Medical Image Analysis for Simplified Ultrasound Protocols

Alexander Darius Gleed

Balliol College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2023

Abstract

Ultrasound is an imaging tool used in obstetrics to identify high-risk pregnancies. However, ultrasound (US) requires a trained operator, who guides a transducer in response to real-time interpretation of video content. In low- and middle-income countries (LMICs), there is a shortage of trained sonographers. In this thesis, we address this key challenge by combining simple US video sweeps with computational algorithms to provide clinical benefit. The sweeps can be taken by an US novice.

First, we design an algorithm that automatically creates an assistive video overlay from a simple video sweep. The overlay assists interpretation of US video to assess placenta location. We describe the design and evaluation of a deep learning-based automatic segmentation model and a statistical data visualisation of 2-D placenta shapes. The data visualisation reveals the spectrum of placenta shapes in this problem space. A probabilistic graphical model is used to improve segmentations with regards to the highly variable placenta shape. From the automatic segmentations, image guidance is created, translating the clinical criteria into assistive visual information.

Second, we explore analysis of multiple video sweeps using graphs. A three-node graph models three video sweeps, where the nodes encode binary sequences representing the fetal head frame-level detection across all video frames in a sweep. To better characterise the sweeps, we perform a statistical analysis of large-scale manual annotations of video sweeps in our dataset. This reveals common patterns of frame-level anatomy occurrence for different video sweep trajectories. Particular insight is gained for patterns that correspond to fetal pose. In this regard, we build a graph convolutional network to automatically classify fetal presentation, using graphs that combine complementary video sweep information relating to fetal pose.

Finally, we demonstrate the feasibility of placenta 3-D reconstruction using multiple video sweeps. We pose this challenging problem as spatio-temporal alignment of US video. We first temporally align video sweeps to represent video content at the same temporal scale. Then, we use affine transformations to spatially align images in temporally aligned video. The results in this chapter are exciting as they show the feasibility of placenta 3-D reconstruction in a simple US sweep system.

Medical Image Analysis for Simplified Ultrasound Protocols

Alexander Darius Gleed

Balliol College

University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2023

Acknowledgements

My sincerest gratitude goes to my supervisor Professor J. Alison Noble for her support and guidance throughout my DPhil studies. I am grateful to have learned so much from your wisdom and experience as a researcher. You have been an outstanding mentor to me. Thank you for all the opportunities.

My gratitude also goes to Professor Shinjini Bhatnagar who supported many of the activities in this thesis and beyond.

I would like to thank all collaborators, particularly Professor Aris T. Papageorghiou and the MCH and ADAPT teams at THSTI who added immense value to my work. I would also like to thank my colleagues at IBME who created a friendly and welcoming atmosphere.

To my dear friends who I have met during my time at Oxford, you have all supported and inspired me in so many different ways. In particular, I would like to thank Ricardo Gonzales, Mohammad Al Sharid, Conan O'Brien, Nayani Jensen, and Elizaveta Savochkina for the many happy times we had together, be it music, food, or new experiences.

To Koundinya Desiraju, Lipsa Panda, and Divyanshu Mishra, you have treated me like family since we met and I am immensely grateful for your kindness and friendship, particularly in times of uncertainty.

Finally, I would like to thank my parents, Fabi and Steve, who have supported and encouraged me at every stage of my life. To my mum, thank you for teaching me an extraordinary work ethic. Everyday I am inspired by the challenges you overcome. To my dad, thank you for having a profound impact on my upbringing, something completely unexpected from a stranger who came later into my life.

I am grateful to the EPSRC for a doctoral training award which made my DPhil studies possible. I am also grateful to Balliol College who supported in-person attendance at an international conference. My thanks are given to Professor Jens Rittscher and Dr Alberto Gomez for their comments on this thesis.

Alexander Gleed
Oxford, June 2023

Contents

List of Figures

List of Tables

List of Abbreviations

BMI . . . . . body mass index

BP . . . . . breech presentation

CALOPUS . . . . . computer assisted low-cost point-of-care ultrasound

CNN . . . . . convolutional neural network

CP . . . . . cephalic presentation

CRF . . . . . conditional random field

CRF-RNN . . . . . conditional random field as a recurrent neural network

CVAT . . . . . computer vision annotation tool

DFD . . . . . discrete Fréchet distance

DTW . . . . . dynamic time warping

FCN . . . . . fully convolutional network

FPS . . . . . frames per second

GA . . . . . gestational age

GCN . . . . . graph convolutional network

GE . . . . . General Electric

GNN . . . . . graph neural network

GPU . . . . . graphical processing unit

IoU . . . . . intersection-over-union

LMIC . . . . . low- and middle-income countries

LSTM . . . . . long short-term memory

MLP . . . . . multi-layer perceptron

MMR . . . . . maternal mortality ratio

MRI . . . . . magnetic resonance imaging

MS COCO . . . . . Microsoft common objects in context

OSP . . . . . obstetric sweep protocol

Pascal VOC . . . . . pascal visual object classes

PASD . . . . . placenta accreta spectrum disorders

PCA . . . . . principal component analysis

PCC . . . . . Pearson correlation coefficient

PULSE . . . . . perception ultrasound by learning sonographic experience

ReLU . . . . . rectified linear unit

RNN . . . . . recurrent neural network

SNE . . . . . stochastic neighbour embedding

SVM . . . . . support vector machine

t-SNE . . . . . t-distributed stochastic neighbour embedding

TP . . . . . transverse presentation

US . . . . . ultrasound

VATIC . . . . . video annotation tool from Irvine, California

VGG . . . . . visual geometry group

VSI . . . . . volume sweep imaging

Notation

In this thesis, matrices are denoted by boldface uppercase letters $\mathbf{X}$ and vectors by boldface lowercase letters $\mathbf{x}$. The i, j-th element of matrix $\mathbf{X}$ is denoted x_{ij}. The i-th element of vector $\mathbf{x}$ is denoted by x_i.

Chapter 4

Image	$\mathbf{I} \in \mathbb{R}^{w \times h}$, w, h, width and height respectively
Video-level shape similarity metric	$\mathbf{I_2} \in \mathbb{R}^N$, N is the number of video frames
Percentage shape error	$\boldsymbol{\varepsilon} \in \mathbb{R}^N$, N is the number of video frames
Labels	$\mathbf{x} \in \mathcal{L}^N$, N is the number of classes
Energy	E
Unary energy	ψ_u
Pairwise energy	ψ_p
Label compatibility function	μ
Unweighted Gaussian kernel	k_G
Weighted Gaussian kernel	k
Negative unary energy	U_i
Partition function	Z_i
Dice similarity coefficient	$Dice$
Shape similarity metrics	I_1, I_2, I_3

