to examine the region of interest (ROI). The transducer feeds the output signal to the ultrasound machine to display the 2D ultrasound image on the screen. The 2D ultrasound image shows the cross-sectional part of the ROI. By using hand-eye coordination, the physician is able to form a mentally constructed volume of that ROI for examination of the organ features and also to estimate the volume of the ROI.

However, the reliance on 2D ultrasound images during an ultrasound scanning session has limitations (see section 2.4). First, mental integration of 2D images to form an impression of the anatomy and pathology in three dimensions may result in inconsistent with reality 3D mental impression and longer examination. Second, conventional 2D probes do not permit viewing of planes parallel to the skin, hence, an operator needs to have the expertise to reposition a probe and view a fetal structure at different angles. Third, the use of 2D US imaging for measurements of organ or lesion volume may be inaccurate [293].

On the other hand, volumetric 3D ultrasound imaging can better and more completely capture a volume of interest. 3D imaging was made clinically feasible by the development of motor-controlled "wobbler" and 2D "matrix" array probes, which allow for acoustic beam steering in the elevation dimension in addition to the azimuth dimension (refer to section 2.4.1 for more details). However, the exact orientation of the volumetric images is not registered with respect to the patient's frame of reference, hence, a high level of skill and experience is still required for

effective volumetric image acquisition and interpretation. That way, the problem of operator dependence remains. In addition, the increased data-processing complexity necessary to achieve volumetric 3D imaging results in higher cost of equipment which can be prohibitively expensive for many clinics that require ultrasound imaging [293].

This chapter explores different ways to overcome the shortcomings of 2D and 3D ultrasound and presents an integrated sensor technology which combines IMU-based orientation sensor technology to determine orientation and positional coordinates of an ultrasound probe at every time stamp. The aim of the technology is to guide trainee and expert sonographers to important structures during the fetal ultrasound scanning providing them with a digital representation of the fetus at every timestamp. This capability reduces reliance on the operator's experience and mental reconstruction skills, ensuring that even less experienced sonographers can perform high-quality scans with greater ease and accuracy.

6.1.1 Motivation

Inspired by a recent success in methods that visually-assist sonographers during an ultrasound examination [190–193], this chapter explores ways to guide a sonographer with probe movement guidance and a synthetic display of probe location on a 3D reconstructed fetus. Synthetically-displayed probe position on a fetus may help sonographers visualise and confirm the actual physical probe location. Such visually-assisted intervention may help locate a fetal ultrasound plane in question faster and ease the learning curve for trainees as well as assist experienced practitioners when navigating around the fetus womb.

For this task, multiple modalities such as ultrasound video/images and probe motion were used. An operator performs routine obstetric US scanning with a standard clinical machine. An ultrasound transducer (probe) has an attached inertial measurement unit (IMU) sensor. An on-board attitude and heading reference system (AHRS) estimates the sensor's orientation in the earth coordinate system. For each US frame, where available, an IMU measures the linear accelerations ($\ddot{x}$) (three-axis accelerometer) and rotational/angular velocities (ω) (three-axis gyroscope),

which can be numerically integrated to obtain 3-D position/orientation of an object by a process called deadreckoning. The motion sensor signal and the machine video signal are the 2 modalities used in reconstruction, where an US frame has a corresponding probe motion signal at each timestamp.

By integrating IMU data, the system provides real-time feedback on the probe's movements and positions, reducing reliance on the operator's expertise to maintain consistent scanning angles and orientations. This approach not only aids in aligning each ultrasound frame with precise 3D coordinates but also compensates for subtle hand movements and positional shifts that could disrupt the continuity of the reconstructed volume. Consequently, it ensures that each 2D ultrasound slice contributes accurately to the overall 3D image.

Ultimately, such approach aims to bridge the gap between conventional 2D ultrasound's limitations and the complexities of volumetric 3D imaging by providing a practical, cost-effective solution that improves scanning consistency and outcomes in clinical practice.

6.1.2 Overview

Each 2D US image represents a slice of a 3D fetal reconstruction. Based on kinematics solvers that combine the relative angle and position data, a relative plane orientation and probe position can be determined. At each timestamp a 2D slice has a different angle, orientation and position with respect to the predefined reference point. Therefore, a combination of 2D slices, each slice with an angular positions can be formed into a 3D reconstructed image. Finally, as the fetus is being reconstructed, where each slice of the fetus is directly related to the probe's 3D relative coordinates, the position of the probe on a virtual fetus can be displayed. Note, due to the processing and analysis of short ultrasound clips, it is assumed that the fetus does not move during data acquisition process.

Figure 6.1 summarizes the method.

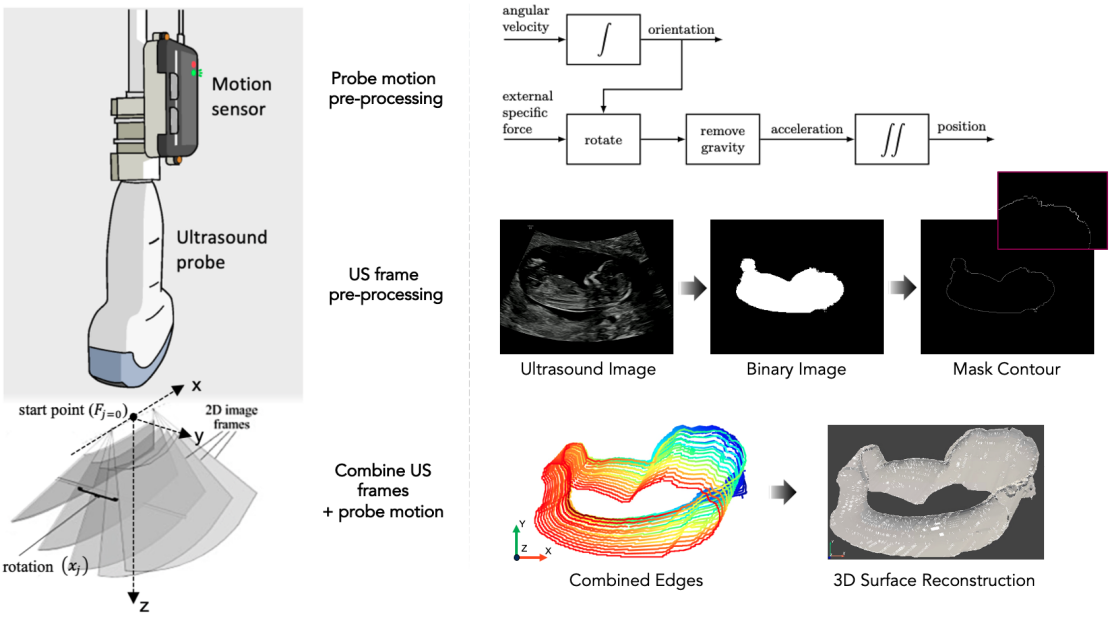


Figure 6.1: An overview of the 3D reconstruction process. A 2D curvilinear ultrasound transducer is represented by a 2D US image slice (convention used is described later in 6.3). At $t = 0$, an initial orientation of the probe is denoted as $F_{j=0}$, where $j = 0$ represents the first 2D slice. To stack 2D slices for 3D reconstruction, 3 representations that define the frame orientation are considered: matrix rotations, Euler coordinate system and quaternion coordinates (refer to section 2.3.3). With a set of 2D ultrasound images F_j^N, N cross-sectional views of a fetus are captured when a probe is rotated about an axis and shifted to a different position relative to the previous frame. Relative rotation and orientation are derived using angular velocity of a gyroscope, and relative position is derived using linear acceleration of an accelerometer. US frames are localised in 3D space and fetal masks are processed to reconstruct a 3D ultrasound fetus with probe position estimated at each frame. Prior to reconstruction a non-fetus region is masked out for better visualisation and representative surface reconstruction.

A reconstructed fetus might not be easy to visualise due to a noisy dataset, and possible failures and errors before and during the reconstruction process. Consequently, different techniques to reduce the noise and improve the 3D reconstructed fetus quality are considered. In addition, some motion sensor devices do not accurately determine real body displacements due to accumulated *integrational drift* over time (refer to section 2.3.1). To accommodate the motion sensor drift, the best results are obtained with techniques that use two sources for the acceleration data (i.e. accelerometer and gyroscope or accelerometer and magnetometer). In most cases, the accelerometer data is corrected using a gyroscope/magnetometer system, (Bayesian) Kalman filter [147, 354] and/or Madgwick filter [165].

A 3D ultrasound reconstruction and simultaneous display of probe location could be potentially beneficial during fetal examination. When performing a scan, operators would be able to note if any sections of a fetus are missing from the virtual 3D fetus. Therefore, the provision of probe location would help sonographers move the probe more freely to fill in the gaps and achieve a better 3D image representation as well as ease the rest of the scanning process. A similar commercial system was deployed in COVID-19 ICU to acquire cardiac ultrasound images [355]. Unlike [355] which provides a self-contained system with an in-built sensor technology, the proposed ultrasound system utilizes a standard 2D ultrasound probe and an attached IMU which tracks the probe movement.

The rest of the chapter is organised with the final aim of 3D fetal reconstruction in mind. A general order is displayed in Figure 6.1. The IMU-assisted probe motion data is pre-processed (section 6.3) followed by the preparation of 2D ultrasound frames described in sections 6.4 and 3.5.7. Next, the processed ultrasound masks are coupled with probe motion data to form a 3D fetal reconstruction (refer to section 6.5). The final 3D surface reconstruction of a fetus is visualised in section 6.6.

6.2 Originality and Individual Role

I extracted and processed all IMU sensor data and corresponding US frames. Ultrasound frames were manually annotated by expert sonographers and engineers for the purpose of classifying the anatomies. I combined human annotations with technical machine annotations and synchronised with the rest of the dataset used in this chapter. I adapted the Madgwick's Sensor Fusion algorithm [165] to be used as part of multi-sensor fusion and sensor error correction.

To extract fetus from 2D ultrasound frames and later produce a 3D reconstruction, Nested Hourglass (NHG) fetal segmentation architecture by Yasrab et al. [356] was used. I processed and merged both modalities, images and probe motion data that led to final 3D surface reconstruction.

6.2.1 Software

Image and contour manipulations are performed with *OpenCV*(3.4.9) library. Static 3D visualisation is performed using *Matplotlib* library. Point cloud data is represented and made accessible for processing and visualization using functionalities of *Open3D* library. *PyVista* library is used as an interactive visualization and graphics tool to represent objects in 3D.

6.3 Preparation of IMU-Based Probe Motion Data

The goal of this subsection is to prepare a probe motion tracking system that can be integrated with 2D ultrasound frames for 3D fetal reconstruction. The system does not rely on any fixed frame of reference and can track the ultrasound probe with an attached IMU sensor.

6.3.1 Gait Tracking

Whilst there are various technologies which can track a body in 3D space, this chapter focuses on inertial measurement units (IMUs) due to their light weight, low cost and ease of use [149]. IMUs are comprised of a tri-axial accelerometer, gyroscope, and magnetometer and are used to track the displacement and orientation of a rigid body in real-time (see section 2.3.1).

The IMU's accelerometer measures the gravitational acceleration. The gyroscope measures the angular velocity (rate of change of orientation), which can be used to calculate the change in orientation using numerical strap-down integration (SDI) [149]. The SDI directly integrates the gyroscope measurements, calculated change in orientation over time and updates the current orientation of the object. However, using the accelerometer, gyroscope, and magnetometer alone may yield poor estimations in terms of accuracy or robustness due to various sources of error [148] (see section 2.3.2 for more details).

Hence, IMU orientation must be first calculated using a sensor fusion algorithm (SFA), also known as attitude and heading reference system (AHRS). An AHRS

algorithm is a sensor fusion algorithm that combines the data from multiple sensors to yield a single measurement of orientation. The AHRS operates on the principle that each of the sensors provides some information about the orientation but fails to provide a complete picture [165].

The most important sensor within an AHRS system is the gyroscope as it provides the angular velocities which can then be integrated over time to measure relative orientation with respect to initial readings. If the initial orientation know and gyroscope data is free from errors, the gyroscope data is sufficient to be used in further motion analysis. However, due to the gradual deviation of the gyroscope readings from the true angular velocity (sensor imperfections and noise), other sensors such as accelerometers are integrated to correct the given state [193].

Sensor fusion algorithms combine the gyroscope and accelerometer data to obtain a more accurate estimation of the rotation. The contribution of each sensor to the updated rotation matrix can be determined by a blending factor α. By adjusting the value of α, the response time and stability of the estimation can be balanced [150]. Further, the accuracy and robustness of rotation estimation can be enhanced using Kalman filters and/or Madgwick filters (see section 6.3.2). These methods leverage additional sensor data, such as magnetometer readings, to improve the estimation and to mitigate against drift.

6.3.2 Madgwick's Sensor Fusion Algorithm

To resolve the accelerometer and gyroscope drift and prepare the motion data to be used for 3D reconstruction, Madgwick's algorithm was used [165] (now called *Fusion*). The IMU and AHRS sensor fusion algorithm includes the implementation of Robert Mayhony's *Direction Cosine Matrix (DCM)* [162] filter in quaternion form (described in detail in section 2.3.3).

The fusion algorithm (also called a *complimentary filter*) combines gyroscope, accelerometer, and magnetometer data into a single measurement of orientation relative to the Earth. It consists of two parts, first is an estimation of the orientation from angular rates measured from gyroscope data (i.e. the changes in time of the

rotation angle around each axis) and second estimation of the orientation from vectors measured from accelerometer (acceleration of the object) and magnetometer (Earth's magnetic field vector).

Madgwick's filter combines information from various sensors to estimate the object's orientation [357]. It employs quaternions to represent orientation and utilizes sensor fusion techniques. The gyroscope provides precise attitude estimates for rapid movements and short time periods. Whereas, integration of accelerometer provides accurate measurements of probe directions, compensating for prolonged gyroscope drift over time. An overview of the fusion algorithm for an IMU implementation is depicted in Figure 6.2.

A more detailed explanation of the process can be found in Madgwick's internal report [358], section 3.

6.3.2.1 Algorithm Features

The algorithm functions as a complimentary filter and calculates orientation by integrating the gyroscope data followed by the addition of a feedback term. The feedback term is computed as the product of the error in the current orientation measurement, determined by other sensors, and a user-defined gain. Consequently, the algorithm blends high-pass filtered gyroscope readings with low-pass filtered data from other sensors, with the corner frequency controlled by the chosen gain value (set to 0.5). A low gain prioritizes the gyroscope's input, making it more susceptible to drifting. On the other hand, a high gain amplifies the influence of other sensors, including the errors arising from accelerations and magnetic distortions. If the gain is set to zero the readings from the other sensors are disregarded and gyroscope data determines the orientation measurement entirely.

The main features which define the workings of the algorithm include initialisation, acceleration and magnetometer rejection, and rejection timeout. The sensor fusion algorithm is initialised at the start of the algorithm or after an acceleration timeout. When an acceleration timeout occurs, the algorithm takes corrective action by initiating a reinitialization process [358]. Acceleration timeout is set to

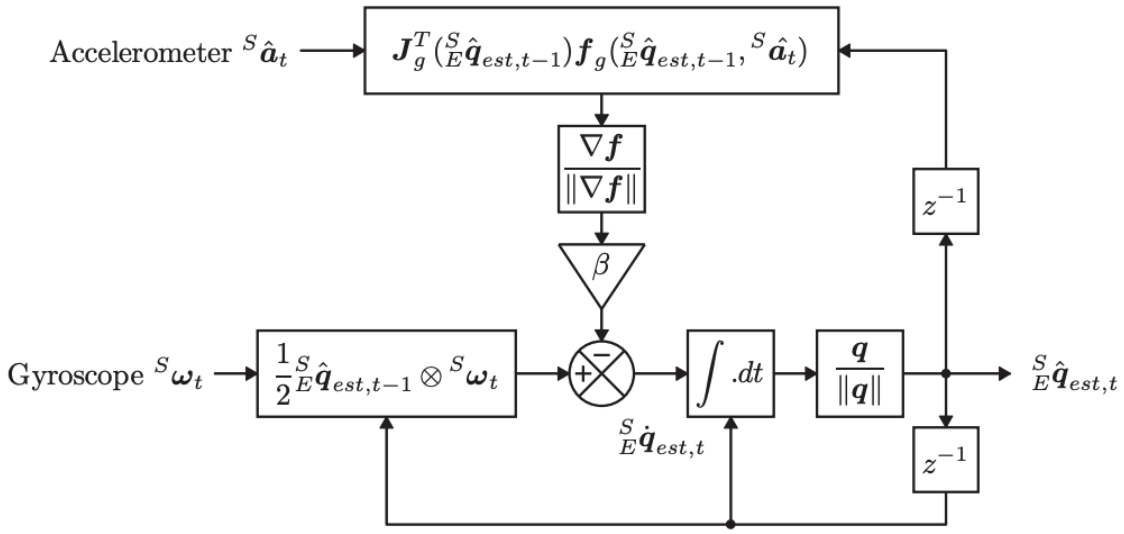


Figure 6.2: Block diagram representation of the complete orientation estimation algorithm for an IMU implementation. Refer to work by Madgwick et al. [165] for a detailed notation and algorithm derivation.

5x *sample rate* where prolonged unreliable accelerations from linear and rotation motion is detected, i.e. larger than $5x$ *sample rate*.

To prevent the timeout, acceleration rejection and magnetic rejection are implemented which help obtain a more robust and accurate results. The focus of sensor fusion is utilization of multiple sensors where each sensor corrects the readings of other sensors with the aim of overall motion enhancement. The acceleration and magnetic rejection compares the accelerometer and magnetometer respective instantaneous measurements against the algorithms output.

If the angular difference between the accelerometer's inclination and the algorithm's output inclination is greater than a certain threshold (set to 10 degrees), the accelerometer's data is deemed unreliable. In such cases, the algorithm ignores the accelerometer's data for that particular update or time step. Consequently, the algorithm relies more on other sensors, such as the gyroscope, which can provide more accurate short-term orientation information.

Magnetometers can also be affected by external magnetic sources (electronic devices or metallic objects) which distort or deviate the readings. Magnetic rejection is applied to magnetometer's heading where if the heading deviates from the output

by 20 degrees, magnetometer's data is ignored. If the readings are ignored more than $5x$ *sample rate*, the algorithm is reinitiated.

The benefit of acceleration and magnetic rejection lies in their ability to mitigate errors that could affect the final orientation estimation [359]. By selectively using or ignoring sensor data based on certain conditions, the sensor fusion algorithm can be more adaptive and resilient in dynamic environments (where sensors may be prone to fluctuations or temporary disturbances).

6.3.2.2 Output Parameterization

Three types of output can be produced by the algorithm: quaternion, linear acceleration, and Earth acceleration. Recall from section 2.3.3.2, quaternions describe the orientation of the sensor relative to the Earth. Unit quaternions can be converted to a rotation matrix or to Euler angles. Linear acceleration is retrieved from an accelerometer, whilst Earth acceleration is the accelerometer's measurement in the Earth coordinate frame. In both cases the gravity term is removed.

6.3.2.3 Convention

The coordinate frame which determines the ultrasound probe orientation used in this chapter needs to be adapted to match the convention under which the fusion algorithm can operate. Figure 6.3 displays the original coordinate frame convention (PULSE probe motion dataset) transformed to match the sensor fusion algorithm convention.

The Madgwick and Mayhony filters work in right-handed coordinate system. Hence, the fusion algorithm supports North-West-Up (NWU), East-North-Up (ENU), and North-East-Down (NED) axes conventions. It is essential to use a consistent and standardized coordinate frame convention to ensure that the ultrasound frames are reconstructed along a correct axis path with correct value rotations. Each convention defines how the axes are oriented with respect to the Earth's surface or a specified reference frame.

In navigation problems, determining the orientation of an object typically involves representing its orientation relative to a global reference frame or a

local coordinate system. The choice of coordinate frame convention is crucial to ensure that all the sensors and algorithms interpret and provide data in a consistent manner [360].

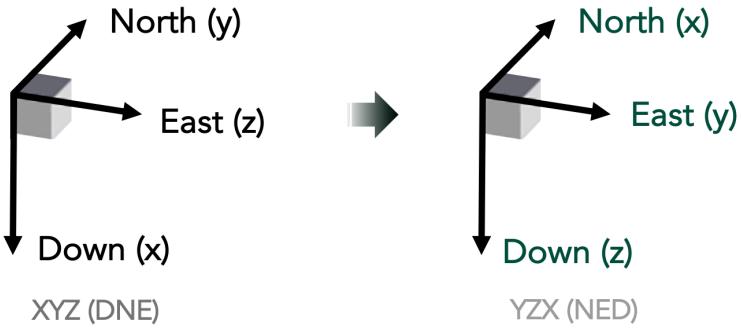


Figure 6.3: Transformation of probe motion coordinate frame convention from X (Down), Y (North), Z (East) to X (North), Y (East), and Z (Down).

The probe motion data recorded by the IMU during an ultrasound examination has the sensor's convention as Down-North-East (XYZ) (see Fig. 6.3). To closely match the original XYZ axes convention with AHRS fusion algorithm convention, a North-East-Down (NED) convention is chosen.

The NED convention is adopted by rearranging the order of the original IMU coordinates. Specifically, the probe motion convention is X (Down), Y(North), and Z (East), whereas, the AHRS coordinates are ordered as X (North), Y (East), and Z (Down). Consequently, the XYZ order of probe motion dataset axis was changed to YZX; the initial X coordinate (Down) changed to the new Z coordinate, Y coordinated (North) became X and Z (East) is a new Y coordinate.

The IMU sensor data is now aligned with the NED convention expected by the AHRS fusion algorithm and can be used for further processing.

6.3.3 Attitude Representations and Transformations

An attitude representation allows the orientation of an object to be mathematically described relative to a reference frame. This mathematical description is a set of parameters or a transformation that describes the orientation of a reference frame with respect to another reference frame [295].

6.3.3.1 Reference Frames

To describe the attitude two reference frames are considered: *world/fixed reference frame* and *body/rotated reference frame* (see Figure 6.4).

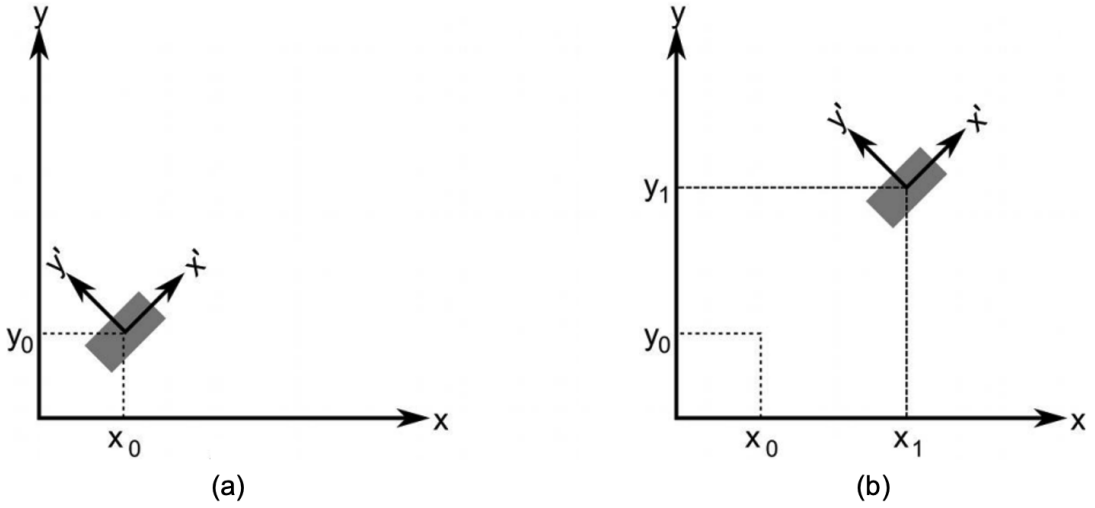


Figure 6.4: Inertial navigation example. In (a) and (b), a force acting along the x'-axis causes acceleration along both the x and y axis, i.e. the body (grey box) moves from (x_0, y_0) to (x_1, y_1) in real world frame.

The world reference frame, also known as the global reference frame, is a coordinate system that provides an external, fixed perspective from which various objects or bodies can be observed and analyzed (x and y reference frame in Figure 6.4).

A body reference frame, also known as a local reference frame, is a coordinate system attached to a specific object or body (denoted as x' and y' in Figure 6.4). It moves with the body, and its axes are often aligned with the body's axes. This reference frame is useful when studying the motion and dynamics of the body from its own perspective. It allows for the description of the body's motion and forces which can be complex whilst undergoing rotations.

6.3.3.2 Attitude Representations

There are multiple ways to express and store an attitude representation, with the main three being rotation matrix or Direct Cosine Matrix (DCM), Euler Angles, and Quaternion.

Rotation matrices and Euler angles were described in detail in section 2.3.3.1, and orientation parameterization using quaternions in section 2.3.3.2. Recall, Euler Angles use a 3 component rotation angles to parameterize the attitude and can be expressed in matrix rotations. Unit quaternions provide a concise notation for completely representing orientation or relative rotation in space, without the issue of gimbal lock associated with Euler angles (see section 2.3.3.1).

The DCM, also known as the direction cosine matrix or rotation matrix, is a mathematical representation that describes the orientation of one coordinate frame (system) with respect to another coordinate frame (system) in three-dimensional space. It is commonly used to represent rotations and transformations in various fields such as aerospace, robotics and computer graphics [295].

The direct cosine matrix is a square matrix that captures the cosine of the angles between the basis vectors of the two coordinate frames. In 3D space, with two coordinate frames (i.e. frame a and frame b), the elements of the DCM are calculated as the dot products of the unit vectors of frame a projected onto the unit vectors of frame b. The DCM projects the reference frame axes onto the rotated frame axes, i.e. when the DCM is applied to a vector expressed in one coordinate frame, it rotates that vector to align with the other coordinate frame. In essence, the DCM uses a 3×3 matrix to represent the linear transform mapping from one coordinate frame to another rotated coordinate frame.

Below is a direct cosine matrix where each element of the matrix is the cosine of the unsigned angle between the frame axes. A pre-subscript denotes the source coordinate frame a and a pre-superscript denotes the destination coordinate frame b.

$$C_a^b = \begin{bmatrix} \cos(\theta_{x^a,x^b}) & \cos(\theta_{y^a,x^b}) & \cos(\theta_{z^a,x^b}) \\ \cos(\theta_{x^a,y^b}) & \cos(\theta_{y^a,y^b}) & \cos(\theta_{z^a,y^b}) \\ \cos(\theta_{x^a,z^b}) & \cos(\theta_{y^a,z^b}) & \cos(\theta_{z^a,z^b}) \end{bmatrix} \tag{6.1}$$

where θ_{x^a,x^b} is the unsigned angle between the x^a axis and the x^b axis. This process effectively rotates the reference frame via the projection.

6.3.3.3 Attitude Transformations

Each attitude representation can be transformed or represented in form of another. Quaternions can be converted to rotation matrix or to Euler angles. In this chapter, quaternions are first calculated using the Madgwick's algorithm to help depict a full range of motion.

For ease of implementation and visualisation, quaternions were converted to a rotation matrix (with DCM applied later). The *Rotation* class provides a convenient interface for working with quaternions and other rotation representations in Python.

6.3.3.4 Vector Transformation

The position of an ultrasound probe needs to be tracked over time to later determine the location of a probe in space. The displacements that are measured by the optical tracker on the probe have to be transformed into displacements relative to the navigation reference frame, which is ultimately defined by the starting position of the probe or the initial coordinates of the probe.

The transformation is accomplished by taking the product of the DCM and the displacement vector.

$$d^b = C_a^b d^a \tag{6.2}$$

where C_a^b is a direct cosine matrix defined in section 6.3.3.2, whilst d^b and d^a are displacement vectors represented in coordinate frames b and a, respectively.

6.3.4 Algorithm Output

The final output of Madgwick's algorithm yields refined coordinates for gyroscope and accelerometer data. The updated readings are used to obtain the acceleration, velocity and position (see Figures 6.5 and 6.6).

To process linear acceleration, the gravity component is removed from accelerometer data. Next, at every stage of integration from acceleration to velocity and from velocity to position, a high pass filter is used to remove the drift (refer to section 6.3.2.1).

The corrected accelerometer data combined with angular rate data (gyroscope readings) are used to calculate the direction and magnitude of displacement that occur at each time step. This allows for accurate estimation of position relative to the initial starting point. The orientation is stored based on the updated gyroscope readings where quaternions are used for attitude calculation and converted to rotation matrices.

For visualisation purposes, quaternions were converted to Euler angles represented using Row, Pitch, Yaw convention which can be viewed in Fig. 6.5, bottom graph. All recorded measurements are stored with an equal time step (30 fps).

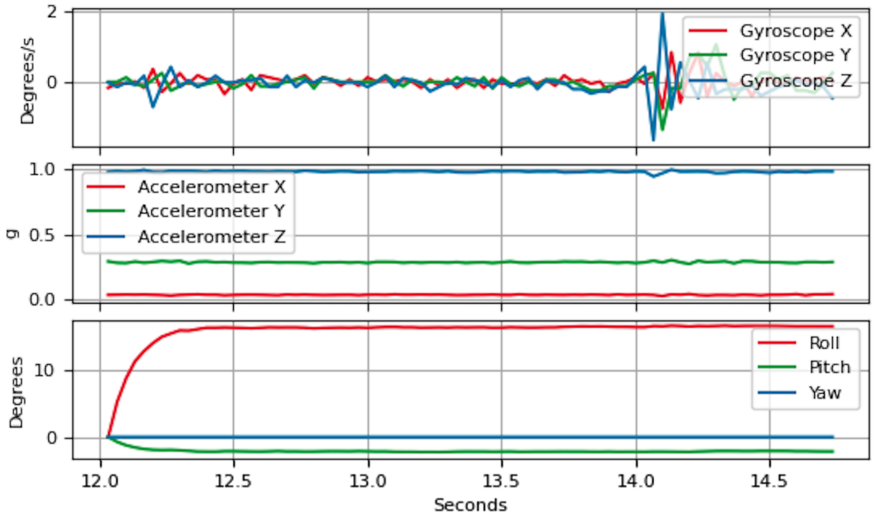


Figure 6.5: Gyroscope and acceleration in X, Y, Z direction, and Euler angles (Row, Pitch, Yaw convention) over a time period of $\sim$ 2.5 seconds with 81 frames.

6.3.5 Probe Motion Algorithm Functionality

Prior to probe motion data processing, updated coordinates of position and orientation, ultrasound frames and timestamps are synchronised. In addition, the probe motion algorithm checks for any gaps (missing entries) in probe motion and ultrasound data and only processes the continuous data. Hence, an example used in this section contains only 81 frames spread over $\sim$ 2.5 seconds of data.

Due to no fixed origin or physical reference point, the IMU data is relative and a probe's absolute coordinates cannot be obtained. A common method to

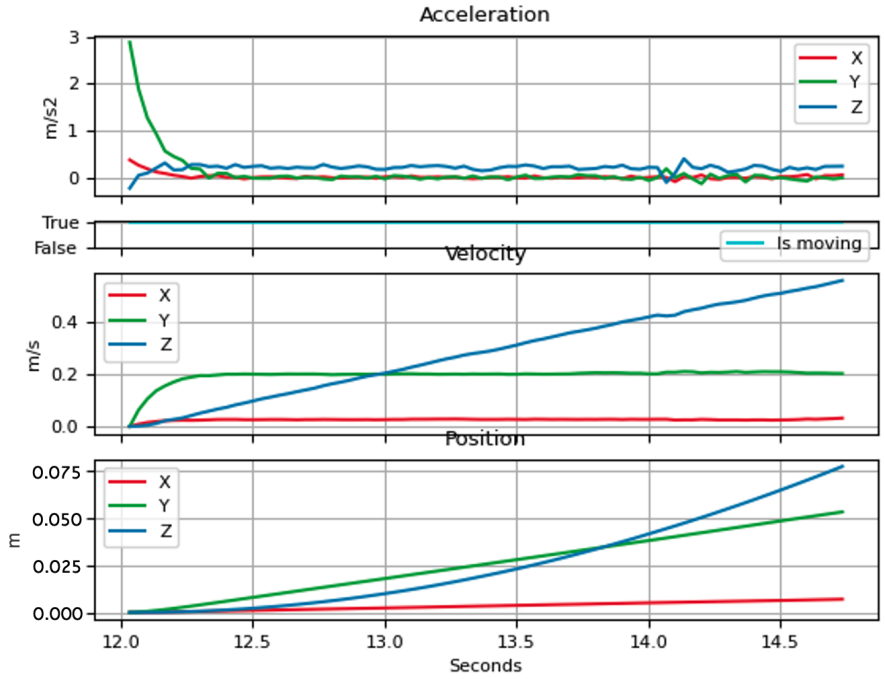


Figure 6.6: Acceleration, velocity and position coordinates in X, Y, Z direction over a time period of $\sim$ 2.5 seconds with 81 frames used.

determine the initial orientation of the probe uses the first accelerometer and magnetometer entries [147].

The initial reference point for the relative motion is set to be the first reading of probe's refined position and orientation data. The tracking system (an ultrasound probe) has no way of determining its orientation, the positions it reports are always relative to the starting point of a scan.

Recall, that the body reference frame is the coordinate system that the sensor readings are expressed in. There are different ways to track a body in space. In this work, the body reference frame is being tracked relative to the previous coordinates. Hence, the motion of an ultrasound probe in space is tracked by fixing the first coordinates of probe motion data to be the initial frame and the consecutive motion steps are calculated relative to a previous timestamp.

From section 6.3.2, acceleration is measured using an accelerometer and is successively integrated to calculate change in velocity and then position. To keep track of the direction in which the accelerometer is pointing, the body's rotational motion is measured using gyroscopes in a quaternion space. Quaternions are broken

down into four components and converted to a rotation matrix parameterized with direct cosine matrix for ease of implementation and visualisation of a final 3D reconstruction.

6.4 Automatic Fetal Segmentation

This section reflects on how 2D ultrasound images are prepared for 3D fetal reconstruction. To reconstruct a fetus in 3D and create a clear-shaped surface of it, 2D ultrasound images are segmented to extract the full length of a fetus.

Recall from section 3.2 (Figure 3.3), there are 6 biometry planes which are acquired during the first trimester ultrasound exam. In section 3.3 the fetal planes were identified through the technical annotations input by sonographers as part of the routine ultrasound scan.

In this chapter, ultrasound frames were manually annotated by expert sonographers and engineers to classify anatomies. Consequently, each ultrasound frame contains a corresponding biometry label. Ultrasound frames were categorised and the anatomical planes identified. Recall from section 3.5.4, out of 6 biometry planes only CRL views were considered for analysis.

To create a 3D fetal surface, first the Nested Hourglass (NHG) fetal segmentation architecture by Yasrab et al. [356] was used. NHG is a nested encoder-decoder semantic segmentation architecture designed to perform pixel-wise segmentation to extract the crown-rump length and nuchal translucency structures.

Recall there are only two standardised imaging planes which are compulsory to analyze and these are scored (refer to section 3.2) using an image guidance table 3.6. This means that the main focus of the sonographer is on CRL and NT measurement which may lead to class imbalance of CRL and NT versus the rest of the 6 biometry planes. Specifically, Yasrab et al. [356] work confirmed the hypothesis stated above and found that there is an extreme foreground-background class imbalance with NT plane as one of the examples prevailing over other defined biometry planes.

To address the imbalance, a class balancing based weighted-loss (WL) function was employed to improve the segmentation performance. Specifically, the WL

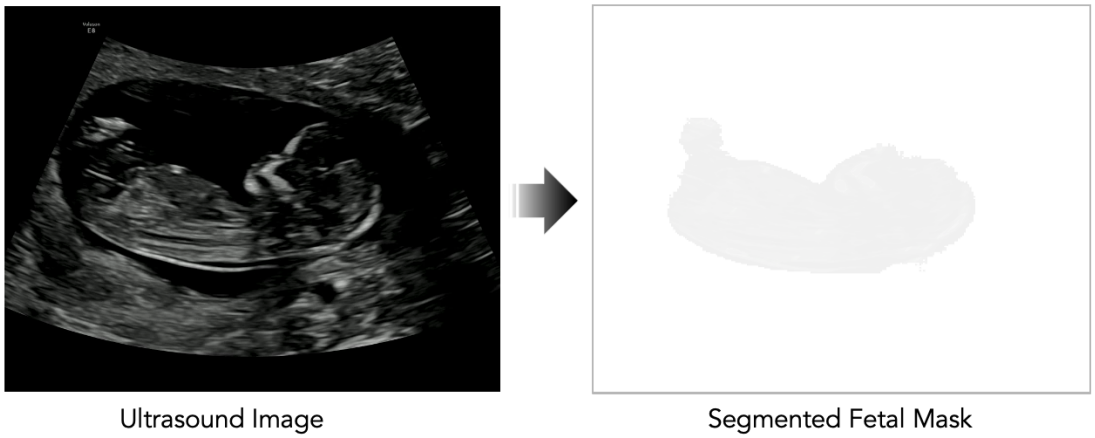


Figure 6.7: Segmented CRL fetal structure using NHG architecture for pixel-wise semantic segmentation [356].

function assigns weights to each class inversely proportional to the median frequency in which that class appears throughout the entire training set [361]. This heuristic loss calculation improves segmentation performance by optimizing the network convergence (adding focus to foreground pixels) without additional trainable parameters. The final predictions are refined using a dense Conditional Random Field (dCRF) model as a post-proceessing step. A dCRF smooths and maximizes agreement between similar neighbouring pixels of the predicted segmentation masks at the inference stage [356].

Anatomy classification was performed (mentioned above) to solely feed CRL planes into the NHG network and automatically segment the fetus from 2D images (see Fig. 6.7).

6.5 2D to 3D Reconstruction: Fetal Masks & Motion Data Combined

To perform a 3D fetal reconstruction, at each timestamp the estimated orientation and rotation of an ultrasound probe are applied to corresponding 2D fetal mask edges to track mask motion in 3D space. This involves processing a series of mask images and applying rotation and translation operations to each of the extracted masks. At a certain angle and distance, mask contours are projected into a 3D space to generate the final 3D representation of the fetus.