Chapter 5

Graph	$\mathcal{G}$
Nodes	$\mathcal{N}$
Edges	$\mathcal{E}$
Adjacency matrix	$\mathbf{A} \in \mathbb{R}^{N \times N}$, N is the number of nodes
Identity-normalised adjacency matrix	$\hat{\mathbf{A}}$
Node feature matrix	$\mathbf{X} \in \mathbb{R}^{N \times D}$, D is the number of features
Neural network layer	$\mathbf{H}$
Learnable weight matrix	$\boldsymbol{\Theta} \in \mathbb{R}^{C \times D}$, C is the number of channels
Degree matrix	$\mathbf{D}$
Degree matrix of $\hat{\mathbf{A}}$	$\hat{\mathbf{D}}$
Node feature descriptor	$\mathbf{x}_i \in \mathbb{R}^D$, D is the number of features
Connectivity between nodes n_i and n_j	A_{ij}
Edge weight between nodes n_i and n_j	w_{ij}
Diagonal-encoded degree of node n_i	D_{ij}
Eigenvalues	λ_n
Discrete Fréchet distance	d_{dF}
Dynamic time warping	d_{tw}
Pearson correlation coefficient	p_{cc}

Chapter 6

Fixed video sequence	$\mathbf{S}$
Moving video sequence	$\mathbf{S}'$
Global spatial transform	$\mathbf{P}_{spatial}$
Global temporal transform	$\mathbf{P}_{temporal}$
Linear operation	$\mathbf{A}$
Translation	$\mathbf{b}$
Space-time point in $\mathbf{S}$	$\mathbf{x} = (x, y, t)$
Corresponding space-time point in $\mathbf{S}'$	$\mathbf{x}' = (x', y', t') = (x + u, y + v, t + w)$
Ratio of frame rate of sensor	s
Normalised ratio of frame rate of sensor	s'

Chapter 1

Introduction and Preliminaries

1.1 Clinical Motivation

Ultrasound (US) is the primary imaging modality in obstetric care. It is real-time, non-invasive, and radiation free. However, US scanning is difficult and requires a trained operator, called a sonographer, who guides an US transducer in response to real-time interpretation of video content.

In low- and middle-income countries (LMICs), there is a shortage of sonographers and a lack of cost-effective US equipment, which affects access to obstetric care [1]. The World Health Organisation estimates that 94% of maternal related deaths occur in LMICs [2], illustrated by Figure 1.1, which shows a global map of the maternal mortality ratio (MMR, number of maternal deaths per $100,000$ live births). Many of these deaths are considered preventable by timely access to evidence-based interventions [3]. For a list of current LMICs see [4].

This thesis is a step towards reducing the technical burden of US scanning, by combining simple US video sweeps with computational algorithms to assess obstetric US video content for clinical benefit. The work here is a step towards universal access to obstetric care, to reduce future maternal and fetal mortalities.

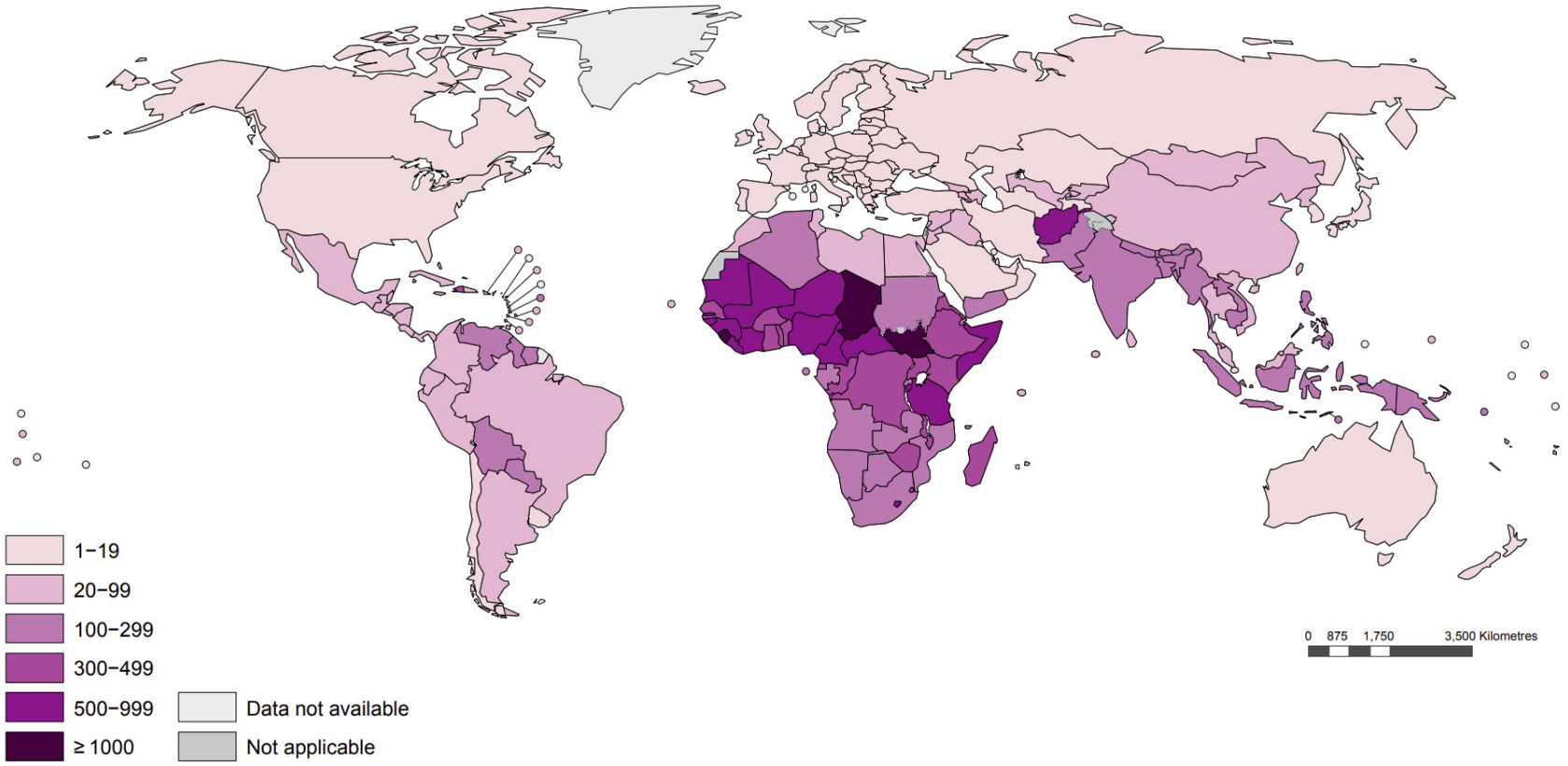

Figure 1.1: Global map showing maternal mortality ratio, (MMR, number of maternal deaths per $100,000$ live births) for 2017 [2].

1.2 CALOPUS Project

The **C**omputer **A**ssisted **LO**w-cost **P**oint-of-care **U**ltra**S**ound (CALOPUS) project is an interdisciplinary, two-site prospective study which investigates automated prediction of pregnancy risk factors using US image analysis and machine learning-based algorithms. CALOPUS is an international collaboration between the Institute of Biomedical Engineering, the Nuffield Department of Women's and Reproductive Health, University of Oxford, UK, and the Translational Health Science and Technology Institute, Faridabad, India. The scientific vision is simple:

To automate prediction of pregnancy risk factors by analysis of simple US video sweeps taken by an US-inexperienced operator.

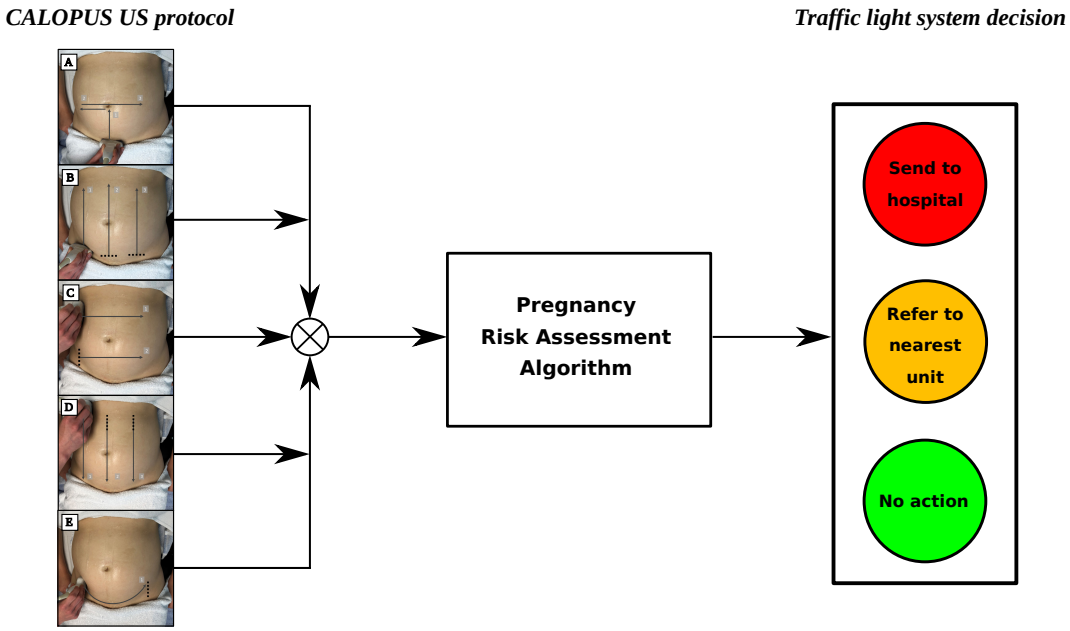

Figure 1.2: CALOPUS project vision.

An important objective of CALOPUS is to increase uptake of US-based assessments by reducing the technical burden of US scanning. We seek to design tools to increase competence of US-inexperienced operators, to enable basic pregnancy assessments using simple US video sweeps combined with machine learning-based algorithms, see Figure 1.3.

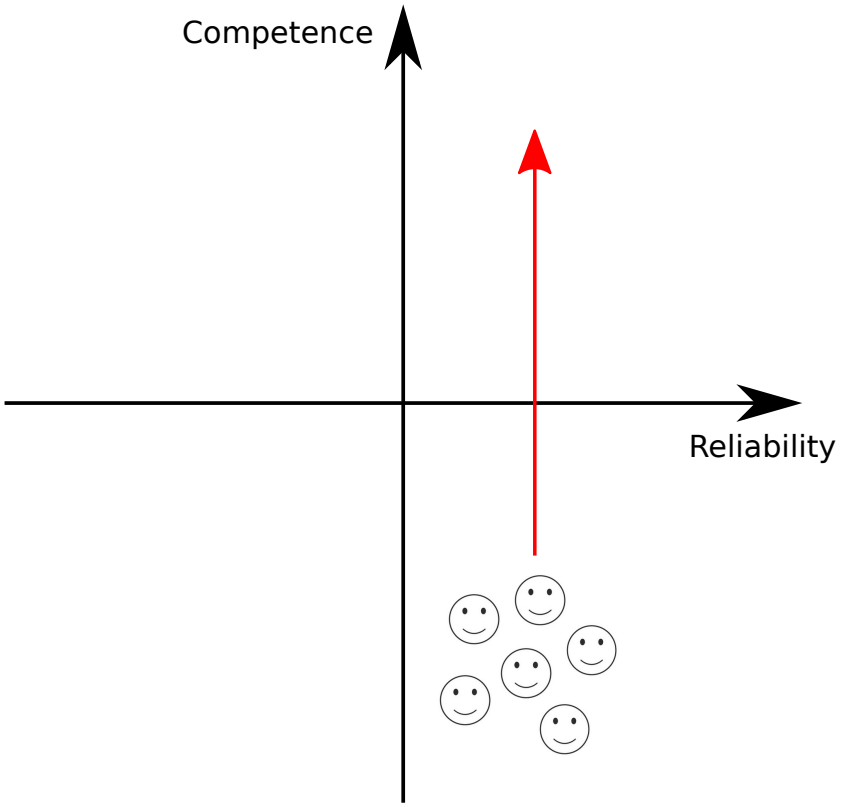

Figure 1.3: CALOPUS competency objective.

1.3 Contribution: Medical Image Analysis for3 Simple Ultrasound Video Sweeps

The contribution of this thesis is threefold, summarised graphically in Figure 1.4:

1. An image guidance algorithm to assess placenta location, including a statistical data visualisation of 2-D placenta shapes (Chapter 4).

2. A statistical characterisation and graph-based analysis of multiple US video sweeps, using fetal pose as a machine learning prior (Chapter 5).

3. A framework to spatio-temporally align multiple, untracked US video sweeps, with early results demonstrating feasibility of placenta 3-D reconstruction (Chapter 6).