6.5.1 Fetal Mask Contours in 2D

Once the fetal masks are segmented from 2D ultrasound images (section 6.4), the fetal mask boundaries are manipulated to remove the false edges and improve the general fetal outline (refer to section 3.5.7).

Processed ultrasound images are grayscale where the mask edge pixel coordinates represent a foreground (white) with background pixels set to be black. To later rotate or translate a fetal edge, the non-zero pixel coordinates (mask edges) need to be extracted and stored as separate x and y variables.

Each point of a fetal mask edge is represented by x and y coordinate in 2D space which is limited by a 784x1008 image size, i.e point A has x, y coordinates (170, 5008). The size of mask contours varies throughout the scan due to fetal movement and segmentation of CRL planes from different view points, i.e. the probe tilted and rotated by 5 degrees from the original position may provide a slightly different segmentation of a fetal map.

If a segmented fetal map A is bigger than a fetal map B, the number of pixels which make up an edge of mask A is larger compared to a mask B. Hence, the array length of each contour differs from another mask edge, whereas, the number of x, y pixels in a single fetal edge is the same.

6.5.2 Fetal Mask Contours in 3D

To project a 2D mask edge into a 3D space, a z-coordinate is added to the 2D mask coordinate space to convert them to 3D space. The mask coordinates are recorded as separate variables x and y and have the same length, i.e. x and y variables each store 2040 pixel points which represent the mask edge coordinates. To add the z variable, the z-coordinates are set to 1 with the same length as x and y.

The mask contours in 3D are now represented by x, y, z variables which are vertically stacked (array of size 3x2040). The z variable is filled with 1s to apply rotation and translation (discussed below).

6.5.3 Combine Motion and Fetal Masks in 3D

After the sensor fusion described in Chapter 3 (section 3.5) and drift elimination (section 6.3), the mask edges can be transformed using corresponding motion data (motion data includes rotation and position coordinates). The transformation is illustrated in Figure 6.8.

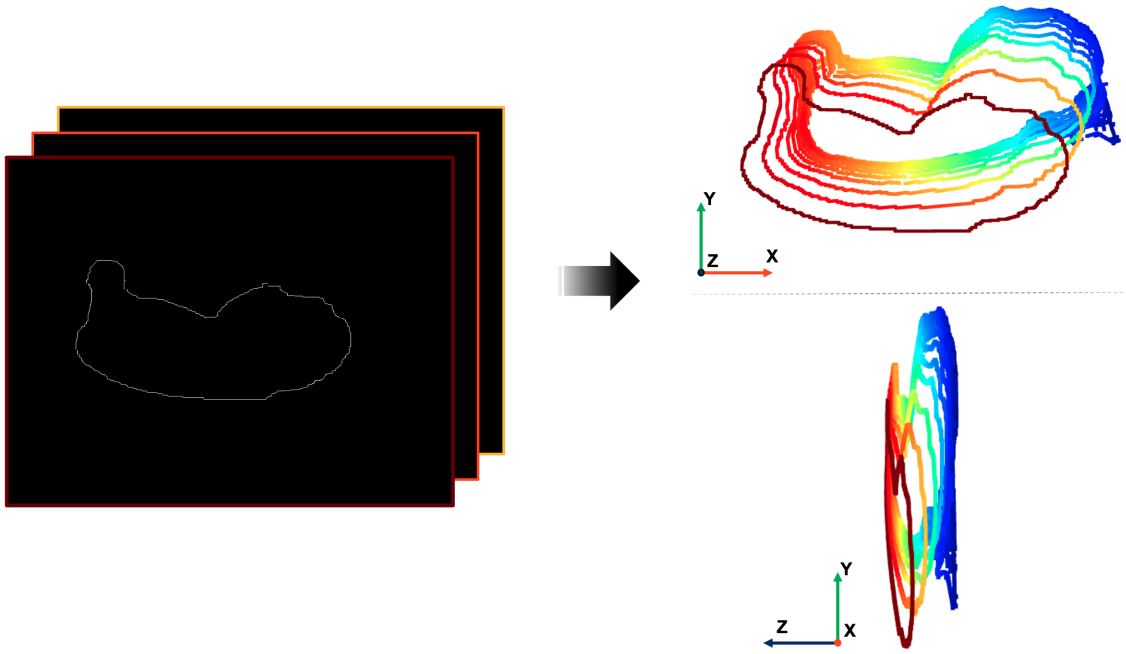


Figure 6.8: Fetal mask transformation from 2D to 3D space. On the left, a sequence of 2D mask contours is prepared for transformation. On the right, the rotated and translated sequence of mask edges is presented to view from different angles/view points.

To transform the orientation of a mask in 3D space, the rotation matrix from sensor fusion is applied to an image mask. The non-zero fetal mask coordinates (x, y, z) are transformed using rotation matrices corresponding to each timestamp. The transformation is calculated using a dot product of the rotation matrix (3x3 matrix) and the initial mask coordinates (3x2040) which results in a 3x2040 matrix. This rotation aligns the masks with the estimated orientation from the sensor data.

The step direction which represents the translation or displacement in 3D space is obtained from a 1D position array (1x3 matrix). The translated mask coordinates are obtained by adding the step direction to each point in the rotated mask coordinates.

6.5.4 Fetal Mask Gap Detection

Fetal masks transformed in 3D space (see Fig. 6.8) display substantial gaps between initial frames. To confirm this qualitatively and quantitatively, first the fetus is reconstructed using a 3D point cloud and Delaunay triangulation (described in section 6.6) to provide a visualisation of a 3D surface (see Fig. 6.9). Next, statistical analysis is performed to detect outlier values and confirm the size of gaps between frames. Note, 3D point cloud and everything related to surface reconstruction is explained later in section 6.6, whilst here it is used solely for visualisation purposes.

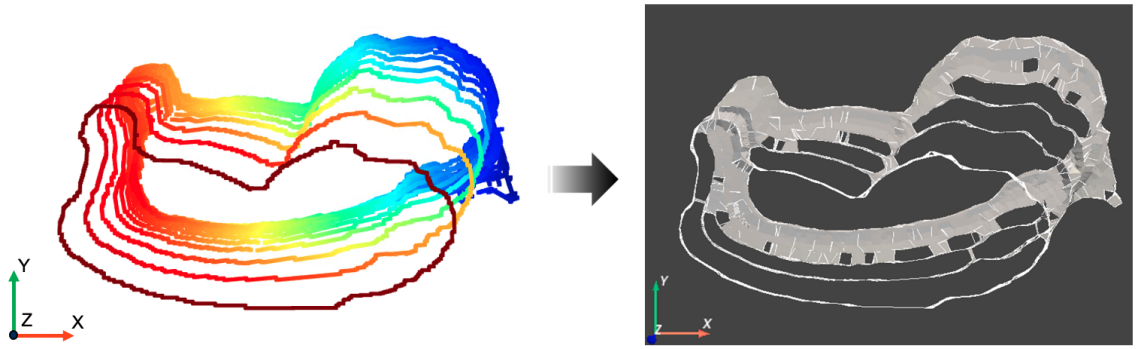


Figure 6.9: Visualisation of gaps between fetal masks through a surface reconstruction created with 3D point cloud and Delaunay triangulation (see a detailed description in section 6.6). Note, this example presents a total of 82 frames which appear over a period of 2.7 seconds. Therefore, the gap size generated in this plot is exaggerated for the purpose of visualisation and further analysis.

Figure 6.9 presents an intermediate transformation stage of 2D masks in 3D space (left to right). The reconstructed surface on the right, confirms the need for interpolation due to large gap sizes between the first $4-6$ fetal masks. The correct number of gaps needed for interpolation is determined below.

Figure 6.10 provides an overview of steps taken to determine the gap size between frames to perform interpolation.

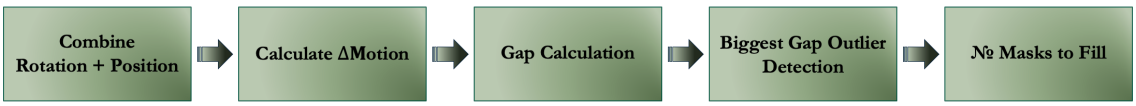


Figure 6.10: A summary of steps taken to determine outlier distances between consecutive frames which create bigger than normal gaps.

	Min	Max	Mode (freq)	Range	Gap
X	-0.0123	0.0914	0.001 (x33)	0.1037	**0.0904**
Y	0.0021	0.0932	0.0067 (x48)	0.0911	0.0865
Z	-0.0125	0.0282	0.0159 (x5)	0.0407	0.0123

Table 6.1: Combined motion change of ultrasound probe around x, y, z axes. Left to right: Minimum (Min), maximum (Max), mode and range values. The largest gap between frames along an axis is depicted in **bold**.

First, to obtain complete information about motion of each frame in x, y, z directions, probe's rotation and position coordinates at every timestamp are combined. Next, a change in motion between the consecutive data points is calculated by subtracting the motion of the current index from the motion of the previous index (x, y, z axes each represent a change in motion at every timestamp). The change in motion is combined in corresponding arrays to determine the range, maximum and mode values, and the maximum gap size in every direction (see Table 6.1). Different values are calculated to analyse the motion change along different directions.

The range is used to analyse how spread out the motion is along each axis, i.e. probe's combined rotation and translation along one axis in comparison to another. Figure 6.8 shows a more dramatic shift along x and y axes which is confirmed with higher range values depicted in Table 6.1). The mode is only used to perform further analysis if the same change in motion appears more than 8 times (arbitrary value based off the visualised motion data).

The gap size or the distance between frames (Equation 6.3) is determined by subtracting the mode value from the maximum value in a given array.

$$Gap_i = Max_i - Mode_i, \tag{6.3}$$

where i represents x, y, z axes. The aim is to find the biggest deviation from the most common motion (rotation + translation) value along an axis. This provides information on how far the single change in motion values is from the most common

change. The highest deviation from the norm informs us of the largest outlier gap which requires interpolation. The exception to the rule stands when the mode value appears less than 8 times, making the gap calculation unreliable. The highest gap value found in motion data is along the x axis (recorded in Table 6.3 in bold).

The axis (i) with the largest gap is used as a prime axis along which the rest of the outlier gaps are identified. The Z-score ($Zscore$) outlier detector method [362] is used to determine the number of outlier gaps that need interpolation (Equations 6.4 and 6.5).

$$Zscore_i = \frac{axis_i - \mu_i}{\sigma_i} \tag{6.4}$$

$$N\_ouliers = Zscore_i > threshold \tag{6.5}$$

The Z-score determines the number of standard deviations an element is away from the mean. The function identifies outliers based on a threshold (default is 0.5), where the Z-score values above the threshold are considered to be outliers. The default threshold is selected based on the initial visualisation of gaps between masks and their intermediate point cloud reconstruction. Based on the calculations, 9 fetal masks and 8 gaps between them require interpolation for a full 3D reconstruction.

The value of the largest gap size which needs no interpolation is found to be $0.0046m$ (or $0.46cm$) and anything above this value requires interpolation. The highest value (Max_value_i) which requires no interpolation is used to determine the number of masks that need to be inserted between the mask gaps (depicted in Equation 6.6 below).

$$N\_masks_i = \frac{Highest\_gap\_value_i}{Max\_value_i} \tag{6.6}$$

Per gap, the highest value ($Highest\_gap\_value_i$) in array X (1x3 matrix per axis) is divided by the maximum value (Max_value_i) which needs no interpolation. This way, the gaps are filled with the *least* amount of extra masks.

In addition, the maximum value that determines the number of masks needed for interpolation can be changed based on the final reconstruction (hole filling

and triangulation method described in section 6.6). Depending on the desired level of detail (i.e. smaller or larger triangles which make up a surface and fill the gaps between masks) and smoothness in the final reconstructed surface, the number of masks that need filling can be changed. Hence, based on the Figure 6.9 and the final reconstruction discussed in section 6.6, gaps between 7 fetal masks require interpolation.

From formula 6.6, the number of extra masks to fill each of the gaps is: $[3, 2, 2, 1, 1, 1, 0, 0]$. Hence, the gap between the mask with index 0 and mask index 1 requires 3 extra masks.

6.5.5 Fetal Mask Interpolation

The masks which require interpolation as well as the number of extra masks to be inserted to fill these gaps have been identified. Now, by iterating through each gap, adjacent mask contours (i.e. mask A and mask B) are used to create new intermediate masks. The same methods used in section 3.5.7.3 which detect, refine and create new contours have been implemented here with addition of midline search (medial axis) [363] and skeletonization [364].

This section discusses mask interpolation process using mask A and mask B as an example. The final interpolation method is applied to all the gaps and adjacent masks in need of interpolation. Figure 6.11 provides an overview of the process for masks A and B.

First, the edges are retrieved with *findContours* (see section 3.5.7.5 for contour analysis). The extracted contours for masks A and B are applied to an empty image. The gaps between the two edges are filled to create a single thick mask outline using OpenCV *findContours* function with $thickness = 5$ for each edge.

Next, a number of methods were explored to create an averaged mask that represents the edge between mask A and mask B (this would represent mask interpolation). These include calculations of distance transforms [365], morphological operations and skeletonization [366].

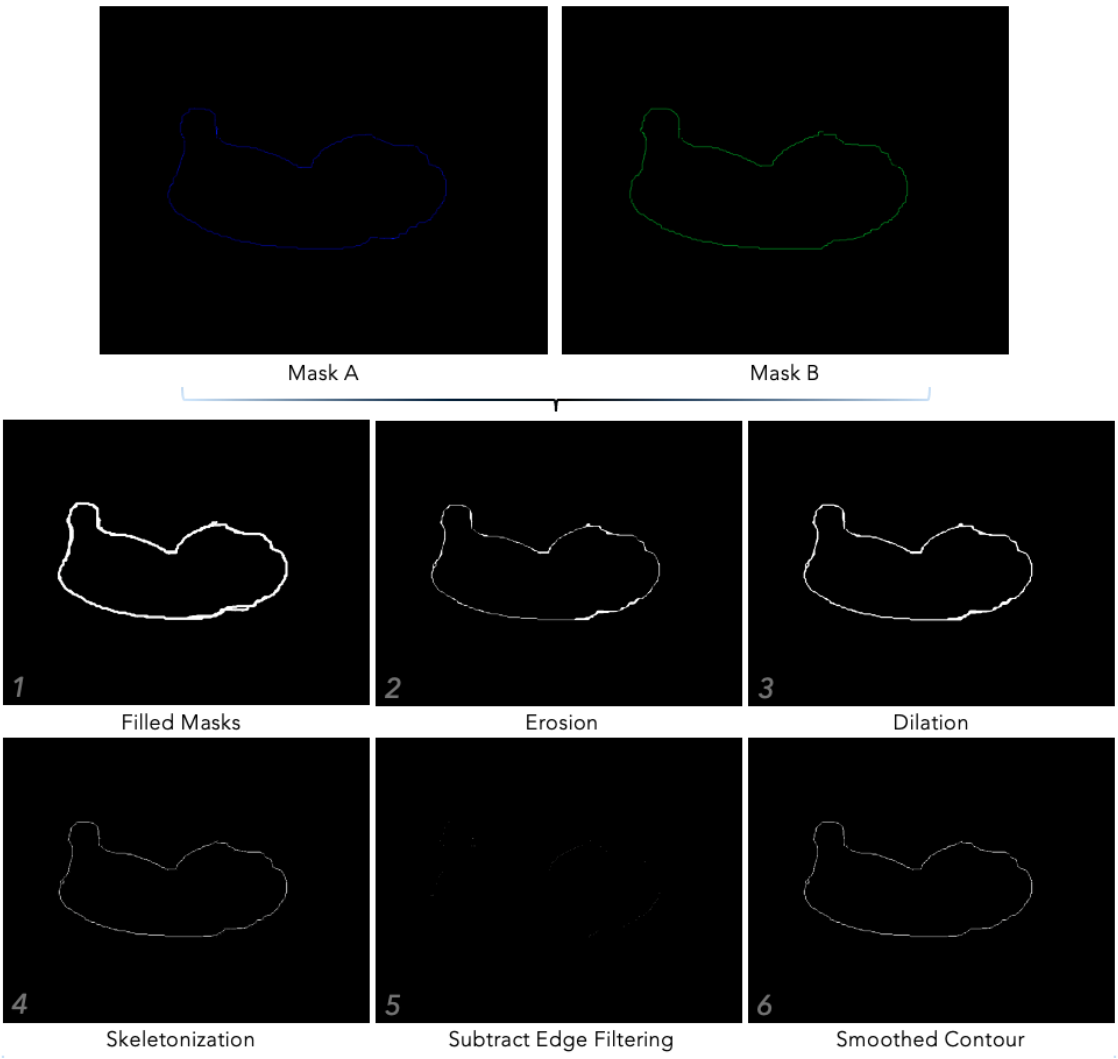


Figure 6.11: Overview of fetal mask interpolation process between two adjacent masks in need of interpolation. The order is as follows: 1. Edges combined to create a single outline, 2. Erosion (4x4), Skeletonization (Erosion), 3. Skeletonization (Dilation), 4. Filtered skeletonization, 5. Pruning or edge filtering (subtract from 4.), 6. Final smoothed contour edge.

Distance transforms and signed distance calculations were explored. However, the resulting mask did not highlight the transition region between the two masks. Instead, morphological thinning and skeletonization techniques were successfully implemented (described below).

First, morphological thinning (erosion) was used to erode pixels from the boundaries using a kernel of size 4 (the same process described in section 3.5.7.4). The resulted edge was then eroded and dilated using topology preserving skeletonization

technique by Vanajakshi et al. [366].

A topology preserving skeleton is a synthetic representation of an object that retains its shape, topology, geometry, connectivity and significant morphological properties. Skeletonization is used to reduce the thickness and size of the combined fetal edge whilst preserving the end points of line segments until no more thinning is possible [367]. The end result approximates a skeleton which is usually connected and its thickness is equal to one pixel.

Skeletonization is done using morphological operations (erosion and dilation) with a new 3x3 cross-shaped structuring element, i.e. 4-connexity. As skeletonization preserves topological and geometrical characteristics of the original edge (in case the original edge needs to be recreated), the resultant boundary contains *spurious spurs* (seen in Figure 6.11).

Pruning [366] or edge filtering can be carried out to remove the spurs. Following the same contour noise removal process as described in section 3.5.7.5, spurious or disconnected edge segments are eliminated. The final contours edge is smoothed out with dilation, followed by erosion with a 3x3 kernel.

Finally, the final mask edge which is a result of interpolated masks A and B is shown in Figure 6.12. Mask A is represented in green color, mask B in blue and the final interpolation result is presented as a white contour.

The same logic is applied when creating more than one intermediate mask with interpolation. Figure 6.13 provides an overview how masks A and B are interpolated and 3 intermediate masks are created.