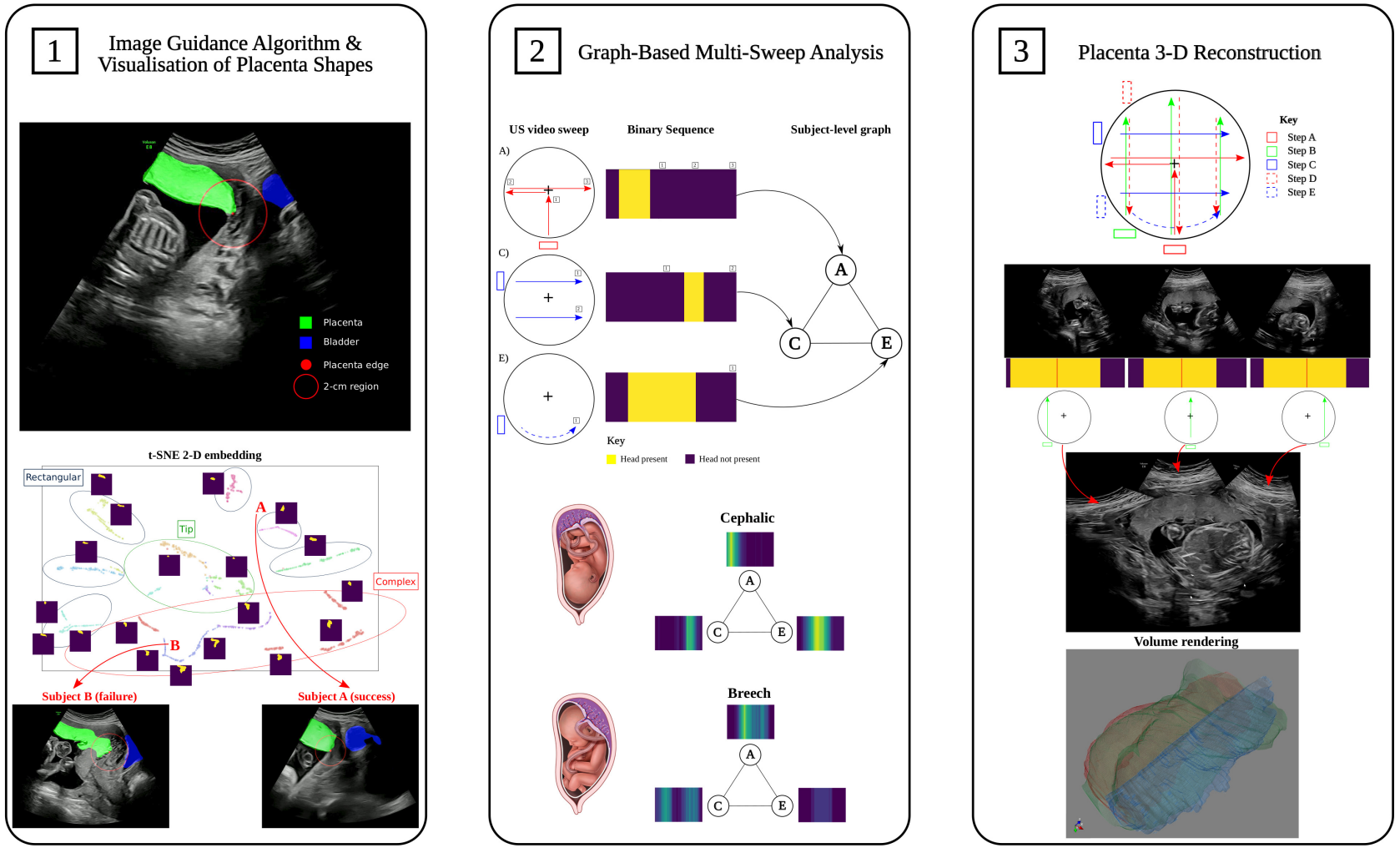

Figure 1.4: Thesis graphical summary. 1, 2, 3 refer to Chapters 4, 5, 6 respectively.

1.4 Thesis Outline

In Chapter 1, we introduce the background, context, and clinical motivation of this thesis. The main research contributions are outlined, along with a statement of originality and a list of publications. In Chapter 2, we review the relevant literature with a focus on simple US sweep protocols and their combination with computational algorithms. Specific papers relevant to a technical contribution chapter are reviewed in that chapter itself. In Chapter 3, we characterise the CALOPUS US video dataset to understand the problem space and quantify the scale of video data recorded. Sample video frames are provided.

Chapters 4, 5, and 6 are all technical contributions. Each chapter follows a set format. We begin with an **introduction** to the problem we are trying to solve, along with a review of any relevant studies. Next, we specify the **dataset** used and subsequent characterisation. We then describe our **methods**, outlining the approach and algorithms developed. Experimental outcomes are provided in **results**. We examine our results in the context of this thesis in a **discussion**. Finally, we give final comments in the **conclusions**. The specific technical contributions of Chapters 4, 5, and 6 have already been summarised in this chapter, see Figure 1.4 for a graphical summary.

In Chapter 7, we give concluding remarks and reflect on the research contributions of this thesis. Specific to each technical chapter, we discuss limitations and possible directions of future work.

1.5 Originality

I declare that I am the sole author of this thesis document unless otherwise stated, and I produced all tables and figures included herein. Critique and comments were provided by Professor J. Alison Noble. Unless otherwise specified, all image and video preprocessing and manipulation, algorithm implementation, and data analysis was my original and individual work. The US videos were acquired by clinical collaborators as part of the CALOPUS project, see Chapter 3. Medical fetus illustrations were taken from [5].

Chapter 3: Manual interpretation of US video was performed by Dr Alice Self. I computed bladder volumes using manual annotations of Dr Alice Self. Statistical sample size estimation was performed by a contracted external team at the University of California, Berkeley.

Chapter 4: Manual anatomy segmentations were produced by myself and Dr James Jackman. The PyTorch implementation of CRF-RNN was written by the authors [6]. The OpenCV library [7] was used to compute shape metrics I_1, I_2, I_3. I wrote a custom PyTorch implementation of U-Net, including a custom ResNet encoder. A SciPy implementation of t-SNE was used. External validation using THSTI site data was performed by Mr Divyanshu Mishra and Mr Varun Chandramohan.

Chapter 5: Manual frame-level anatomy annotations were produced by clinical collaborators [8]. I used the PyTorch Geometric library [9] to implement custom graph structures and a graph convolutional network. The VGG-16 model was implemented in PyTorch by Dr Richard Droste and trained by Dr Zeyu Fu. Automatic annotations were generated using THSTI site data by Mr Divyanshu Mishra.

Chapter 6: Manual frame-level placenta annotations were produced by clinical collaborators. I used the FFmpeg multimedia framework, OpenCV [7] and Mayavi libraries [10] for custom video and image manipulation and 3-D visualisation.

1.6 Publications

The publications arising from this thesis are:

First Author

A. D. Gleed, Q. Chen, J. Jackman, D. Mishra, V. Chandramohan, A. Self, S. Bhatnagar, A. T. Papageorghiou, J. A. Noble. "Automatic image guidance for assessment of placenta location in ultrasound video sweeps". In: *Ultrasound in Medicine and Biology* 49.1 (2023), pp. 106–121. – (*featured on journal front cover*)

A. D. Gleed, D. Mishra, V. Chandramohan, Z. Fu, A. Self, S. Bhatnagar, A. T. Papageorghiou, and J. A. Noble. "Towards multi-sweep ultrasound video understanding: Application in detection of breech position using statistical priors". In: *International Symposium on Biomedical Imaging*. IEEE, 2023. – (2*nd place, best oral presentation*)

Contributing Author

A. Self, Q. Chen, B. K. Desiraju, S. Dhariwal, **A. D. Gleed**, D. Mishra, R. Thiruvengadam et al. "Developing clinical artificial intelligence for obstetric ultrasound to improve access in underserved Regions: Protocol for a Computer-Assisted LOw-cost Point-of-care UltraSound (CALOPUS) Study". In: *JMIR Research Protocols* 11(9) (2022), p.e37374.

A. Self, **A. D. Gleed**, S. Bhatnagar, J. A. Noble, and A. T. Papageorghiou. "VP18. 01: Machine learning applied to the standardised six-step approach for placental localisation in basic obstetric ultrasound". In: *Ultrasound in Obstetrics and Gynecology* 58 (2021), pp.172 – 172.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we review the relevant literature to contextualise this thesis. In particular, we focus on studies that have designed simple obstetric US sweep protocols for use in LMICs and their combination with computational algorithms for clinical benefit. We include a broad review of studies that focus on the human placenta, freehand US, and deep learning in general.