First, masks A and B are interpolated and the skeleton (midline) is extracted which becomes an intermediate mask edge (Mask C). As 3 extra masks are required to be inserted between masks, the same process is repeated between masks A and C, and masks C and B. Once both pairs have been interpolated and generated new intermediate mask edges, mask D and mask E, respective to each pair. The final combination follows of masks follows the order: A, D, C, E, B, where masks A and B are the prime boundary edges and masks D, C, E are the intermediate mask edges.

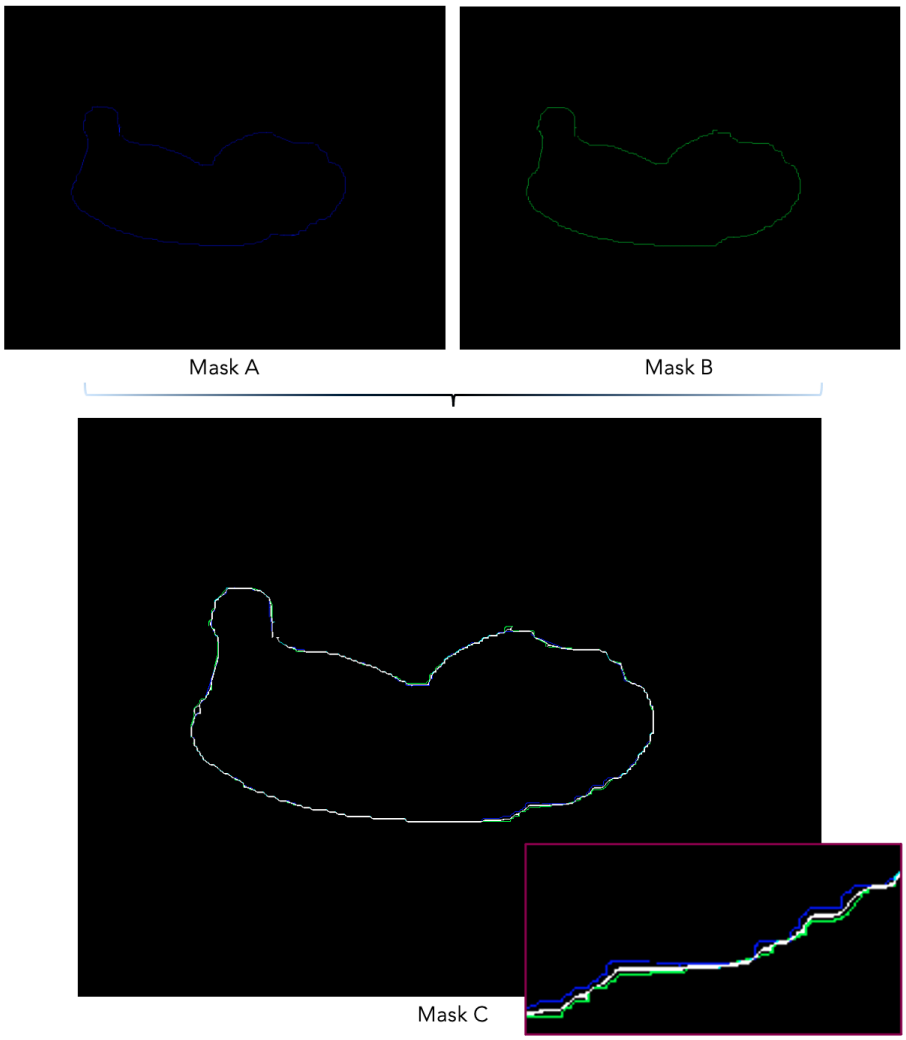


Figure 6.12: Final interpolated mask A (blue) and mask B (green) with generated intermediate mask C (white).

The same method is applied to the rest of the fetal masks which require edge interpolation and gap filling.

6.5.6 Motion Interpolation: Position and Rotation

This section determines the placement of fetal masks between the masks (and gaps between them) in need of interpolation. Again, the method is explained with an example of masks A and B, and later applied to the rest of the dataset.

The same masks A and B which were used for mask interpolation are now used to perform interpolation between two motion matrices, motion A and motion B. Each mask edge has a corresponding rotation and translation coordinates. The

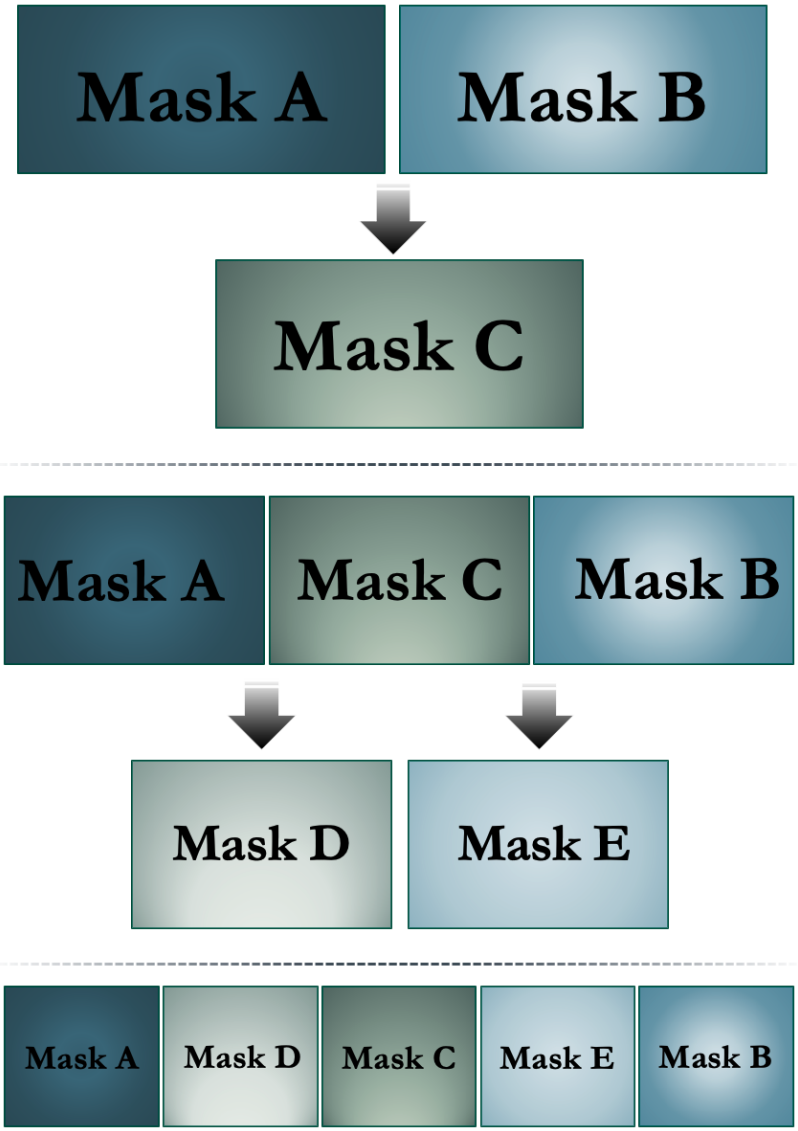


Figure 6.13: Interpolation process of masks A and B to generate 3 intermediate masks D, C, and E.

intermediate motion matrices can be calculated based on the number of extra masks required to be inserted between A and B. The distance between each frame and its placement directly depends on the quantity of intermediate masks.

To follow the same example, 3 extra masks need to be inserted between masks A and B. The weights are determined based on the number of extra masks, where each weight represents the interpolation factor for each intermediate mask. Since 3 masks (N_masks) need to be fitted, the total number of masks including motion A and B is 5. Hence, the weights are calculated as follows:

$$Weights = \frac{N\_masks - i}{N\_masks + 1}, \tag{6.7}$$

where i is the index of the weight in the range [1,3]. That way, 5 masks have the following weights, W $= [0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1]$.

The weights are assigned to new masks based on their position (distance) from masks A or mask B. The interpolation factor gradually shifts from motion A to motion B based on the weights. This is calculated by interpolating motion and iterating over the list of weights (see Equation 6.8).

$$Interpolated\_motion = (1 - weight_k) * motionA \tag{6.8}$$

$$+ weight_k * motionB \tag{6.9}$$

where $weight_k$ is the index of the weights (W).

6.5.7 Combine Interpolated Motion and Fetal Masks

Following the same process as in section 6.5.3, interpolated motion matrices are applied to newly created fetal mask edges (described in section 6.5.5). The transformation is calculated using a dot product of the motion matrix and the corresponding mask coordinates. Interpolated motion aligns the masks with the estimated rotation and orientation in 3D space.

6.5.8 Combine Original & Interpolated Masks with Motion in 3D

This section combines all calculations and analysis from sections 6.3, 6.4, 3.5.7 and 6.5 into a final stack of fetal mask edges in 3D space with corresponding motion data. Original masks and motion data are combined with interpolated data points to create a final reconstruction in section 6.6.

Figure 6.14 shows combined interpolated and original mask contours which represent a fetal outline in 3D space.

Each data point of a fetal contour is stored in a list and can now be used to create a point cloud (discussed below).

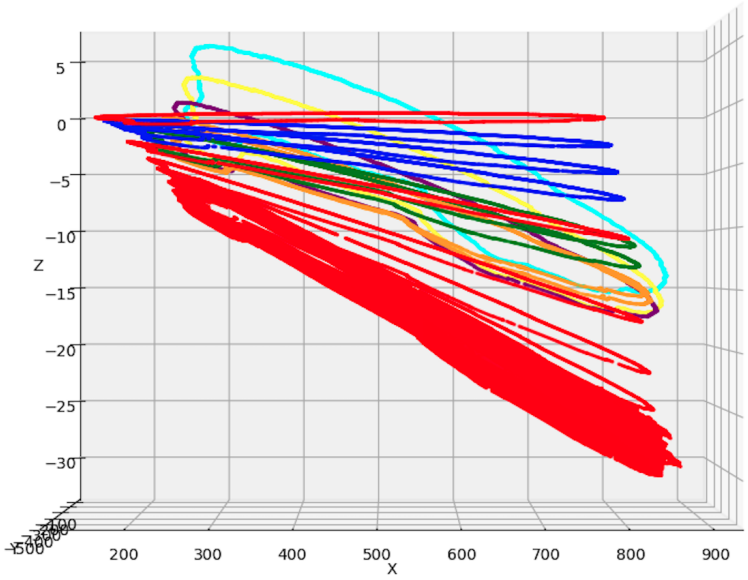


Figure 6.14: Combination of interpolated and original fetal contours with corresponding motion data, all used to represent fetus in 3D space. Original ultrasound mask contours are represented in red color and other colors represent intermediate edges. The fetal mask edge in 3D space has a maximum frame size (1008x784x35) in x, y, and z directions. Note, the step size between frames and motion around $z - axis$ looks exaggerated due to the size of a plot, here $z - axis$ is scaled by a factor of 22.

6.6 Visualisation of Fetal Surface Reconstruction

This is a final section which provides a comprehensive review of the visualisation techniques used to showcase 3D fetal reconstruction.

6.6.1 Point Cloud Representation

To create a surface reconstruction, the stored data points are used to form a point cloud representation of the fetus. A point cloud consists of a collection of 3D coordinates, i.e. stored x, y, and z coordinates of fetal masks. Cartesian coordinates are assigned to each point position to represent its location in 3D space. Point clouds provide a detailed representation of the shape, characteristics and spatial distribution of objects, allowing for accurate measurements, surface reconstruction, and other data processing tasks [368].

Figure 6.15 presents point cloud visualisation using Open3D.

Point clouds maintain the spatial relationship between points, allowing for analysis and computations based on proximity, distance, or connectivity. In

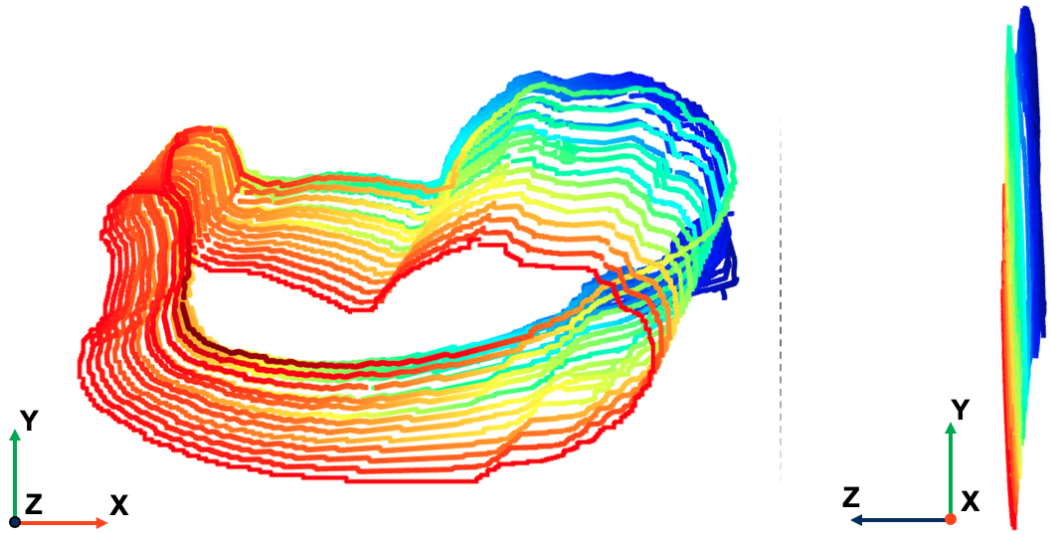


Figure 6.15: Point cloud representation of the fetus in 3D space. Rainbow-like colors of each mask correspond to a different Z-coordinate value (or height of each point).

particular, a point cloud can be used as input for a surface reconstruction algorithm because they provide a dense set of 3D points that represent the surface of an object or a scene [369]. That way, a dense set of 3D coordinates that represent the fetal surface geometry can be generated.

In this work, since ultrasound images have been segmented and the location of the final mask edges have been recorded in 3D space, a 3D point cloud file can be created. To convert the sequence of mask edges with their corresponding motion coordinates to point cloud format, the Open3D library is used.

First, the list of x, y, and z coordinates is flattened and an array representation is created with shape (num_points, 3). The Open3D library is leveraged for handling and analyzing 3D geometric data (see Equation 6.10). Next, the points of the flattened array (*point_cloud*) are assigned to *pcd* object as $Vector3dVector$ (Equation 6.11). By using the $Vector3dVector$ class, the point cloud data is efficiently represented and made accessible for further processing and visualization using functionalities of Open3D library.

$$pcd = o3d.geometry.PointCloud() \tag{6.10}$$

$$pcd.points = o3d.utility.Vector3dVector(point\_cloud) \tag{6.11}$$

6.6.2 Delaunay Triangulation Surface Reconstruction

Once the point cloud is generated, it can either be directly rendered and inspected or converted to a variety of representations including polygon and triangle mesh models [370], NURBS surface models [371] and others.

Surface reconstruction algorithms can be applied to create a continuous and smooth representation of the fetal surface. The algorithms analyze the spatial distribution and connectivity of the points to estimate the underlying surface geometry. The result is a reconstructed surface that approximates fetal shape and structure based on the input point cloud.

There are many techniques which convert a point cloud to a 3D surface, these include approaches such as Delaunay triangulation [372], alpha shapes [373], and ball pivoting [374]. The approaches build a network of triangles over the existing vertices of the point cloud. Other approaches convert the point cloud into a volumetric distance field and reconstruct the implicit surface, i.e. a marching cubes algorithm.

In this work, fetal surface reconstruction is used to estimate the underlying surface geometry from the scattered points and create a detailed and realistic representation of the fetus in space; such reconstruction aids in further analysis, visualization and interaction with the reconstructed fetus.

First, using previously flattened point cloud array, a PyVista *PolyData* object is created which stores geometric data (Equation 6.12). PyVista library interpolates the surfaces to create a smooth 3D appearance. Next, the Delaunay triangulation is performed on the point cloud data (geometric data) with the specified alpha (see Equation 6.13).

Delaunay triangulation is a method for creating a surface mesh from a set of points in 3D space where no point should lie within the circumcircle of any triangle formed by the points. The *alpha* value represents the distance parameter. A smaller *alpha* value creates a denser mesh with more triangles, closely following the shape of the point cloud. Whilst larger *alpha* value generates a sparser mesh with fewer triangles, i.e. a smoother representation of the surface [373]. In essence,

a smaller *alpha* value leads to a more detailed representation, while a larger *alpha* results in a simplified and smoother surface.

Next, surface geometry is extracted (see Eq. 6.14) to analyse the connectivity and topology of the surface mesh and identify regions that are considered to be holes. These regions typically have open boundaries or incomplete faces. The function $fill\_holes$ automatically detects and fills in the holes in the fetal surface geometry. This is done by specifying the size of the hole filling variable (see Equation 6.15).

$$cloud = pv.PolyData\,(point\_cloud) \tag{6.12}$$

$$surf = cloud.delaunay\_3d\,(alpha = 10) \tag{6.13}$$

$$shell = surf.extract\_geometry\,() \tag{6.14}$$

Depending on the gaps and holes between fetal contours, the size of the hole filling function can be changed based on the proximity of each contour to the next one. The higher the interpolation (in section 6.5.4), the smaller the size of the hole filling function is set.

$$filled\_surface = \; shell.fill\_holes\,(hole\_size = 90) \tag{6.15}$$

The filled surface representation is stored and the reconstructed fetal surface is visualised using PyVista library. Figure 6.16 presents a final 3D fetal surface reconstruction with variation shown in terms of interpolation intensity, distance parameter *alpha* which defines the *thickness* of the edge and the size of surface hole filling.

6.6.3 End-to-End 3D Surface Reconstruction

This section provides an end-to-end visualisation of the surface reconstruction process, starting from non-interpolated fetal masks and ending with a full 3D surface reconstruction. Figure 6.17 presents 3D point cloud generation followed by Delaunay triangulation to perfrom surface reconstruction.

First, the sequence of fetal edges are converted to point cloud representation. The mask edges contain noise and outliers as well as the final 3D representation

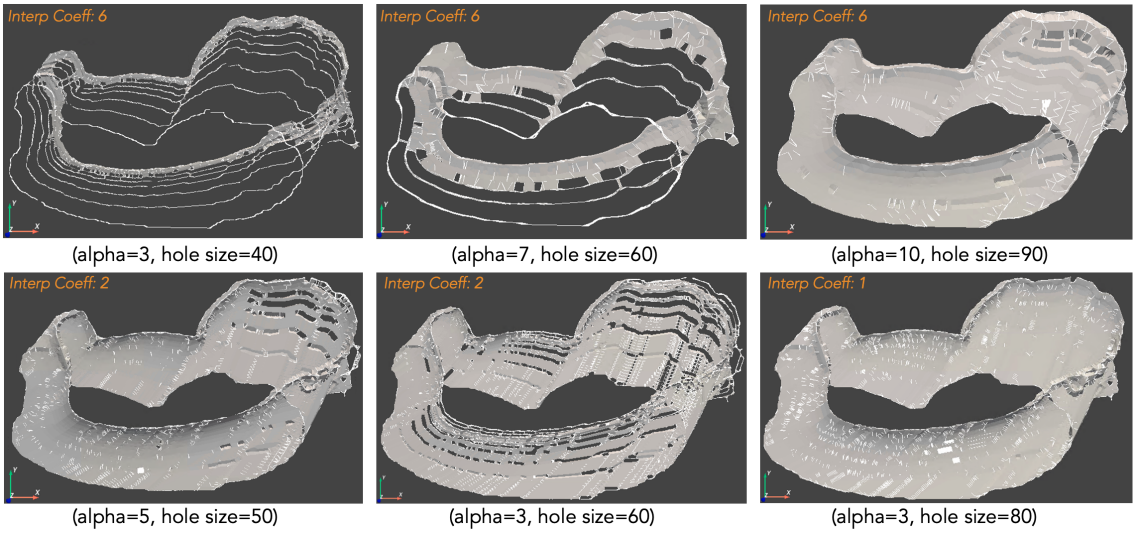


Figure 6.16: Visualisation of 3D fetal surface reconstruction using PyVista and Delaunay triangulation. Each frame represents a 3D reconstruction, the top row reconstruction is performed with interpolation coefficient of 6 (denoted in orange in the top left corner of each frame) and a variety of alpha and hole filling coefficients. Bottom row presents reconstruction made with most or all intermediate frames. Higher number of intermediate frames generated after interpolation provides a more detailed reconstruction (seen in the bottom right corner) with smaller size triangles filling the surface holes.