2.2 US Protocols in LMICs

There are three studies that have developed simple obstetric US sweep protocols for use in LMICs [11, 12, 13]. The overall aim is to reduce the technical burden of taking an US scan by providing a grid which describes US transducer movement. The grid uses body landmarks, such as the umbilicus or pelvis, and is simple for an operator to follow. An operator is not expected to adjust imaging parameters such as depth or width. Training an operator takes a single day [14]. When paired with US outreach programmes, these studies aim to increase uptake of US-based obstetric assessments and help women access evidence-based interventions.

To date, there have been two design approaches to clinical interpretation of video from simple US sweep protocols. The six-step approach [11] pairs each video sweep (*i.e.* step) with a diagnostic criterion. This reduces complexity for an operator, as the obtained US video is manually interpreted for a single diagnostic criterion, such as placenta location. In contrast, the volume sweep imaging (VSI) protocol [12] and obstetric sweep protocol (OSP) [13] use telemedicine for clinical interpretation. US video is obtained at one site and uploaded to an online system for remote interpretation by a trained sonographer at another site. The major disadvantage of telemedicine is the associated time delay which removes the real-time advantage of US. The three protocols are reviewed in detail next.

2.2.1 Video Sweep Protocols

The six-step approach of Abuhamad et al. [11] provides a six step checklist, which pairs each simple video sweep with a diagnostic criterion. Figure 2.1 shows the protocol grid. The six step checklist assesses:

1. Fetal presentation
2. Fetal cardiac activity
3. Number of fetuses
4. Location and position of placenta
5. Amniotic fluid volume
6. Fetal biometric measurements

The pairing of diagnostic criteria with simple video sweeps reduces complexity for an US-inexperienced operator, as they understand expected findings in practice. A criticism of that study is that interpretation of US video still requires expertise that comes directly from training. No solution is offered for this, beyond training operators with suggested outreach programmes.

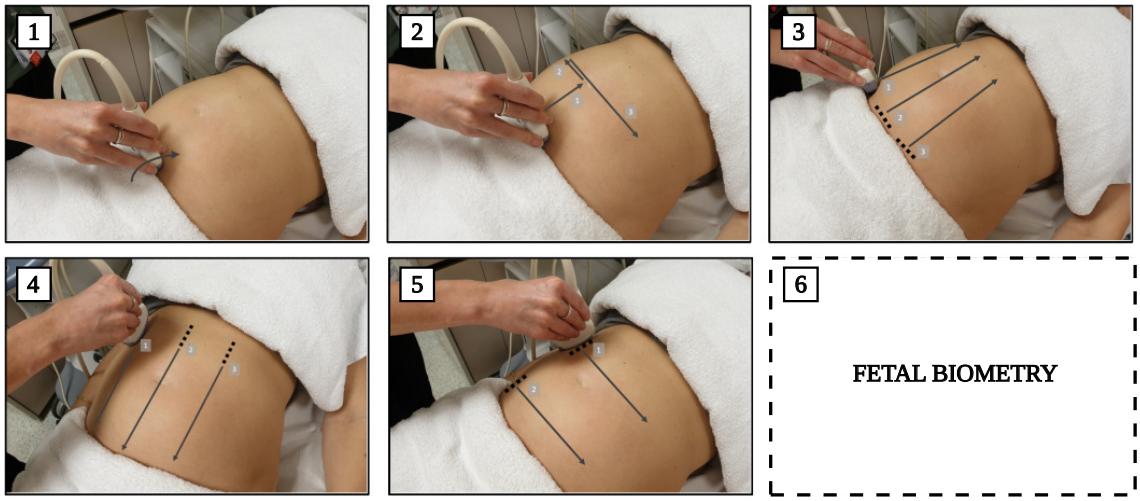

Figure 2.1: Six-step approach of Abuhamad et al. [11]. Step 1 is pure rotation of the transducer.

The six-step approach shows some redundancy between steps, compare linear sweep of step 2 and middle sweep of step 3. Step 6 is not a true simple video sweep but instead instructs a user to take fetal biometric measurements, similar to a freehand US scan taken by a trained operator. It is not clear how this step can be taken by trainee operators. In this thesis, the CALOPUS US protocol is based on a modified six-step approach. We discuss the CALOPUS protocol later in §3.1 (*CALOPUS Ultrasound Protocol*).

The VSI [12] and OSP [13] are reviewed next. These protocols use telemedicine for remote interpretation of US video. Figure 2.2 shows the relevant protocol grids. Both protocols consists of six sweeps that capture axial and sagittal US videos. For the VSI, an additional low sweep captures US video of the lower uterine segment, a region important in assessing placenta location. To aid comparison of multiple sweep protocols, we colour code Figure 2.2 to highlight similarities.

A significant disadvantage of all studies is that video interpretation is required by a trained operator, a major barrier to wider uptake. A possible solution is automated interpretation using computational algorithms that combine machine learning and medical image analysis. We review studies that have explored this approach next.

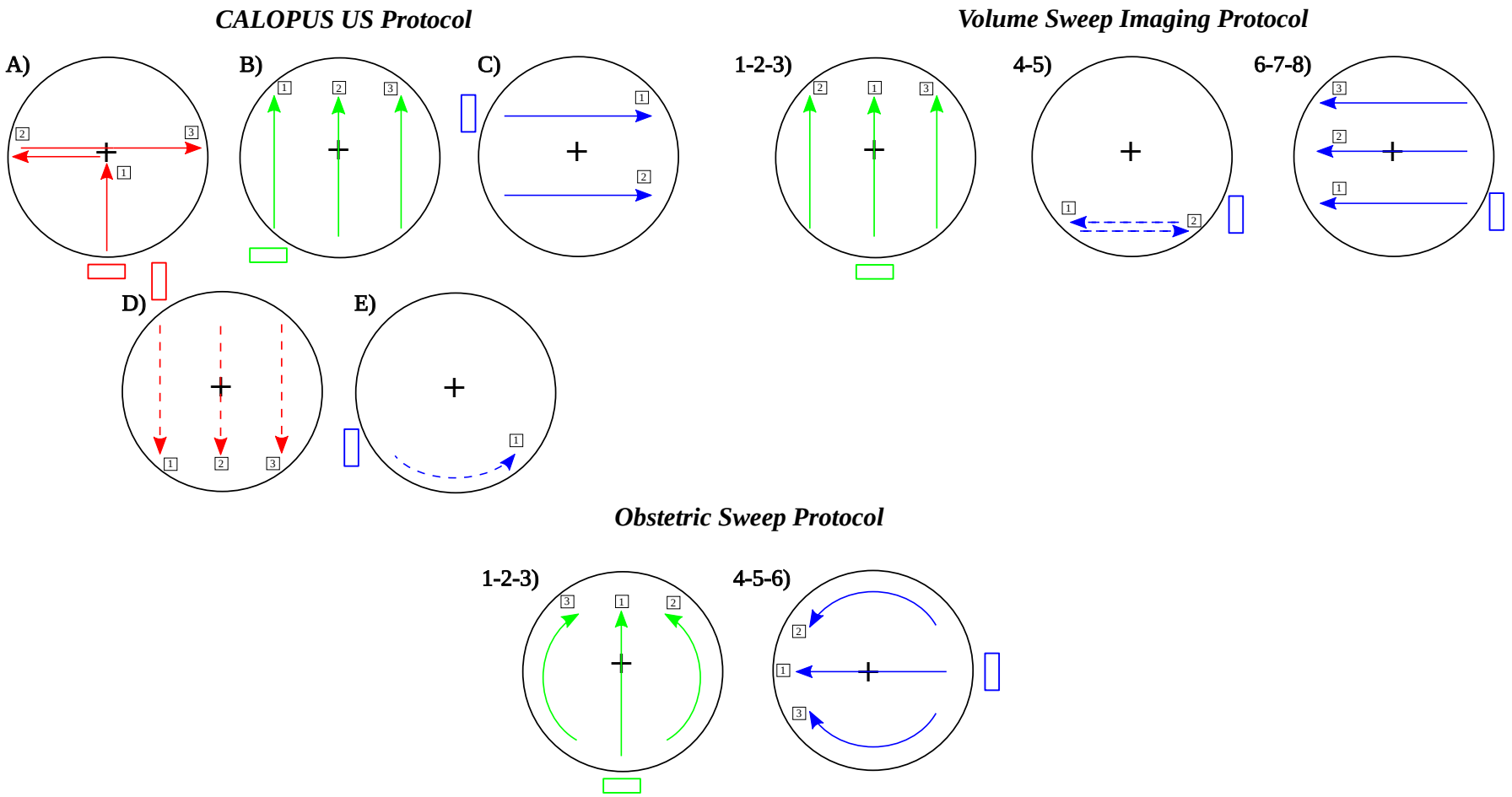

Figure 2.2: Comparison of simple sweep protocols to the CALOPUS US protocol used in this thesis. Rectangles show transducer orientation, fixed in each step, and + the umbilicus. The VSI protocol and OSP have been colour coded to allow visual comparison to the CALOPUS protocol.

2.2.2 Algorithms

We review studies that use algorithms to assess linear US video sweeps (*e.g.* straight line sweeps) or video from any of the previously discussed protocols. A model-based method by Maraci et al. [15] used a support vector machine (SVM) with Fisher vector features to automatically classify US video frames containing the fetal head or heart. Video frames from linear US video sweeps were classified, while a conditional random field (CRF) regularised the class predictions by modelling the inter-frame temporal relationship. Subsequently, the viability of the fetal heart was determined using a kernelised linear dynamical model.

A further study by Maraci, Xie, and Noble [16] used a deep learning-based AlexNet [17] with a long short-term memory (LSTM) model [18] to classify US video frames containing the fetal head, abdomen, or heart. Linear US video sweeps were used again. That study successfully modelled long-range dependencies present in US video sweeps that showed challenging anatomical variations.

Gao, Maraci, and Noble [19] used transfer learning on a deep learning-based model for the same task as [16]. Pretraining a model on ImageNet [20] was found to improve classification accuracy by 3.6%, compared to a baseline model trained on a small number of US images only, showing that convolutional kernels from models pretrained on natural images generalise to US images. That study reported that the most challenging class to classify was the fetal heart, likely due to its small, motion-induced appearance and natural class imbalance in US video.

A study by Self et al. [21] investigated automated classification of fetal presentation from linear US video sweeps. A fetal anatomical frame classifier was combined with a clinical decision-making matrix, which used the timing and order of when fetal anatomical structures were detected in video. That study assessed algorithm performance using 120 subjects and reported 94.2% classification accuracy. Classification accuracy increased to 100% when the gestational age (GA) was only > 30 weeks.

Studies by van den Heuvel et al. have explored automated interpretation of OSP video. First, [22] developed a deep learning-based fetal head classifier using a visual geometry group (VGG) model [23] to classify frames as present, not present, or partially present. The model was first trained on data from a high-cost US device and tested on data from a low-cost device. No classification accuracy was reported, aside from representative examples, making it challenging to assess algorithm performance in the broader context. Interestingly, that study showed an example of classification of a twin pregnancy, where fetal heads were recognised in two distinct locations.

Next, in [14], automated fetal head biometric measurements were made using a combined frame classifier and segmentation model. A U-Net [24] segmented the fetal head in frames classified as fully present (*e.g.* containing the full fetal head). Using fetal growth curves, a fitted ellipse was used to estimate GA. A subsequent ablation study found that frames could be downsampled up to fourteen times before classification accuracy decreased, an interesting result in the context of computationally efficient algorithms.

Finally, Heuvel, Korte, and Ginneken [25] implemented a random forest classifier [26] to predict the number of fetuses and fetal presentation. That study used 247 singleton pregnancies, with GAs between $14 - 28$ weeks, using 216 and 31 cephalic and breech presenting fetuses respectively. A classification accuracy of 99.5% and 100% was reported for cephalic and breech presentations respectively. Predicting the number of fetuses was more challenging. A further 33 twin pregnancies were assessed however only 20 were correctly predicted.

A systems approach by Arroyo et al. [5] explored automated fetal biometric measurements and classification of fetal presentation and placenta location. Video from the VSI protocol was used with a U-Net model to automatically segment the fetal head and placenta. That study and of Maraci et al. [15] are the first examples that have combined several automated diagnostic tasks into one framework. Following

[5], a study by Gomes et al. [27] automated classification of fetal presentation and prediction of GA. Interestingly, prediction of GA used US images only, instead of automated segmentation, biometric measurements and subsequent mapping to fetal growth charts.

Relevant to Chapter 4, there are three studies that automate assessment of placenta location using simple US video sweeps. Saavedra et al. [28] proposed automated detection of placenta previa using automated placenta segmentation using video from the VSI protocol. Schilpzand et al. [29] followed a similar approach, but used video from the OSP. In [29], placenta locations were classified as normal or abnormal, rather than a clinically defined placenta previa. Abnormal placenta location consisted of both low-lying placentae and placenta previa. In both studies, the location of placenta segmentations are assessed against the first video frame in a sweep. Placenta segmentations too close to the first frame are classified as abnormal or placenta previa. Finally, the systems approach of Arroyo et al. [30] classified placenta position as anterior, posterior or fundal. By placenta position, that study means the position description using anatomical directions, different to whether the placenta is described as low or normal. All studies only evaluated a single U-Net model.

2.3 Placenta-Based Algorithms

Relevant to Chapters 4 and 6, we review placenta-based and automated US placenta segmentation studies. The human placenta is a discoid, haemochorial organ that appears during pregnancy and provides the fetus with nutrients, gases, and facilitates waste exchange [31]. The placenta exists only for a short window of time in comparison to a human lifetime and its function is poorly understood.

A random walk-based algorithm by Stevenson et al. [32] semi-automated placenta segmentation using US images. An operator was required to manually initialise the algorithm by providing a closed loop that covered placenta pixels. The semi-

automated segmentations were compared to manual segmentations of a clinician and a semi-automated commercial method, virtual organ computer-aided analysis (VOCAL, GE Healthcare). The algorithm reported a mean Dice coefficient of 0.87 ± 0.13 and a shorter mean initialisation time compared to VOCAL, (175s against 301s, $p \leq 0.0001$). That study reported that algorithm failures were observed when the fetus lay close to the placenta. For such a case, the proposed solution was to modify the operator initialisation from a closed loop to a single line.