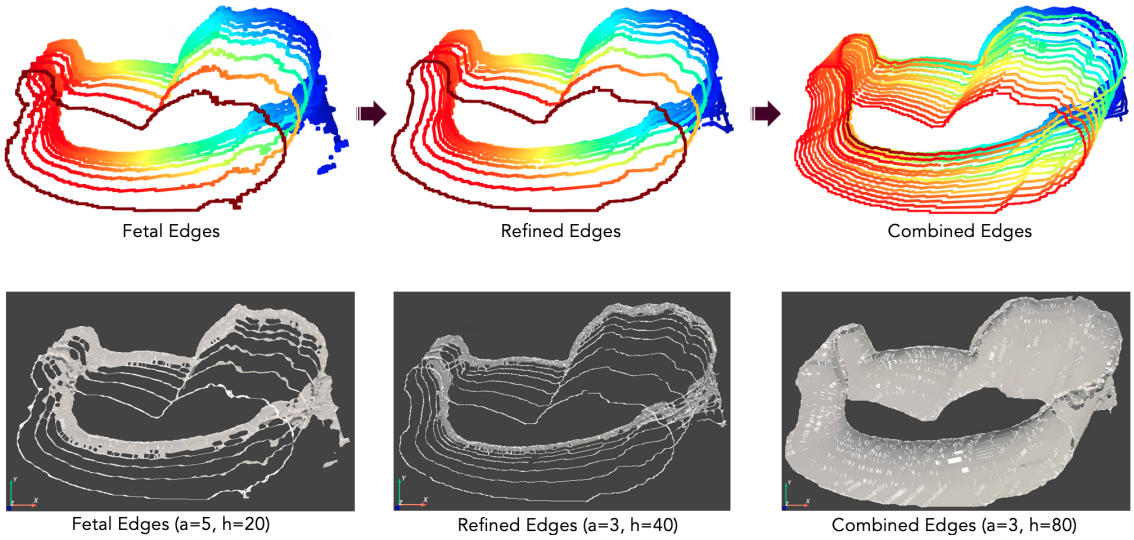


Figure 6.17: End-to-end 3D surface reconstruction. Left to right: original fetal edges before contour refinement, refined edges, a combination of intermediate & original masks after interpolation. Top to bottom: point cloud representation followed by Delaunay triangulation 3D surface reconstruction (alpha denotes as a and hole size as h).

contains large gaps between mask edges. The fetal edges are also reconstructed with Delaunay triangulation for comparison purposes. All final reconstructions contain no reduction in the number of intermediate masks produces after interpolation.

Second, the mask contours are refined which can be seen by comparing the first and the second columns of Figure 6.17. The edges are smoothed out and the noise reduced. Finally, the mask edges are interpolated and intermediate masks are combined with the original mask edges (Fig. 6.17, column on the right). It can also be noted that the final interpolated edge (see the dark blue noise in the top right corner) contains less noise. Prior to interpolation and reconstruction, the mask with the highest number of foreground pixels was removed (it contained the most noise). That way, the final reconstruction (last column) contains less noise compared to *Refined Edges* frames (blue noise in top right corner).

6.7 Discussion

The initial prototype tracking system for freehand 3D ultrasound presented in this work offers several potential clinical benefits. By integrating IMU-based orientation sensors with standard 2D ultrasound imaging, the system assists sonographers by providing continuous feedback on probe orientation and position. This feature reduces the cognitive load on operators, particularly benefiting trainees who may struggle to anticipate how probe movements affect the ultrasound image.

Clinically, this system could bridge the gap between traditional 2D ultrasound and more advanced 3D imaging techniques without the high costs and complexities associated with motor-controlled or matrix array probes. For trainees, it acts as a valuable teaching tool, offering real-time feedback to develop better probe-handling skills. Experienced sonographers can also benefit from enhanced workflow and additional probe orientation information during scans.

The system leverages the placement of 2D ultrasound fetal masks relative to the probe's contact point with the skin. Analyzing and processing short clips for 3D reconstruction assumes minimal fetal movement, allowing straightforward modeling based on rotation and positional data. However, to extend this to longer scans, motion correction algorithms are necessary (see section 7.2 for a detailed description of future work). All algorithms estimate the motion and apply corrections during the

reconstruction. Integrating these algorithms would ensure that the reconstructed 3D image remains reliable and accurate in dynamic scenarios.

6.8 Conclusion

We presented an initial prototype tracking system for freehand 3D ultrasound. The integrated sensor technology combines IMU-based orientation sensor to determine orientation and positional coordinates of an ultrasound probe at every time stamp. Simultaneously, corresponding 2D ultrasound frames are used to segment the fetal structure and extract a fetal contour. Fetal mask edges and probe motion coordinates are combined to form a fetal surface reconstruction in 3D space.

Far and away the best prize that life has to offer is the chance to work hard at work worth doing.

— Theodore Roosevelt [375]

7 Conclusion

Contents

This thesis has presented a series of automatic image analysis algorithms to visually-assist and guide sonographers during a first trimester ultrasound examination. This chapter summarizes the main contributions of the research and some directions for future research.

7.1 Conclusions

The main contributions of this thesis are (in chronological order):

1. Preparation of a first trimester fetal ultrasound dataset with ultrasound videos and frames, real-time sonographer gaze tracking and IMU-assisted probe motion data.

Chapter 3 presented the research, methodology and techniques all related to data used in this thesis. The chapter began to describe PULSE fetal dataset curation and ultrasound equipment used to collect the data, how the main three modalities used in this thesis were acquired and processed for further analysis and research. The

modalities included ultrasound videos and frames, probe motion tracking achieved using an IMU and sonographer eye gaze recorded with an eye-tracker. Chapter 3 provided a description of the subset of data for each research contribution in this thesis, including the data acquisition, data preparation, annotation (if any), and the detailed train/validation/test sets partitioning. Data visualisation was also provided for better understanding of the dataset.

2. A stochastic augmentation policy search method to improve segmentation performance when predicting where sonographer is going to look next.

Chapter 4 presented a single frame saliency prediction algorithm for first trimester ultrasound images. A proposed augmentation strategy was implemented to expand the size of medical training data, combat class imbalance and data shortage for a specific, and increase model generalization. The augmentation policy search consisted of a simple grid search and different types of augmentation transformations applied with different intensities where ultrasound images and ground truth saliency maps were transformed producing new artificial image examples. Though the algorithm was originally applied to second-trimester ultrasound images, we showed that our algorithm can work with two modalities (ultrasound images and eye gaze data) and improve segmentation performance in the first trimester by predicting the eye gaze of sonographers. Results showed that stochastic augmentation policy search method can alleviate over-fitting and improve accuracy of saliency map prediction in first trimester US video.

3. A spatio-temporal convolutional network for video saliency prediction with stochastic augmentation.

Chapter 5 explored temporal connectivity of ultrasound frames and sonographer eye gaze pattern for video saliency prediction. A convolutional LSTMU-Net algorithm was designed on top of a single frame saliency prediction model to account for naturally temporal ultrasound frames in a video and to learn their intra-dependence in a sequence. In addition, sonographer eye gaze pattern recorded prior to prediction of saliency on a single frame was found to aid in the analysis

of eye gaze trajectory (or a distribution of saliency maps) on the future frames. Put simply, based on the gaze pattern from previous frames, the future trajectory of where sonographer may look next is calculated. The architecture design was built to find an optimal number of consecutive video frames that would record sufficient eye-gaze pattern variation to track changes in sonographer gaze well. After evaluation, it was found that a video clip of 6 consecutive frames accounts for a good gaze variation (with more frames adding little useful information). The addition of temporal information before the frame that the saliency map was predicted for provided the best model performance. This resulting model may be suitable for automatic guidance mechanism for real-time first trimester US scanning where the saliency predictions direct sonographer gaze to important anatomy. This is being investigated in an on-going translational study called the PURFECT study.

4. Initial prototype tracking system for freehand 3D ultrasound imaging to visually-assist a sonographer during an ultrasound examination.

Chapter 6 proposed a novel system which provides sonographers with a digital representation of the fetus and probe location at every timestamp. The system design began from processing and synchronising IMU probe motion data with 2D ultrasound images and corresponding human and technical annotations. Human annotations were used to filter out CRL biometry views, i.e. the full size fetus displayed on a frame. Next, 2D fetal masks were segmented using a Nested Hourglass (NHG) fetal segmentation architecture and fetal contours extracted. Due to IMU motion sensor drift and general hardware inconsistencies, multi-sensor fusion and sensor error correction were performed using Madgwick's sensor fusion algorithm. Finally, 2D mask edges were projected into a 3D space and transformed using rotation and position coordinates of IMU at every timestamp. To produce a representative 3D fetal reconstruction, a 3D point cloud was formed to analyze the placement of fetal masks in space. This allowed to identify gaps between masks that needed interpolation to produce a full 3D surface. Delaunay triangulation was used to represent a final 3D surface reconstruction with an addition of an ultrasound probe location.

The contribution of this work is in integration of IMU sensor where the probe motion data is processed and synchronized with 2D ultrasound images and annotations. The transformation of 2D mask edges into a 3D space, followed by the creation of a 3D point cloud and interpolation to fill gaps between masks, represents a novel approach to generating a comprehensive 3D fetal surface. This allows for a more detailed and accurate representation of the fetus. The framework can be extended to perform real-time scanning and provide clinicians with dynamic 3D representations during ultrasound scans.

The application of the proposed system has the potential to significantly reduce the cognitive load on sonographers, the time and effort required to mentally reconstruct and analyse a fetal womb from solely visualizing 2D ultrasound scans, and help digitally pin-point the probe location during a scan. The system may improve the efficiency and accuracy of fetal assessments during ultrasound scans.

7.2 Future Work

All of the research in this thesis is performed with the final aim of translating the work into clinical practice to help guide sonographers to important structures whilst freely navigating around the maternal womb. Below we discuss some possible next steps towards this goal.

Translation of video saliency prediction network.

Video saliency prediction model is related to chapter 5 that builds on chapter 4. The next steps are already on the way. This model may be suitable for automatic guidance mechanism for real-time first trimester US scanning where the saliency predictions direct sonographer gaze to important anatomy. This is being investigated in an on-going translational project called the PURFECT study with the model embedded into an ultrasound system.

Fetal motion correction for 3D reconstruction.

Chapter 6 proposed an initial prototype system which takes as input 2D ultrasound images with IMU-assisted probe motion data and outputs a 3D reconstruction. The pipeline took advantage of the appearance of 2D ultrasound fetal masks and

their placement with respect to a point of contact of a probe with skin. Due to the processing and analysis of only short ultrasound clips to produce a 3D reconstruction, the assumption was that the fetus does not move during the data acquisition process. This allowed for direct 3D modelling by extracting a fetal outline and projecting it into 3D space solely relying on the rotation and positional data.

In video clips recorded over a longer period of time, automated real-time fetal alignment algorithms would be needed due to non-rigid motion. These may be based on the existing motion correction algorithms used in ultrasound imaging to track fetal movement including speckle tracking (analysis of the speckle pattern within the ultrasound images to track tissue movement) [376], block matching (patches within the ultrasound frames are compared to find the best matching block in subsequent frames) [377], optical flow (calculation of the motion vector field between frames based on the intensity changes of pixel values) [378] and others. The developed algorithm would need to estimate the fetal motion and apply corrections during the reconstruction in real-time.

The subsequent significant phase involves a translational study with the objective of integrating the prototype system into an ultrasound machine. This will be followed by a series of experiments to evaluate the system's value and feasibility in a clinical setting.

7.3 Summary

The research developed in this thesis focused on building deep learning based assistive models to assist a sonographer to perform a first trimester scan. The work has explored traditional image processing algorithms, signal processing techniques, and cutting edge deep learning algorithms. A suit of algorithms and self-contained prototype systems were developed with potential for clinical use which may simplify, visually-assist and guide sonographers during a first trimester ultrasound examination. During the developmental stages, a variety of challenges were faced and taken into account to create robust pipelines which can generalise well on unseen data. Future work will aim to further improve the presented approaches through

investigating technical details of the established frameworks, further validating the performance and feasibility in clinical setting.

References

[1] Mourad Gridach et al. "Self-Knowledge Distillation for First Trimester Ultrasound Saliency Prediction". In: *International Workshop on Advances in Simplifying Medical Ultrasound*. Springer. 2022, pp. 117–127.

[2] 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*. Springer. 2015, pp. 234–241.

[3] E Savochkina et al. "First trimester gaze pattern estimation using stochastic augmentation policy search for single frame saliency prediction". In: *MIUA*. Springer, LNCS. 2021, pp. 361–374.

[4] NHS. *FASP ultrasound handbook April 2015*. URL: `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/443865/FASP_ultrasound_handbook_July_2015_090715.pdf`. (accessed: 02.11.2020).

[5] E Albert Reece et al. "The safety of obstetric ultrasonography: Concern for the fetus." In: *Obstetrics and gynecology* 76.1 (1990), pp. 139–146.

[6] Ian Donald, John Macvicar, and TG Brown. "Investigation of abdominal masses by pulsed ultrasound". In: *The Lancet* 271.7032 (1958), pp. 1188–1195.

[7] DAVE FORNELL. *5 Key Trends in New Ultrasound Technology*. URL: `https://www.itnonline.com/article/5-key-trends-new-ultrasound-technology`. (accessed: 02.11.2020).

[8] Ali Luffman. *Advancements in Ultrasound Technology*. URL: `https://www.healthtechzone.com/topics/healthcare/articles/2019/12/19/444053-advancements-ultrasound-technology.htm`. (accessed: 02.11.2020).

[9] philips. *Affinity*. URL: `https://www.philips.co.uk/healthcare/sites/affinity`. (accessed: 02.11.2020).

[10] HealthManagement.org. *RSNA 2018: Radiological Society of North America Meeting*. URL: `https://healthmanagement.org/c/imaging/event/rsna-2018-radiological-society-of-north-america-meeting`. (accessed: 02.11.2020).

[11] Konica Minolta. *SONIMAGE HS1*. URL: `https://www.konicaminolta.com/global-en/healthcare/products/us/hs1/index.html`. (accessed: 02.11.2020).

[12] Jehan N Karim et al. “Systematic review of first-trimester ultrasound screening for detection of fetal structural anomalies and factors that affect screening performance”. In: *Ultrasound in Obstetrics & Gynecology* 50.4 (2017), pp. 429–441.

[13] Christian F Baumgartner et al. “Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 203–211.

[14] Christos P Loizou et al. “Comparative evaluation of despeckle filtering in ultrasound imaging of the carotid artery”. In: *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 52.10 (2005), pp. 1653–1669.

[15] Zhi Yang and Martin D Fox. “Speckle reduction and structure enhancement by multichannel median boosted anisotropic diffusion”. In: *EURASIP Journal on Advances in Signal Processing* 2004 (2004), pp. 1–11.

[16] Ruud JG Van Sloun et al. “Deep learning for super-resolution vascular ultrasound imaging”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 1055–1059.

[17] Kirsten Christensen-Jeffries et al. “Super-resolution ultrasound imaging”. In: *Ultrasound in medicine & biology* 46.4 (2020), pp. 865–891.

[18] Mohammad Ali Maraci et al. “Searching for structures of interest in an ultrasound video sequence”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2014, pp. 133–140.

[19] Mohammad Yaqub et al. “A deep learning solution for automatic fetal neurosonographic diagnostic plane verification using clinical standard constraints”. In: *Ultrasound in Medicine & Biology* 43.12 (2017), pp. 2925–2933.

[20] Christian F Baumgartner et al. “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.

[21] Hao Chen et al. “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.

[22] Y Gao, Mohammad Ali Maraci, and J Alison Noble. “Describing ultrasound video content using deep convolutional neural networks”. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 787–790.

[23] J Alison Noble. *Reflections on ultrasound image analysis*. 2016.

[24] Karl Oliver Kagan et al. “Impact of bias in crown–rump length measurement at first-trimester screening for trisomy 21”. In: *Ultrasound in obstetrics & gynecology* 40.2 (2012), pp. 135–139.

[25] Hosuk Ryou et al. “Automated 3D ultrasound image analysis for first trimester assessment of fetal health”. In: *Physics in Medicine & Biology* 64.18 (2019), p. 185010.

[26] Laura Detti et al. “Early pregnancy ultrasound measurements and prediction of first trimester pregnancy loss: A logistic model”. In: *Scientific Reports* 10.1 (2020), pp. 1–10.

[27] Yifan Cai et al. “SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1475–1478.

[28] Richard Droste et al. “Ultrasound image representation learning by modeling sonographer visual attention”. In: *International conference on information processing in medical imaging*. Springer. 2019, pp. 592–604.

[29] I Sarris et al. “Intra-and interobserver variability in fetal ultrasound measurements”. In: *Ultrasound in obstetrics & gynecology* 39.3 (2012), pp. 266–273.

[30] Ali Borji. “Saliency prediction in the deep learning era: An empirical investigation”. In: *arXiv preprint arXiv:1810.03716* 10 (2018).

[31] Elizabeth Huynh et al. “Artificial intelligence in radiation oncology”. In: *Nature Reviews Clinical Oncology* 17.12 (2020), pp. 771–781.

[32] Ian Goodfellow et al. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[33] Laura J Brattain et al. “Machine learning for medical ultrasound: status, methods, and future opportunities”. In: *Abdominal radiology* 43 (2018), pp. 786–799.

[34] Jiajun Zhang and Chengqing Zong. “Deep neural networks in machine translation: An overview”. In: *IEEE Intelligent Systems* 5 (2015), pp. 16–25.

[35] Geoffrey Hinton et al. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”. In: *IEEE Signal processing magazine* 29.6 (2012), pp. 82–97.

[36] Ronan Collobert and Jason Weston. “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 160–167.

[37] Yann LeCun et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[39] Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[40] Ruslan Salakhutdinov and Geoffrey Hinton. “Deep boltzmann machines”. In: *Artificial intelligence and statistics*. 2009, pp. 448–455.