Looney et al. [33] developed two deep learning-based models for automated placenta segmentation. DeepMedic [33] used a two-stream convolutional neural network (CNN) and was trained using segmentations produced by the random walk-based algorithm of [32]. OxNNet [34] used an architecture similar to U-Net [24], leveraging skip connections. Over-segmentation was the main failure mode as it was difficult to delineate myometrium and placenta pixels.

Hu et al. [35] addressed acoustic shadow detection in automated placenta segmentation. Non-anterior placenta positions, such as posterior, suffer from additional imaging artefacts in US such as attenuation, shadowing, and occlusion. Physically, this is because an US transducer is placed on the anterior side of the maternal abdomen so non-anterior placentae are further away. That study used a U-Net with a scan-line entropy analysis [36] to interpolate segmentation masks across regions of acoustic shadows. The algorithm correctly interpolated segmentation masks across shadowed placenta regions, seen in the posterior position. Similar challenges to [34] were reported, where the algorithm failed to delineate myometrium and placenta pixels.

Zimmer et al. [37] developed an automated placenta segmentation and multi-view fusion algorithm using US video acquired at late GA. That study designed a bespoke transducer holder which integrates two or three commercial US transducers to obtain multi-view US videos. The aim was to achieve whole placenta imaging at late GA,

where the placenta is too large to entirely fit in the field-of-view of a commercial US transducer. The holder constrains the spatial relationship between transducers, which is used to register images. Voxel-wise fusion was used to create multi-view fusion images. That study presented multi-view fusion images of the entire placenta in third-trimester. It is not clear how the bespoke holder size generalises to women of different body sizes.

Qi, Collins, and Noble developed a series of algorithms to quantify US image features correlated with placenta accreta spectrum disorders (PASD), a type of pathology where the placenta grows invasively into maternal tissue [38, 39, 40]. In [38], a weakly-supervised, deep learning-based classifier labelled US image patches, which contained multiple anatomies. The objective was to preprocess a dataset of US images to localise anatomies implicated with PASD. That study reported a top-1 error rate of 0.086. Next, in [39], an image analysis algorithm was developed to automate localisation of placental lacunae. Placental lacunae are sonolucent spaces which are larger, more frequent, and irregular in PASD [41]. However, there is a lack of image-based, quantitative tools to understand placental lacunae features. Automated confidence maps from dot annotations were generated using simple linear iterative clustering superpixels. Finally, [40] developed a deep learning-based model to automate detection of the utero-placental interface. The model used global context [42] and convolutional group-wise enhancement blocks [43]. These ensured non-local dependency and learning of global semantic features respectively.

2.4 Freehand US Algorithms

Relevant to Chapters 5 and 6, we review freehand obstetric US algorithms because of their relevance to simple US video sweeps. Prevost et al. [44] used a deep learning-based model to estimate spatial transformations between successive US frames. Speckle decorrelation [45] was modelled as a series of common operations in

convolutional layers (convolution, max-pooling, non-linear activation). That study achieved a median normalised drift of 5.2%, using 800 US sweeps, but reported that it was necessary to disable speckle filtering in the US hardware, something that may not be possible in all commercial US machines.

Li et al. [46] estimated 3-D spatial transformations between US frames using a conventional 2-D US transducer with an optical tracking camera system and a recurrent neural network (RNN) [47]. The optical tracker system required a large, external rod containing markers attached to the transducer. EfficientNet [48] was combined with a LSTM-based module [18] to model long-range video dependencies. A maximum drift of 46-mm was reported. In [44, 46], both studies used linear US transducers to scan across human forearms.

Yeung et al. [49] used a deep learning-based model to predict the position of a 2-D US image in a 3-D atlas of the fetal brain. The atlas was constructed from prior work [50] using volumes taken with a 3-D US transducer. Images were sampled from the atlas, using Fibonacci sphere sampling [51], which allowed supervised training of the model, as the frame position was always known. Similarly, Yeung et al. [52] used neural radiance fields [53] to synthesise 2-D images of the fetal brain given an input plane. A deep learning-based model learned an implicit volume representation by supervised learning of greyscale pixel intensities and their associated positions. The atlas of [50] was used again. That study showed that visual quality improved by over 30%, compared to the baseline of [49]. It is unclear how the approaches in [49, 52] can be used without an accurate 3-D atlas.

Luo et al. [54] estimated 3-D spatial transformations between US frames by modelling US transducer movements as four distinct types: linear, loop, fast-slow, and sector. A position-tracked US transducer obtained video that was used to train a combined deep learning and LSTM-based model [18] using Edge maps of US frames. That study showed representative examples of 3-D fetus reconstruction.

Oktay et al. [55] used an anatomically constrained U-Net to automate segmentation of the left ventricular endocardium in 3-D US volumes. That study reported a mean Dice coefficient and Hausdorff distance of 0.912 and 7-mm respectively and demonstrated state-of-the-art results with only 15 3-D US volumes. Leclerc et al. [56] automated multi-class segmentation of the fetal heart using 2-D US images of two- and four-chamber views. That study showed an improvement in the mean Dice coefficient of [55] with a reported score of 0.939 ± 0.043.

Baumgartner et al. [57] developed SonoNet, a deep learning-based model that automated detection of fetal standard planes in 2-D US images, with the aim to guide an inexperienced operator in real-time. SonoNet automated localisation of anatomical structures using the generated image-level labels as weak supervision and reported an accuracy of 77.9%. Similarly, Komatsu et al. [58] used a deep learning-based model to automate localisation of cardiac substructures to diagnose cardiac abnormalities. A video-level abnormality score was generated using the combined set of labelled substructures and guidance was presented to an operator in barcode-like detection timeline.

2.5 Deep Learning-Based Methods

We next review deep learning-based studies as the previous section has demonstrated the frequent use of deep learning in the medical image analysis literature. A general review is presented here. In Chapters 4, 5, and 6, we review specific deep learning methods relevant to each chapter in detail there.

2.5.1 Classification

The first convolutional neural network (CNN) used in 1998 was LeNet, which automated classification of handwritten digits on postal letters, see Figure 2.3 [59]. Next, AlexNet won the ImageNet large scale visual recognition challenge 2012, with

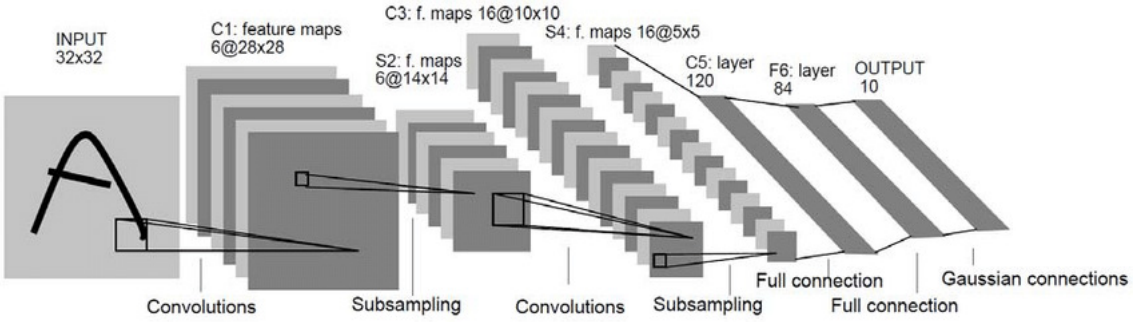

Figure 2.3: LeNet-5 architecture, adapted from [59].