[41] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[42] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. “Bidirectional LSTM networks for improved phoneme classification and recognition”. In: *International Conference on Artificial Neural Networks*. Springer. 2005, pp. 799–804.

[43] Andrej Karpathy. *Convolutional Neural Networks (CNNs / ConvNets)*. URL: https://cs231n.github.io/convolutional-networks/. (accessed: 01.11.2020).

[44] Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. "Avoiding overfitting: A survey on regularization methods for convolutional neural networks". In: *ACM Computing Surveys (CSUR)* 54.10s (2022), pp. 1–25.

[45] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1 (2019), pp. 1–48.

[46] Ekin D Cubuk et al. "Randaugment: Practical automated data augmentation with a reduced search space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 702–703.

[47] Asaduz Zaman et al. "Generative approach for data augmentation for deep learning-based bone surface segmentation from ultrasound images". In: *International journal of computer assisted radiology and surgery* 15 (2020), pp. 931–941.

[48] Lok Hin Lee, Yuan Gao, and J. Alison Noble. *Principled Ultrasound Data Augmentation for Classification of Standard Planes*. 2021. arXiv: 2103.07895 [cs.CV].

[49] Sungbin Lim et al. "Fast autoaugment". In: *Advances in Neural Information Processing Systems* 32 (2019).

[50] Cecilia Summers and Michael J Dinneen. "Improved mixed-example data augmentation". In: *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2019, pp. 1262–1270.

[51] Hiroshi Inoue. "Data augmentation by pairing samples for images classification". In: *arXiv preprint arXiv:1801.02929* (2018).

[52] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. "Data augmentation using random image cropping and patching for deep CNNs". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.9 (2019), pp. 2917–2931.

[53] Joseph Lemley, Shabab Bazrafkan, and Peter Corcoran. "Smart augmentation learning an optimal data augmentation strategy". In: *Ieee Access* 5 (2017), pp. 5858–5869.

[54] Ekin D Cubuk et al. "Autoaugment: Learning augmentation policies from data". In: *arXiv preprint arXiv:1805.09501* (2018).

[55] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[56] Daniel Ho et al. "Population based augmentation: Efficient learning of augmentation policy schedules". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2731–2741.

[57] Christopher Bowles et al. "Gan augmentation: Augmenting training data using generative adversarial networks". In: *arXiv preprint arXiv:1810.10863* (2018).

[58] Carl Doersch. "Tutorial on variational autoencoders". In: *arXiv preprint arXiv:1606.05908* (2016).

[59] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[60] Barret Zoph and Quoc V Le. "Neural architecture search with reinforcement learning". In: *arXiv preprint arXiv:1611.01578* (2016).

[61] Dan Ciresan et al. "Deep neural networks segment neuronal membranes in electron microscopy images". In: *Advances in neural information processing systems*. 2012, pp. 2843–2851.

[62] Jifeng Dai, Kaiming He, and Jian Sun. "Convolutional feature masking for joint object and stuff segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3992–4000.

[63] Feng Ning et al. "Toward automatic phenotyping of developing embryos from videos". In: *IEEE Transactions on Image Processing* 14.9 (2005), pp. 1360–1371.

[64] Bharath Hariharan et al. "Simultaneous detection and segmentation". In: *European Conference on Computer Vision*. Springer. 2014, pp. 297–312.

[65] Saurabh Gupta et al. "Learning rich features from RGB-D images for object detection and segmentation". In: *European conference on computer vision*. Springer. 2014, pp. 345–360.

[66] Abdullah F Al-Battal et al. "A CNN segmentation-based approach to object detection and tracking in ultrasound scans with application to the vagus nerve detection". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 3322–3327.

[67] Samir M Badawy et al. "Automatic semantic segmentation of breast tumors in ultrasound images based on combining fuzzy logic and deep learning—A feasibility study". In: *PLoS One* 16.5 (2021), e0251899.

[68] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[69] Lei Cai, Jingyang Gao, and Di Zhao. "A review of the application of deep learning in medical image classification and segmentation". In: *Annals of translational medicine* 8.11 (2020).

[70] Michal Drozdzal et al. "The importance of skip connections in biomedical image segmentation". In: *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[71] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[72] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[73] Bharath Hariharan et al. "Hypercolumns for object segmentation and fine-grained localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 447–456.

[74] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[75] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[76] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[77] Liang-Chieh Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[78] Alberto Garcia-Garcia et al. “A review on deep learning techniques applied to semantic segmentation”. In: *arXiv preprint arXiv:1704.06857* (2017).

[79] Yan Zeng et al. “Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated V-Net”. In: *Journal of Digital Imaging* 34 (2021), pp. 134–148.

[80] Vahid Ashkani Chenarlogh et al. “Fast and accurate U-net model for fetal ultrasound image segmentation”. In: *Ultrasonic Imaging* 44.1 (2022), pp. 25–38.

[81] Shervin Minaee et al. “Image segmentation using deep learning: A survey”. In: *arXiv preprint arXiv:2001.05566* (2020).

[82] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[83] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[84] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[85] KJ Dai and YL R-FCN. “Object detection via region-based fully convolutional networks. arxiv preprint”. In: *arXiv preprint*. 2016.

[86] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[87] Laurent Itti and Christof Koch. “Computational modelling of visual attention”. In: *Nature reviews neuroscience* 2.3 (2001), pp. 194–203.

[88] Lucie Lévêque et al. “State of the art: Eye-tracking studies in medical imaging”. In: *Ieee Access* 6 (2018), pp. 37023–37034.

[89] Mohammad Alsharid et al. “Gaze-assisted automatic captioning of fetal ultrasound videos using three-way multi-modal deep neural networks”. In: *Medical Image Analysis* 82 (2022), p. 102630.

[90] Roy S Hessels et al. "Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers". In: *Royal Society open science* 5.8 (2018), p. 180502.

[91] Alexander Thiele et al. "Neural mechanisms of saccadic suppression". In: *Science* 295.5564 (2002), pp. 2460–2462.

[92] Louisa V Kulke, Janette Atkinson, and Oliver Braddick. "Neural differences between covert and overt attention studied using EEG with simultaneous remote eye tracking". In: *Frontiers in human neuroscience* 10 (2016), p. 592.

[93] Michael F Land and Mary Hayhoe. "In what ways do eye movements contribute to everyday activities?" In: *Vision research* 41.25-26 (2001), pp. 3559–3565.

[94] John M Henderson. "Human gaze control during real-world scene perception". In: *Trends in cognitive sciences* 7.11 (2003), pp. 498–504.

[95] Yashas Rai and Patrick Le Callet. "Visual attention, visual salience, and perceived interest in multimedia applications". In: *Academic Press Library in Signal Processing, Volume 6*. Elsevier, 2018, pp. 113–161.

[96] Neil DB Bruce et al. "On computational modeling of visual saliency: Examining what's right, and what's left". In: *Vision research* 116 (2015), pp. 95–112.

[97] Amelia R Hunt and Alan Kingstone. "Covert and overt voluntary attention: linked or independent?" In: *Cognitive Brain Research* 18.1 (2003), pp. 102–105.

[98] Ziad M Hafed and James J Clark. "Microsaccades as an overt measure of covert attention shifts". In: *Vision research* 42.22 (2002), pp. 2533–2545.

[99] Hee-kyoung Ko, Martina Poletti, and Michele Rucci. "Microsaccades precisely relocate gaze in a high visual acuity task". In: *Nature neuroscience* 13.12 (2010), pp. 1549–1553.

[100] Tilke Judd, Frédo Durand, and Antonio Torralba. "A benchmark of computational models of saliency to predict human fixations". In: (2012).

[101] Ali Borji and Laurent Itti. "State-of-the-art in visual attention modeling". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 185–207.

[102] Ali Borji et al. "Analysis of scores, datasets, and models in visual saliency prediction". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 921–928.

[103] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet". In: *arXiv preprint arXiv:1411.1045* (2014).

[104] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[105] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. "DeepGaze II: Reading fixations from deep features trained on object recognition". In: *arXiv preprint arXiv:1610.01563* (2016).

[106] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[107] Xun Huang et al. “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 262–270.

[108] Junting Pan et al. “Shallow and deep convolutional networks for saliency prediction”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 598–606.

[109] Matthias Kummerer et al. “Understanding low-and high-level contributions to fixation prediction”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4789–4798.

[110] Wenguan Wang and Jianbing Shen. “Deep visual attention prediction”. In: *IEEE Transactions on Image Processing* 27.5 (2017), pp. 2368–2378.

[111] Shaojing Fan et al. “Emotional attention: A study of image sentiment and visual attention”. In: *Proceedings of the IEEE Conference on computer vision and pattern recognition*. 2018, pp. 7521–7531.

[112] Marcella Cornia et al. “A deep multi-level network for saliency prediction”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 3488–3493.

[113] Zhaohui Che et al. “How is gaze influenced by image transformations? dataset and model”. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 2287–2300.

[114] Alexander Kroner et al. “Contextual encoder–decoder network for visual saliency prediction”. In: *Neural Networks* 129 (2020), pp. 261–270.

[115] Sen Jia and Neil DB Bruce. “Eml-net: An expandable multi-layer network for saliency prediction”. In: *Image and vision computing* 95 (2020), p. 103887.

[116] Zoya Bylinskii et al. “Mit saliency benchmark”. In: (2015).

[117] Jennifer T Coull. “fMRI studies of temporal attention: allocating attention within, or towards, time”. In: *Cognitive Brain Research* 21.2 (2004), pp. 216–226.

[118] Jin Chen et al. “Video Saliency Prediction Using Enhanced Spatiotemporal Alignment Network”. In: *arXiv preprint arXiv:2001.00292* (2020).

[119] Wenguan Wang, Jianbing Shen, and Ling Shao. “Video salient object detection via fully convolutional networks”. In: *IEEE Transactions on Image Processing* 27.1 (2017), pp. 38–49.

[120] Mohammad Shokri, Ahad Harati, and Kimya Taba. “Salient object detection in video using deep non-local neural networks”. In: *Journal of Visual Communication and Image Representation* 68 (2020), p. 102769.

[121] Meijun Sun et al. “SG-FCN: A motion and memory-based deep learning model for video saliency detection”. In: *IEEE transactions on cybernetics* 49.8 (2018), pp. 2900–2911.

[122] Panagiotis Linardos et al. “Simple vs complex temporal recurrences for video saliency prediction”. In: *arXiv preprint arXiv:1907.01869* (2019).

[123] Souad Chaabouni et al. “Deep learning for saliency prediction in natural video”. In: *arXiv preprint arXiv:1604.08010* (2016).

[124] Richard Droste, Jianbo Jiao, and J Alison Noble. “Unified Image and Video Saliency Modeling”. In: *arXiv preprint arXiv:2003.05477* (2020).

[125] Ruth Jiang, Robin Kleer, and Frank T Piller. "Predicting the future of additive manufacturing: A Delphi study on economic and societal implications of 3D printing for 2030". In: *Technological Forecasting and Social Change* 117 (2017), pp. 84–97.

[126] Alexey Dosovitskiy et al. "Flownet: Learning optical flow with convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2758–2766.

[127] Wenguan Wang et al. "Revisiting video saliency: A large-scale benchmark and a new model". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4894–4903.

[128] X Wu et al. "SalSAC: A video saliency prediction model with shuffled attentions and correlation-based ConvLSTM". In: *AAAI*. Vol. 34. 07. 2020, pp. 12410–12417.

[129] Marcella Cornia et al. "Predicting human eye fixations via an lstm-based saliency attentive model". In: *IEEE Transactions on Image Processing* 27.10 (2018), pp. 5142–5154.

[130] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. 2015, pp. 2048–2057.

[131] Hugo Larochelle and Geoffrey E Hinton. "Learning to combine foveal glimpses with a third-order Boltzmann machine". In: *Advances in neural information processing systems*. 2010, pp. 1243–1251.

[132] Yaran Chen et al. "A visual attention based convolutional neural network for image classification". In: *2016 12th World Congress on Intelligent Control and Automation (WCICA)*. IEEE. 2016, pp. 764–769.

[133] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. "Recurrent models of visual attention". In: *Advances in neural information processing systems*. 2014, pp. 2204–2212.

[134] Yan Luo, Ming Jiang, and Qi Zhao. "Visual attention in multi-label image classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

[135] Craig A Beam et al. "The place of medical image perception in 21st-century health care". In: *Journal of the American College of Radiology* 3.6 (2006), pp. 409–412.

[136] Harold L Kundel. "History of research in medical image perception". In: *Journal of the American college of radiology* 3.6 (2006), pp. 402–408.

[137] Calvin F Nodine and Harold L Kundel. "Using eye movements to study visual search and to improve tumor detection." In: *Radiographics* 7.6 (1987), pp. 1241–1250.

[138] Sophie Voisin et al. "Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography". In: *Medical physics* 40.10 (2013), p. 101906.

[139] Georgia Tourassi et al. "Investigating the link between radiologists' gaze, diagnostic decision, and image content". In: *Journal of the American Medical Informatics Association* 20.6 (2013), pp. 1067–1075.

[140] Raymond Bertram et al. “The effect of expertise on eye movement behaviour in medical image perception”. In: *PloS one* 8.6 (2013), e66169.

[141] Sheng Wang et al. “Follow my eye: using gaze to supervise computer-aided diagnosis”. In: *IEEE Transactions on Medical Imaging* 41.7 (2022), pp. 1688–1698.

[142] Arijit Patra et al. “Efficient ultrasound image analysis models with sonographer gaze assisted distillation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 394–402.

[143] Lior Drukker et al. “Safety indices of ultrasound: adherence to recommendations and awareness during routine obstetric ultrasound scanning”. In: *Ultraschall in der Medizin-European Journal of Ultrasound* 41.02 (2020), pp. 138–145.

[144] Chang Peng et al. “Recent Advances in Tracking Devices for Biomedical Ultrasound Imaging Applications”. In: *Micromachines* 13.11 (2022), p. 1855.

[145] Alessandro Filippeschi et al. “Survey of motion tracking methods based on inertial sensors: A focus on upper limb human motion”. In: *Sensors* 17.6 (2017), p. 1257.

[146] Qianqian Cai et al. “Inertial Measurement Unit Assisted Ultrasonic Tracking System for Ultrasound Probe Localization”. In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* (2022).

[147] Manon Kok, Jeroen D Hol, and Thomas B Schön. “Using inertial sensors for position and orientation estimation”. In: *arXiv preprint arXiv:1704.06053* (2017).

[148] Renu Bhardwaj, Neelesh Kumar, and Vipan Kumar. “Errors in micro-electro-mechanical systems inertial measurement and a review on present practices of error modelling”. In: *Transactions of the Institute of Measurement and Control* 40.9 (2018), pp. 2843–2854.

[149] Milad Nazarahari and Hossein Rouhani. “40 years of sensor fusion for orientation tracking via magnetic and inertial measurement units: Methods, lessons learned, and future challenges”. In: *Information Fusion* 68 (2021), pp. 67–84.

[150] Ariadna Valldeperes et al. “Wireless inertial measurement unit (IMU)-based posturography”. In: *European Archives of Oto-Rhino-Laryngology* 276 (2019), pp. 3057–3065.

[151] William E Lorensen and Harvey E Cline. “Marching cubes: A high resolution 3D surface construction algorithm”. In: *ACM siggraph computer graphics* 21.4 (1987), pp. 163–169.

[152] Gerhard Kurz et al. “Recursive estimation of orientation based on the Bingham distribution”. In: *Proceedings of the 16th International Conference on Information Fusion*. IEEE. 2013, pp. 1487–1494.

[153] William Rowan Hamilton. “Xi. on quaternions; or on a new system of imaginaries in algebra”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 33.219 (1848), pp. 58–60.

[154] Jack B Kuipers. *Quaternions and rotation sequences: a primer with applications to orbits, aerospace, and virtual reality*. Princeton university press, 1999.

[155] Jeroen D Hol. “Sensor fusion and calibration of inertial sensors, vision, ultra-wideband and GPS”. PhD thesis. Linköping University Electronic Press, 2011.

[156] John-Olof Nilsson and Isaac Skog. "Inertial sensor arrays—A literature review". In: *2016 European Navigation Conference (ENC)*. IEEE. 2016, pp. 1–10.

[157] Norhafizan Ahmad et al. "Reviews on various inertial measurement unit (IMU) sensor applications". In: *International Journal of Signal Processing Systems* 1.2 (2013), pp. 256–262.

[158] Marco Caruso et al. "Analysis of the accuracy of ten algorithms for orientation estimation using inertial and magnetic sensing under optimal conditions: One size does not fit all". In: *Sensors* 21.7 (2021), p. 2543.

[159] Roberto G Valenti, Ivan Dryanovski, and Jizhong Xiao. "Keeping a good attitude: A quaternion-based orientation filter for IMUs and MARGs". In: *Sensors* 15.8 (2015), pp. 19302–19330.

[160] Angelo M Sabatini. "Quaternion-based extended Kalman filter for determining orientation by inertial and magnetic sensing". In: *IEEE transactions on Biomedical Engineering* 53.7 (2006), pp. 1346–1356.

[161] Henk J Luinge and Peter H Veltink. "Inclination measurement of human movement using a 3-D accelerometer with autocalibration". In: *IEEE Transactions on neural systems and rehabilitation engineering* 12.1 (2004), pp. 112–121.

[162] Robert Mahony, Tarek Hamel, and Jean-Michel Pflimlin. "Nonlinear complementary filters on the special orthogonal group". In: *IEEE Transactions on automatic control* 53.5 (2008), pp. 1203–1218.

[163] Alexander D Young. "Comparison of orientation filter algorithms for realtime wireless inertial posture tracking". In: *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*. IEEE. 2009, pp. 59–64.

[164] Hassen Fourati et al. "Complementary observer for body segments motion capturing by inertial and magnetic sensors". In: *IEEE/ASME transactions on Mechatronics* 19.1 (2012), pp. 149–157.

[165] Sebastian OH Madgwick, Andrew JL Harrison, and Ravi Vaidyanathan. "Estimation of IMU and MARG orientation using a gradient descent algorithm". In: *2011 IEEE international conference on rehabilitation robotics*. IEEE. 2011, pp. 1–7.

[166] Alberto Olivares et al. "Using frequency analysis to improve the precision of human body posture algorithms based on Kalman filters". In: *Computers in Biology and Medicine* 72 (2016), pp. 229–238.

[167] Michael B Del Rosario et al. "Computationally efficient adaptive error-state Kalman filter for attitude estimation". In: *IEEE Sensors Journal* 18.22 (2018), pp. 9332–9342.

[168] Mahdi Abolfazli Esfahani et al. "OriNet: Robust 3-D orientation estimation with a single particular IMU". In: *IEEE Robotics and Automation Letters* 5.2 (2019), pp. 399–406.

[169] Álvaro Deibe et al. "A Kalman Filter for nonlinear attitude estimation using time variable matrices and quaternions". In: *Sensors* 20.23 (2020), p. 6731.

[170] Rudolph Emil Kalman. "A new approach to linear filtering and prediction problems". In: (1960).

[171] Elena Bergamini et al. “Estimating orientation using magnetic and inertial sensors and different sensor fusion approaches: Accuracy assessment in manual and locomotion tasks”. In: *Sensors* 14.10 (2014), pp. 18625–18649.

[172] Karina Lebel et al. “Inertial measures of motion for clinical biomechanics: Comparative assessment of accuracy under controlled conditions–changes in accuracy over time”. In: *PloS one* 10.3 (2015), e0118361.

[173] Luca Ricci, Fabrizio Taffoni, and Domenico Formica. “On the orientation error of IMU: Investigating static and dynamic accuracy targeting human motion”. In: *PloS one* 11.9 (2016), e0161940.

[174] Simone A Ludwig and Kaleb D Burnham. “Comparison of Euler estimate using extended Kalman filter, Madgwick and Mahony on quadcopter flight data”. In: *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE. 2018, pp. 1236–1241.

[175] Young Soo Suh. “Inertial sensor-based smoother for gait analysis”. In: *Sensors* 14.12 (2014), pp. 24338–24357.

[176] Huu Toan Duong and Young Soo Suh. “A simple smoother for attitude and position estimation using inertial sensor”. In: *International Journal of Control, Automation and Systems* 14.6 (2016), pp. 1626–1630.

[177] Vadim Indelman et al. “Factor graph based incremental smoothing in inertial navigation systems”. In: *2012 15th International Conference on Information Fusion*. IEEE. 2012, pp. 2154–2161.

[178] S Zafer, G Sinan, and G Ismail. “Ultra-wideband positioning systems: Theoretical limits”. In: *Ranging Algorithms Protoc* 10 (2008), pp. 1–5.

[179] John Mattingley and Stephen Boyd. “Real-time convex optimization in signal processing”. In: *IEEE Signal processing magazine* 27.3 (2010), pp. 50–61.

[180] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[181] Young Soo Suh, Nguyen Ho Quoc Phuong, and Hee Jun Kang. “Distance estimation using inertial sensor and vision”. In: *International Journal of Control, Automation and Systems* 11.1 (2013), pp. 211–215.

[182] Keisuke Fujii. “Extended kalman filter”. In: *Refernce Manual* 14 (2013).

[183] Bradley M Bell. “The iterated Kalman smoother as a Gauss–Newton method”. In: *SIAM Journal on Optimization* 4.3 (1994), pp. 626–636.

[184] Andrew H Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[185] Peter Corke, Jorge Lobo, and Jorge Dias. *An introduction to inertial and visual sensing*. 2007.

[186] Jeroen D Hol et al. “Robust real-time tracking by fusing measurements from inertial and vision sensors”. In: *Journal of Real-Time Image Processing* 2 (2007), pp. 149–160.

[187] Mingyang Li and Anastasios I Mourikis. “High-precision, consistent EKF-based visual-inertial odometry”. In: *The International Journal of Robotics Research* 32.6 (2013), pp. 690–711.

[188] Agostino Martinelli. "Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination". In: *IEEE Transactions on Robotics* 28.1 (2011), pp. 44–60.

[189] Daniel Roetenberg et al. "Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation". In: *IEEE Transactions on neural systems and rehabilitation engineering* 13.3 (2005), pp. 395–405.

[190] Richard Droste et al. "Automatic probe movement guidance for freehand obstetric ultrasound". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 583–592.

[191] Cheng Zhao et al. "Visual-Assisted Probe Movement Guidance for Obstetric Ultrasound Scanning Using Landmark Retrieval". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 670–679.

[192] Pak-Hei Yeung et al. "Learning to map 2D ultrasound images into 3D space with minimal human annotation". In: *Medical Image Analysis* 70 (2021), p. 101998.

[193] Xiaoping Yun et al. "Self-contained position tracking of human movement using small inertial/magnetic sensor modules". In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE. 2007, pp. 2526–2533.

[194] Taehyung Kim et al. "Versatile low-cost volumetric 3D ultrasound imaging using gimbal-assisted distance sensors and an inertial measurement unit". In: *Sensors* 20.22 (2020), p. 6613.

[195] Niko Pagoulatos, David R Haynor, and Yongmin Kim. "A fast calibration method for 3-D tracking of ultrasound images using a spatial localizer". In: *Ultrasound in medicine & biology* 27.9 (2001), pp. 1219–1229.

[196] Ashrani Aizzuddin Abdul Rahni, I Yahya, and SM Mustaza. "2D translation from a 6-DOF MEMS IMU's orientation for freehand 3D ultrasound scanning". In: *4th Kuala Lumpur International Conference on Biomedical Engineering 2008: BIOMED 2008 25–28 June 2008 Kuala Lumpur, Malaysia*. Springer. 2008, pp. 699–702.

[197] Mohammad Hamed Mozaffari and Won-Sook Lee. "Freehand 3-D ultrasound imaging: a systematic review". In: *Ultrasound in medicine & biology* 43.10 (2017), pp. 2099–2124.

[198] Aaron Fenster, Donal B Downey, and H Neale Cardinal. "Three-dimensional ultrasound imaging". In: *Physics in medicine & biology* 46.5 (2001), R67.

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

[200] Ralf E Gebhard, Treniece N Eubanks, and Rachel Meeks. "Three-dimensional ultrasound imaging". In: *Current opinion in anaesthesiology* 28.5 (2015), pp. 583–587.

[201] Andrew Gee et al. "Engineering a freehand 3D ultrasound system". In: *Pattern Recognition Letters* 24.4-5 (2003), pp. 757–777.

[202] Laurence Mercier et al. “A review of calibration techniques for freehand 3-D ultrasound systems”. In: *Ultrasound in medicine & biology* 31.4 (2005), pp. 449–471.

[203] Po-Wei Hsu et al. “Freehand 3D ultrasound calibration: a review”. In: *Advanced imaging in biology and medicine: technology, software environments, applications* (2009), pp. 47–84.

[204] Ole Vegard Solberg et al. “Freehand 3D ultrasound reconstruction algorithms—a review”. In: *Ultrasound in medicine & biology* 33.7 (2007), pp. 991–1009.

[205] Honggang Yu et al. “A 3D freehand ultrasound system for multi-view reconstructions from sparse 2D scanning planes”. In: *Biomedical engineering online* 10.1 (2011), pp. 1–22.

[206] Harshita Sharma et al. “Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 987–990.

[207] Cai et al. “Spatio-temporal visual attention modelling of standard biometry plane-finding navigation”. In: *MIA* 65 (2020), p. 101762.

[208] Richard Droste et al. “Discovering Salient Anatomical Landmarks by Predicting Human Gaze”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1711–1714.

[209] Haitao Gao et al. “Wireless and sensorless 3D ultrasound imaging”. In: *Neurocomputing* 195 (2016), pp. 159–171.

[210] Ben-Mei Chen, Ling-Wei Xia, and Rui-Qin Zhao. “Determination of NG, NG-dimethylarginine in human plasma by high-performance liquid chromatography”. In: *Journal of Chromatography B: Biomedical Sciences and Applications* 692.2 (1997), pp. 467–471.

[211] Richard W Prager, AH Gee, and Laurence Berman. “Stradx: real-time acquisition and visualisation of freehand 3D ultrasound”. In: (1998).

[212] Yipei Wang et al. “Differentiating operator skill during routine fetal ultrasound scanning using probe motion tracking”. In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer. 2020, pp. 180–188.

[213] Neeraj Sharma and Lalit M Aggarwal. “Automated medical image segmentation techniques”. In: *Journal of medical physics/Association of Medical Physicists of India* 35.1 (2010), p. 3.

[214] Harshita Sharma et al. “Multi-modal learning from video, eye tracking, and pupillometry for operator skill characterization in clinical fetal ultrasound”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1646–1649.

[215] Harshita Sharma et al. “Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos”. In: *Medical Image Analysis* 69 (2021), p. 101973.

[216] Harshita Sharma et al. “Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging”. In: *Computers in biology and medicine* 135 (2021), p. 104589.

[217] Farhan Mohamed and C Vei Siang. “A survey on 3D ultrasound reconstruction techniques”. In: *Artificial Intelligence—Applications in Medicine and Biology* (2019), pp. 73–92.

[218] Paola Clauser et al. “Comparison between different imaging techniques in the evaluation of malignant breast lesions: can 3D ultrasound be useful?” In: *La radiologia medica* 119 (2014), pp. 240–248.

[219] DO Watermann et al. “Three-dimensional ultrasound for the assessment of breast lesions”. In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 25.6 (2005), pp. 592–598.

[220] D Rotten, JM Levaillant, and L Zerat. “Analysis of normal breast tissue and of solid breast masses using three-dimensional ultrasound mammography”. In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 14.2 (1999), pp. 114–124.

[221] GC Meyberg-Solomayer et al. “Does 3-D sonography bring any advantage to noninvasive breast diagnostics?” In: *Ultrasound in medicine & biology* 30.5 (2004), pp. 583–589.

[222] Nariya Cho et al. “Differentiating benign from malignant solid breast masses: comparison of two-dimensional and three-dimensional US”. In: *Radiology* 240.1 (2006), pp. 26–32.

[223] Rafal Zenon Slapa et al. “Advantages and disadvantages of 3D ultrasound of thyroid nodules including thin slice volume rendering”. In: *Thyroid research* 4 (2011), pp. 1–12.

[224] Janet I Vaughan. “Against–3D ultrasound in first and second trimester pregnancy–hype or helpful?” In: *Australasian Journal of Ultrasound in Medicine* 12.3 (2009), p. 32.

[225] Matthew R Morgan et al. “Versatile low-cost volumetric 3-D ultrasound platform for existing clinical 2-D systems”. In: *IEEE transactions on medical imaging* 37.10 (2018), pp. 2248–2256.

[226] J-M Bureau, W Steichen, and G Lebail. “A two-dimensional transducer array for real-time 3D medical ultrasound imaging”. In: *1998 IEEE Ultrasonics Symposium. Proceedings (Cat. No. 98CH36102)*. Vol. 2. IEEE. 1998, pp. 1065–1068.

[227] ED Light et al. “Progress in 2-D arrays for real time volumetric imaging”. In: *Ultrasonic imaging* 20 (1998), pp. 235–250.

[228] Edward D Light et al. “Update of two dimensional arrays for real time volumetric and real time intracardiac imaging”. In: *1999 IEEE Ultrasonics Symposium. Proceedings. International Symposium (Cat. No. 99CH37027)*. Vol. 2. IEEE. 1999, pp. 1217–1220.

[229] Jesse T Yen, Jordan P Steinberg, and Stephen W Smith. “Sparse 2-D array design for real time rectilinear volumetric imaging”. In: *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 47.1 (2000), pp. 93–110.

[230] Jesse T Yen and Stephen W Smith. "Real-time rectilinear 3-D ultrasound using receive mode multiplexing". In: *ieee transactions on ultrasonics, ferroelectrics, and frequency control* 51.2 (2004), pp. 216–226.

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

[232] Graham M Treece et al. "High-definition freehand 3-D ultrasound". In: *Ultrasound in medicine & biology* 29.4 (2003), pp. 529–546.

[233] François Rousseau, Pierre Hellier, and Christian Barillot. "Confhusius: A robust and fully automatic calibration method for 3D freehand ultrasound". In: *Medical image analysis* 9.1 (2005), pp. 25–38.

[234] Mahani Hafizah, Tan Kok, and EKO Supriyanto. "Development of 3D image reconstruction based on untracked 2D fetal phantom ultrasound images using VTK". In: *WSEAS transactions on signal processing* 6.4 (2010), pp. 145–154.

[235] Yipeng Hu et al. "Development and phantom validation of a 3-D-ultrasound-guided system for targeting MRI-visible lesions during transrectal prostate biopsy". In: *IEEE Transactions on Biomedical Engineering* 64.4 (2016), pp. 946–958.

[236] Kenneth Strømmen et al. "Volume estimation of small phantoms and rat kidneys using three-dimensional ultrasonography and a position sensor". In: *Ultrasound in medicine & biology* 30.9 (2004), pp. 1109–1117.

[237] Odd Helge Gilja et al. "In vitro evaluation of three-dimensional ultrasonography based on magnetic scanhead tracking". In: *Ultrasound in medicine & biology* 24.8 (1998), pp. 1161–1167.

[238] Paul R Detmer et al. "3D ultrasonic image feature localization based on magnetic scanhead tracking: in vitro calibration and validation". In: *Ultrasound in medicine & biology* 20.9 (1994), pp. 923–936.

[239] Krishnaswamy Chandrasekaran et al. "Three-dimensional volumetric ultrasound imaging of arterial pathology from two-dimensional intravascular ultrasound: an in vitro study". In: *Angiology* 45.4 (1994), pp. 253–264.

[240] Harm-Gerd Blaas et al. "In-vivo three-dimensional ultrasound reconstructions of embryos and early fetuses". In: *The Lancet* 352.9135 (1998), pp. 1182–1186.

[241] Andreas Wahle et al. "Fusion of angiography and intravascular ultrasound in vivo: establishing the absolute 3-D frame orientation". In: *IEEE Transactions on Biomedical Engineering* 46.10 (1999), pp. 1176–1180.

[242] AW Welsh et al. "Freehand three-dimensional Doppler demonstration of monochorionic vascular anastomoses in vivo: a preliminary report". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 18.4 (2001), pp. 317–324.

[243] Yiannis S Chatzizisis et al. "In-vivo accuracy of geometrically correct three-dimensional reconstruction of human coronary arteries: is it influenced by certain parameters?" In: *Coronary artery disease* 17.6 (2006), pp. 545–551.

[244] Thomas J MacGillivray et al. "3D freehand ultrasound for in vivo determination of human skeletal muscle volume". In: *Ultrasound in medicine & biology* 35.6 (2009), pp. 928–935.

[245] Dominic James Farris et al. “Differential strain patterns of the human Achilles tendon determined in vivo with freehand three-dimensional ultrasound imaging”. In: *Journal of Experimental Biology* 216.4 (2013), pp. 594–600.

[246] Steven J Obst, Richard Newsham-West, and Rod S Barrett. “In vivo measurement of human achilles tendon morphology using freehand 3-D ultrasound”. In: *Ultrasound in medicine & biology* 40.1 (2014), pp. 62–70.

[247] Elyse Passmore et al. “Measuring femoral torsion in vivo using freehand 3-D ultrasound imaging”. In: *Ultrasound in Medicine & Biology* 42.2 (2016), pp. 619–623.

[248] Steven J Obst et al. “Reliability of Achilles tendon moment arm measured in vivo using freehand three-dimensional ultrasound”. In: *Journal of Applied Biomechanics* 33.4 (2017), pp. 300–304.

[249] Kerem Karadayi, Ravi Managuli, and Yongmin Kim. “Three-dimensional ultrasound: from acquisition to visualization and from algorithms to systems”. In: *IEEE Reviews in Biomedical Engineering* 2 (2009), pp. 23–39.

[250] Daniel H Turnbull and F Stuart Foster. “Fabrication and characterization of transducer elements in two-dimensional arrays for medical ultrasound imaging”. In: *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 39.4 (1992), pp. 464–475.

[251] QH Huang et al. “Development of a portable 3D ultrasound imaging system for musculoskeletal tissues”. In: *Ultrasonics* 43.3 (2005), pp. 153–163.

[252] Dónal B Downey, Aaron Fenster, and Jacqueline C Williams. “Clinical utility of three-dimensional US”. In: *Radiographics* 20.2 (2000), pp. 559–571.

[253] Mohammad I Daoud et al. “Freehand 3D ultrasound imaging system using electromagnetic tracking”. In: *2015 International Conference on Open Source Software Computing (OSSCOM)*. IEEE. 2015, pp. 1–5.

[254] Tiexiang Wen et al. “An accurate and effective FMM-based approach for freehand 3D ultrasound reconstruction”. In: *Biomedical Signal Processing and Control* 8.6 (2013), pp. 645–656.

[255] Hengtao Guo et al. “Sensorless freehand 3D ultrasound reconstruction via deep contextual learning”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer. 2020, pp. 463–472.

[256] Pak-Hei Yeung et al. “ImplicitVol: sensorless 3D ultrasound reconstruction with deep implicit representation”. In: *arXiv preprint arXiv:2109.12108* (2021).

[257] Mingyuan Luo et al. “Deep Motion Network for Freehand 3D Ultrasound Reconstruction”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*. Springer. 2022, pp. 290–299.

[258] Xiankang Chen et al. “Reconstruction of freehand 3D ultrasound based on kernel regression”. In: *Biomedical engineering online* 13.1 (2014), pp. 1–15.

[259] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era". In: *IEEE transactions on pattern analysis and machine intelligence* 43.5 (2019), pp. 1578–1604.

[260] Raphael Prevost et al. "Deep learning for sensorless 3D freehand ultrasound imaging". In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II*. Springer. 2017, pp. 628–636.

[261] Wolfgang Wein et al. "Three-dimensional thyroid assessment from untracked 2D ultrasound clips". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer. 2020, pp. 514–523.

[262] Hassan Rivaz, Emad Boctor, and Gabor Fichtinger. "A robust meshing and calibration approach for sensorless freehand 3d ultrasound". In: *Medical Imaging 2007: Ultrasonic Imaging and Signal Processing*. Vol. 6513. SPIE. 2007, pp. 378–385.

[263] Koichi Ito et al. "A probe-camera system for 3D ultrasound image reconstruction". In: *Imaging for Patient-Customized Simulations and Systems for Point-of-Care Ultrasound: International Workshops, BIVPCS 2017 and POCUS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings*. Springer. 2017, pp. 129–137.

[264] Theresa A Tuthill et al. "Automated three-dimensional US frame positioning computed from elevational speckle decorrelation." In: *Radiology* 209.2 (1998), pp. 575–582.

[265] Ruey-Feng Chang et al. "3-D US frame positioning using speckle decorrelation and image registration". In: *Ultrasound in medicine & biology* 29.6 (2003), pp. 801–812.

[266] Narges Afsham et al. "Nonlocal means filter-based speckle tracking". In: *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 62.8 (2015), pp. 1501–1515.

[267] Catherine Laporte and Tal Arbel. "Learning to estimate out-of-plane motion in ultrasound imagery of real tissue". In: *Medical image analysis* 15.2 (2011), pp. 202–213.

[268] Richard W Prager et al. "Sensorless freehand 3-D ultrasound using regression of the echo intensity". In: *Ultrasound in medicine & biology* 29.3 (2003), pp. 437–446.

[269] 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.

[270] Raphael Prevost et al. "3D freehand ultrasound without external tracking using deep learning". In: *Medical image analysis* 48 (2018), pp. 187–202.

[271] Richard James Housden et al. "Calibration of an orientation sensor for freehand 3D ultrasound and its use in a hybrid acquisition system". In: *BioMedical Engineering OnLine* 7 (2008), pp. 1–13.

[272] Hengtao Guo et al. “Transducer Adaptive Ultrasound Volume Reconstruction”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 511–515.

[273] Hengtao Guo et al. “Ultrasound Volume Reconstruction From Freehand Scans Without Tracking”. In: *IEEE Transactions on Biomedical Engineering* (2022).

[274] Mingyuan Luo et al. “Self context and shape prior for sensorless freehand 3D ultrasound reconstruction”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*. Springer. 2021, pp. 201–210.

[275] Hongliang Ren et al. “Multisensor data fusion in an integrated tracking system for endoscopic surgery”. In: *IEEE Transactions on Information Technology in Biomedicine* 16.1 (2011), pp. 106–111.

[276] Qing-Hua Huang et al. “Linear tracking for 3-D medical ultrasound imaging”. In: *IEEE transactions on cybernetics* 43.6 (2013), pp. 1747–1754.

[277] Florin Tatar. “Ultrasound 3D positioning system for surgical instruments”. In: (2006).

[278] Dyah Ekashanti Octorina Dewi et al. “Position tracking systems for ultrasound imaging: A survey”. In: *Medical imaging technology: Reviews and computational applications* (2015), pp. 57–89.

[279] Sevald Berg et al. “Dynamic three-dimensional freehand echocardiography using raw digital ultrasound data”. In: *Ultrasound in medicine & biology* 25.5 (1999), pp. 745–753.

[280] Warren S Edwards, Christian Deforge, and Yongmin Kim. “Interactive three-dimensional ultrasound using a programmable multimedia processor”. In: *International Journal of Imaging Systems and Technology* 9.6 (1998), pp. 442–454.

[281] SW Hughes et al. “Volume estimation from multiplanar 2D ultrasound images using a remote electromagnetic position and orientation sensor”. In: *Ultrasound in medicine & biology* 22.5 (1996), pp. 561–572.

[282] Daniel F Leotta, Paul R Detmer, and Roy W Martin. “Performance of a miniature magnetic position sensor for three-dimensional ultrasound imaging”. In: *Ultrasound in medicine & biology* 23.4 (1997), pp. 597–609.

[283] Stephen Meairs, Jens Beyer, and Michael Hennerici. “Reconstruction and visualization of irregularly sampled three-and four-dimensional ultrasound data for cerebrovascular applications”. In: *Ultrasound in medicine & biology* 26.2 (2000), pp. 263–272.

[284] Karen Altmann et al. “Comparison of three-dimensional echocardiographic assessment of volume, mass, and function in children with functionally single left ventricles with two-dimensional echocardiography and magnetic resonance imaging”. In: *The American journal of cardiology* 80.8 (1997), pp. 1060–1065.

[285] Donald L King, Donald L King Jr, and MY Shao. “Evaluation of in vitro measurement accuracy of a three-dimensional ultrasound scanner.” In: *Journal of Ultrasound in Medicine* 10.2 (1991), pp. 77–82.

[286] Donald L King et al. “Three-dimensional echocardiography. Advances for measurement of ventricular volume and mass.” In: *Hypertension* 23.1_supplement (1994), p. I172.

[287] Andrei State et al. “Observing a volume rendered fetus within a pregnant patient”. In: *Proceedings Visualization'94*. IEEE. 1994, pp. 364–368.

[288] Jason W Trobaugh, Darin J Trobaugh, and William D Richard. “Three-dimensional imaging with stereotactic ultrasonography”. In: *Computerized Medical Imaging and Graphics* 18.5 (1994), pp. 315–323.

[289] Ryutarou Ohbuchi, David Chen, and Henry Fuchs. “Incremental volume reconstruction and rendering for 3-D ultrasound imaging”. In: *Visualization in Biomedical Computing'92*. Vol. 1808. SPIE. 1992, pp. 312–323.

[290] Yakang Dai et al. “A qualitative and quantitative interaction technique for freehand 3D ultrasound imaging”. In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2006, pp. 2750–2753.

[291] Zhenping Chen and Qinghua Huang. “Real-time freehand 3D ultrasound imaging”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.1 (2018), pp. 74–83.

[292] J Nerney Welch et al. “A real-time freehand 3D ultrasound system for image-guided surgery”. In: *2000 IEEE Ultrasonics Symposium. Proceedings. An International Symposium (Cat. No. 00CH37121)*. Vol. 2. IEEE. 2000, pp. 1601–1604.

[293] Carl D Herickhoff et al. “Low-cost volumetric ultrasound by augmentation of 2D systems: Design and prototype”. In: *Ultrasonic imaging* 40.1 (2018), pp. 35–48.

[294] Carl Herickhoff, Junhong Lin, and Jeremy Dahl. “Low-cost sensor-enabled freehand 3D ultrasound”. In: *2019 IEEE International Ultrasonics Symposium (IUS)*. IEEE. 2019, pp. 498–501.

[295] AM Goldsmith, PC Pedersen, and TL Szabo. “An inertial-optical tracking system for portable, quantitative, 3D ultrasound”. In: *2008 IEEE ultrasonics symposium*. IEEE. 2008, pp. 45–49.

[296] Raúl San José-Estépar et al. “A theoretical framework to three-dimensional ultrasound reconstruction from irregularly sampled data”. In: *Ultrasound in medicine & biology* 29.2 (2003), pp. 255–269.

[297] Pierrick Coupé et al. “Probe trajectory interpolation for 3d reconstruction of freehand ultrasound”. In: *Medical image analysis* 11.6 (2007), pp. 604–615.

[298] Robert Rohling, Andrew Gee, and Laurence Berman. “A comparison of freehand three-dimensional ultrasound reconstruction techniques”. In: *Medical image analysis* 3.4 (1999), pp. 339–359.

[299] Wayne Y Zhang, Robert N Rohling, and Dinesh K Pai. “Surface extraction with a three-dimensional freehand ultrasound system”. In: *Ultrasound in medicine & biology* 30.11 (2004), pp. 1461–1473.

[300] CD Barry et al. “Three-dimensional freehand ultrasound: image reconstruction and volume analysis”. In: *Ultrasound in medicine & biology* 23.8 (1997), pp. 1209–1224.

[301] João M Sanches and Jorge S Marques. "A Rayleigh reconstruction/interpolation algorithm for 3D ultrasound". In: *Pattern recognition letters* 21.10 (2000), pp. 917–926.

[302] Qing-Hua Huang and Yong-Ping Zheng. "Volume reconstruction of freehand three-dimensional ultrasound using median filters". In: *Ultrasonics* 48.3 (2008), pp. 182–192.

[303] Pierrick Coupé et al. "3D freehand ultrasound reconstruction based on probe trajectory". In: *8th International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 3749. Springer. 2005, pp. 597–604.

[304] Lu**********s F Gonçalves et al. "Three-and 4-dimensional ultrasound in obstetric practice: does it help?" In: *Journal of Ultrasound in Medicine* 24.12 (2005), pp. 1599–1624.

[305] Harm-Gerd Karl Blaas, Sturla Hall Eik-Nes, and Sevald Berg. "Three-dimensional fetal ultrasound". In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 14.4 (2000), pp. 611–627.

[306] Thomas R Nelson and Dolores H Pretorius. "Interactive acquisition, analysis, and visualization of sonographic volume data". In: *International Journal of Imaging Systems and Technology* 8.1 (1997), pp. 26–37.

[307] Robert Rohling, Andrew Gee, and Laurence Berman. "Three-dimensional spatial compounding of ultrasound images". In: *Medical Image Analysis* 1.3 (1997), pp. 177–193.

[308] James A Sethian. "Fast marching methods". In: *SIAM review* 41.2 (1999), pp. 199–235.

[309] Shi Sherebrin et al. "Freehand three-dimensional ultrasound: implementation and applications". In: *Medical Imaging 1996: Physics of Medical Imaging*. Vol. 2708. SPIE. 1996, pp. 296–303.

[310] Qinghua Huang et al. "A new adaptive interpolation algorithm for 3D ultrasound imaging with speckle reduction and edge preservation". In: *Computerized Medical Imaging and Graphics* 33.2 (2009), pp. 100–110.

[311] Qing-Hua Huang and Yong-Ping Zheng. "An adaptive squared-distance-weighted interpolation for volume reconstruction in 3D freehand ultrasound". In: *Ultrasonics* 44 (2006), e73–e77.

[312] Qinghua Huang et al. "Speckle suppression and contrast enhancement in reconstruction of freehand 3D ultrasound images using an adaptive distance-weighted method". In: *Applied Acoustics* 70.1 (2009), pp. 21–30.

[313] Thomas R Nelson and T Todd Elvins. "Visualization of 3D ultrasound data". In: *IEEE Computer Graphics and Applications* 13.6 (1993), pp. 50–57.

[314] Richard W Prager et al. "Three-dimensional ultrasound imaging". In: *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 224.2 (2010), pp. 193–223.

[315] Chung-Wai James Cheung et al. "Ultrasound volume projection imaging for assessment of scoliosis". In: *IEEE transactions on medical imaging* 34.8 (2015), pp. 1760–1768.

[316] David G Gobbi and Terry M Peters. “Interactive intra-operative 3D ultrasound reconstruction and visualization”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part II 5*. Springer. 2002, pp. 156–163.

[317] Jianhao Tan et al. “Design of 3D visualization system based on VTK utilizing marching cubes and ray casting algorithm”. In: *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. Vol. 2. IEEE. 2016, pp. 192–197.

[318] Christian Barillot. “Surface and volume rendering techniques to display 3-D data”. In: *IEEE Engineering in Medicine and Biology Magazine* 12.1 (1993), pp. 111–119.

[319] Marc Levoy. “Display of surfaces from volume data”. In: *IEEE Computer graphics and Applications* 8.3 (1988), pp. 29–37.

[320] Marcos Balsa Rodrı**********guez et al. “State-of-the-art in compressed GPU-based direct volume rendering”. In: *Computer graphics forum*. Vol. 33. 6. Wiley Online Library. 2014, pp. 77–100.

[321] Qi Zhang, Roy Eagleson, and Terry M Peters. “Volume visualization: a technical overview with a focus on medical applications”. In: *Journal of digital imaging* 24 (2011), pp. 640–664.

[322] Bhumika N Parmar and T Bhatt. “Volume visualization using marching cubes algorithms: Survey & analysis”. In: *International Journal of Innovative Research in Technnology* 2.11 (2016), pp. 21–25.

[323] Eric Keppel. “Approximating complex surfaces by triangulation of contour lines”. In: *IBM Journal of research and development* 19.1 (1975), pp. 2–11.

[324] Gregory M Nielson and Bernd Hamann. “The asymptotic decider: resolving the ambiguity in marching cubes.” In: *IEEE visualization*. Vol. 91. 1991, pp. 83–91.

[325] David Mackay. “Robust contour based surface reconstruction algorithms for applications in medical imaging”. In: (2019).

[326] Nicole Schubert et al. “D Polarized Light Imaging Portrayed: Visualization of Fiber Architecture Derived from 3D-PLI”. In: (2018).

[327] Lior Drukker et al. “Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video”. In: *Scientific Reports* 11.1 (2021), p. 14109.

[328] Aris T Papageorghiou et al. “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.

[329] Robail Yasrab et al. “A Machine Learning Method for Automated Description and Workflow Analysis of First Trimester Ultrasound Scans.” In: *IEEE Transactions on Medical Imaging* (2022).

[330] Andrew K Mackenzie and Julie M Harris. “A link between attentional function, effective eye movements, and driving ability.” In: *Journal of experimental psychology: human perception and performance* 43.2 (2017), p. 381.

[331] James H Elder and Steven W Zucker. "Local scale control for edge detection and blur estimation". In: *IEEE Transactions on pattern analysis and machine intelligence* 20.7 (1998), pp. 699–716.

[332] Pierre Chatelain et al. "Evaluation of gaze tracking calibration for longitudinal biomedical imaging studies". In: *IEEE transactions on cybernetics* 50.1 (2018), pp. 153–163.

[333] Ian T Young and Lucas J Van Vliet. "Recursive implementation of the Gaussian filter". In: *Signal processing* 44.2 (1995), pp. 139–151.

[334] John Varley. *Persistence of Vision*. Penguin, 1988.

[335] Junqiao Zhao et al. "Mathematical morphology-based generalization of complex 3D building models incorporating semantic relationships". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 68 (2012), pp. 95–111.

[336] Yunfei Zhang et al. "A hybrid method to incrementally extract road networks using spatio-temporal trajectory data". In: *ISPRS International Journal of Geo-Information* 9.4 (2020), p. 186.

[337] Zach Eaton-Rosen et al. "Improving data augmentation for medical image segmentation". In: (2018).

[338] Behnaz Abdollahi, Naofumi Tomita, and Saeed Hassanpour. "Data augmentation in training deep learning models for medical image analysis". In: *Deep learners and deep learner descriptors for medical applications* (2020), pp. 167–180.

[339] Maayan Frid-Adar et al. "Synthetic data augmentation using GAN for improved liver lesion classification". In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 289–293.

[340] Hanlin Chen and Peng Cao. "Deep learning based data augmentation and classification for limited medical data learning". In: *2019 IEEE international conference on power, intelligent computing and systems (ICPICS)*. IEEE. 2019, pp. 300–303.

[341] Dakai Jin et al. "CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer. 2018, pp. 732–740.

[342] Walid Abdullah Al and Il Dong Yun. *Reinforcing Medical Image Classifier to Improve Generalization on Small Datasets*. 2019. arXiv: `1909.05630 [cs.LG]`.

[343] F Dubost et al. "Data Augmentation for Regression Neural Networks. arXiv 2019". In: *arXiv preprint arXiv:1807.04798* ().

[344] Amy Zhao et al. "Data augmentation using learned transformations for one-shot medical image segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 8543–8553.

[345] A Kosevoi-Tichie et al. *THU0583 does eye gaze tracking have the ability to assess how rheumatologists evaluate musculoskeletal ultrasound images?* 2015.

[346] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[347] Linjun Zhang et al. “How does mixup help with robustness and generalization?” In: *arXiv preprint arXiv:2010.04819* (2020).

[348] Richard Droste et al. “Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction”. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2019, pp. 174–186.

[349] Zoya Bylinskii et al. “What do different evaluation metrics tell us about saliency models?” In: *IEEE transactions on pattern analysis and machine intelligence* 41.3 (2018), pp. 740–757.

[350] A Salvadoret al. “Recurrent neural networks for semantic instance segmentation”. In: *arXiv:1712.00617* (2017).

[351] F Xu et al. “LSTM multi-modal U-Net for brain tumor segmentation”. In: *ICIVC*. IEEE. 2019, pp. 236–240.

[352] Martin G Tolsgaard et al. “Sustained effect of simulation-based ultrasound training on clinical performance: a randomized trial”. In: *Ultrasound in Obstetrics & Gynecology* 46.3 (2015), pp. 312–318.

[353] Rani Ahmad et al. “Impact of high-fidelity transvaginal ultrasound simulation for radiology on residents’ performance and satisfaction”. In: *Academic radiology* 22.2 (2015), pp. 234–239.

[354] Puian Tadayon et al. “Fusion of inertial and magnetic sensors for 3D position and orientation estimation”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016, pp. 3362–3365.

[355] Baljash S Cheema et al. “Artificial intelligence–enabled POCUS in the COVID-19 ICU: A new spin on cardiac ultrasound”. In: *Case Reports* 3.2 (2021), pp. 258–263.

[356] Robail Yasrab et al. “End-to-end first trimester fetal ultrasound video automated CRL and NT segmentation”. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–5.

[357] Paul Berghold. “Sensor shield and fusion algorithm evaluation of DIY Attitude and Heading Reference Systems (AHRS)”. In: ().

[358] Sebastian Madgwick et al. “An efficient orientation filter for inertial and inertial/magnetic sensor arrays”. In: *Report x-io and University of Bristol (UK)* 25 (2010), pp. 113–118.

[359] Bingfei Fan, Qingguo Li, and Tao Liu. “How magnetic disturbance influences the attitude and heading in magnetic and inertial sensor-based orientation estimation”. In: *Sensors* 18.1 (2017), p. 76.

[360] Robert Bieda and Krzysztof Jaskot. “Determinig of an object orientation in 3D space using direction cosine matrix and non-stationary Kalman filter”. In: *Archives of Control Sciences* 26.2 (2016).

[361] Robail Yasrab et al. “RootNav 2.0: Deep learning for automatic navigation of complex plant root architectures”. In: *GigaScience* 8.11 (2019), giz123.

[362] Abdulmalik Shehu Yaro, Filip Maly, and Pavel Prazak. “Outlier Detection in Time-Series Receive Signal Strength Observation Using Z-Score Method with S n Scale Estimator for Indoor Localization”. In: *Applied Sciences* 13.6 (2023), p. 3900.

[363] Diego Nehab. "Medial Axis". In: ().

[364] Dakai Jin et al. "A robust and efficient curve skeletonization algorithm for tree-like objects using minimum cost paths". In: *Pattern recognition letters* 76 (2016), pp. 32–40.

[365] Tilo Strutz. *The Distance Transform and its Computation*. 2023. arXiv: 2106.03503 [cs.CV].

[366] B Vanajakshi, B Sujatha, and K Sri Rama Krishna. "An Analysis of Thinning & Skeletonization for Shape Representation". In: *International Journal of Computer Communication and Information System (IJCCIS)* 2.1 (2010), pp. 250–253.

[367] Punam K Saha, Gunilla Borgefors, and Gabriella Sanniti di Baja. "Skeletonization and its applications–a review". In: *Skeletonization* (2017), pp. 3–42.

[368] Matthew Berger et al. "A survey of surface reconstruction from point clouds". In: *Computer graphics forum*. Vol. 36. 1. Wiley Online Library. 2017, pp. 301–329.

[369] Juho-Pekka Virtanen et al. "Interactive dense point clouds in a game engine". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 163 (2020), pp. 375–389.

[370] Mario Botsch et al. *Polygon mesh processing*. CRC press, 2010.

[371] Charles W Anderson and Steward Crawford-Hines. "Fast generation of nurbs surfaces from polygonal mesh models of human anatomy". In: *Colorado State University Computer Science Technical Report CS-99* 101 (2000).

[372] Frédéric Cazals and Joachim Giesen. "Delaunay triangulation based surface reconstruction". In: *Effective computational geometry for curves and surfaces*. Springer, 2006, pp. 231–276.

[373] Marek Teichmann and Michael Capps. "Surface reconstruction with anisotropic density-scaled alpha shapes". In: *Proceedings Visualization'98 (Cat. No. 98CB36276)*. IEEE. 1998, pp. 67–72.

[374] Fausto Bernardini et al. "The ball-pivoting algorithm for surface reconstruction". In: *IEEE transactions on visualization and computer graphics* 5.4 (1999), pp. 349–359.

[375] Theodore Roosevelt. *The key to success in life*. URL: https://www.theodorerooseveltcenter.org/Research/Digital-Library/Record?libID=o283099.

[376] Jan-Wiebe H Korstanje et al. "Development and validation of ultrasound speckle tracking to quantify tendon displacement". In: *Journal of biomechanics* 43.7 (2010), pp. 1373–1379.

[377] Spyretta Golemati et al. "Comparison of block matching and differential methods for motion analysis of the carotid artery wall from ultrasound images". In: *IEEE Transactions on Information Technology in Biomedicine* 16.5 (2012), pp. 852–858.

[378] Tae-Jin Nam, Rae-Hong Park, and Jae-Ho Yun. "Optical flow based frame interpolation of ultrasound images". In: *Image Analysis and Recognition: Third International Conference, ICIAR 2006, Póvoa de Varzim, Portugal, September 18-20, 2006, Proceedings, Part I 3*. Springer. 2006, pp. 792–803.