a top-5 error rate of 15.3%, an error reduction of 42% from the runner-up [17]. AlexNet implemented several key architectural changes to LeNet, including parallelisation with graphical processing units (GPUs), non-linear activation functions (ReLU), and parameter dropout schemes. Simonyan and Zisserman [23] developed VGG-Net, which increased network depth to a maximum of 19 layers. This was practically achieved by stacking multiple 3×3 convolutional kernels. Szegedy et al. [60] designed GoogLeNet, leveraging an inception module which enabled a network-in-network approach to allow multi-scale convolutions. Typically, this introduces the vanishing gradient problem where backpropagated gradients tend to zero before reaching early layers of a network. To address this, GoogLeNet introduced auxiliary loss functions. ResNet was introduced by He et al. [61], achieving a maximum network depth of 152 layers. Residual learning allowed convolutional layers to learn a residual which aided model convergence and performance.

To date, other architectures include ResNeXt [62], DenseNet [63], Network in Network [64], MobileNet [65], and SqueezeNet [66]. For class imbalanced scenarios, the corresponding FixResNeXt and FixSqueezeNet have been proposed [67]. Recently, convolution-based models are being combined with attention-based vision transformers [68, 69] to create hybrid models, such as CoAtNet [70]. These models require increasingly large amounts of training data to converge.

2.5.2 Segmentation

Segmentation refers to dense classification of pixels in an image. Long, Shelhamer, and Darrell [71] modified standard classification architectures by replacing fully connected layers with fully convolutional layers. The fully convolutional network (FCN) takes an arbitrary sized input and outputs pixel predictions. Skip connections also concatenated features at different scales to refine pixel predictions. Ronneberger, Fischer, and Brox [24] developed U-Net for automated biomedical image segmentation. U-Net used a encoder-decoder architecture, leveraging skip connections to fuse features between corresponding layers. Many variants of U-Net now exist because of its superior performance and flexibility in automated biomedical image segmentation [72]. Byra et al. [73] developed a selective kernel U-Net for breast mass lesion segmentation modulating kernel dimensions using an attention mechanism [73]. It is unclear how the attention mechanism generalises across different domains. Kohl et al. [74] developed a probabilistic U-Net using a conditional variational autoencoder [75] that segmented ambiguous images, such as US prostate images. Zeng et al. [76] posed automated segmentation as a self-supervised task, where a teacher and a student network share knowledge. Similarly, Huang, Noble, and Namburete [77] proposed omni-supervised learning, which combined a small manually annotated dataset with a teacher and a student network to increase effective training data.

He et al. [78] developed Mask R-CNN, a network that uses fast region proposals to automate segmentation of natural images. A region-of-interest align layer preserves spatial locations of features used to generate segmentations. Maninis et al. [79] semi-automated segmentation of objects in natural images using manually input extreme points. Gaussian functions were used to determine object boundaries. It is unclear if the method generalises to US or medical images. Ke et al. [80] used multi-task learning to iteratively refine a set of coarse manual segmentation, using recursive approximation tasks. Interestingly, that study used US images.

2.5.3 Edge Detection

Edge detection refers to classification of pixels corresponding to boundaries of objects. Perceptually, edges are regions of high intensity change. In the context of US, edge features have been used for freehand 3-D US reconstruction [54] and detection of the fetal head [81]. Classical edge detection algorithms include the Canny edge detector and Sobel operator [82, 83]. Xie and Tu [84] developed HED-Net using whole image inputs, where multi-scale edge features are fused to produce final edge predictions. For multi-class edge detection, Yu et al. [85] developed CASE-Net allowing predicted edges to belong to multiple classes simultaneously, which improved model convergence and performance. As discussed previously, Qi, Collins, and Noble [40] developed UPI-Net, which explored edge detection of the utero-placental interface in placental US images.

2.5.4 Dilated Convolution

Studies have investigated modifications to standard convolutional kernels, with the aim of reducing the number of kernel parameters and achieve fast and efficient feature representations. For US video, dilated convolutions have been used to effectively model video dynamics [16] and US image representations using human gaze [86]. Yu and Koltun [87] developed dilated convolution allowing exponential expansion of a kernel's receptive field while only linearly increasing the number of parameters. With a dilated kernel, multi-scale features are aggregated in the kernel itself, instead of using architectural modifications such as fused side outputs [40] or feature pyramid networks [88].

2.5.5 Graph CNNs

Recently, graph convolutional networks (GCNs) have enabled geometric deep learning on non-Euclidean structured data [89]. By non-Euclidean, we mean data which is

not arranged in a square $n \times n$ grid. GCNs operate on relational data structures (*e.g.* graphs), which most simply model relationships between objects. This flexibility has seen GCNs used across a variety of domains including: new antibiotic discovery [90], social network modelling [91], and within computer vision and medical image analysis [92]. Chen et al. [93] demonstrated automated video classification using sequence modelling with a GCN to classify whole videos in the YouTube-8M dataset. Videos were analysed at multiple scales, beginning at the frame-level, then iteratively to shot, event, and finally video. Parisot et al. [94] performed automated node classification of cohort graphs in autism spectrum disorder and Alzheimer's disease. That study used graphs to encode imaging-based data as node feature descriptors and phenotypic information as edge weights, demonstrating how complementary data can be arranged in cohort graphs. Ktena et al. [95] used similarity metric learning on subject-level graphs using siamese GCNs. That study used graphs to encode fMRI time series data as node feature descriptors. Lu et al. [96] used multi-scale, spatio-temporal GCNs for cardiac image analysis, using graphs to model left ventricle motion patterns. Lu et al. [96] used multi-scale, spatio-temporal GCNs for cardiac image analysis, using graphs to model left ventricle motion patterns. To date, graphs have been used to model multi-modal representations, such as for US transducer guidance [97], combining US video, transducer motion, and human gaze.

2.6 Conclusion

The literature relevant to this thesis has been reviewed. For algorithms designed to analyse video content from simple obstetric US video sweeps, there are multiple studies that have explored automation of relevant clinical tasks including: classification of fetal presentation and number of fetuses, assessment of placenta location, and estimation of fetal GA. A common approach is to train machine learning-based algorithms to understand video content from a single video sweep type and thus au-

tomate a diagnostic task. This simplifies the complexity of the video analysis, but means that complementary video data from other sweeps is unused. Further, as all studies have prioritised full automation, it is unclear what would happen in a catastrophic failure when a human is removed from the clinical decision-making process. An example would be algorithm misclassification of fetal presentation. There is an important need for automated solutions to feedback when a clinical task cannot be achieved with certainty, something which none of the current studies do. There are currently no studies that explore providing automated assistance to an operator to assess a diagnostic criterion, in the context of simple US obstetric sweep protocols.

For deep learning, there are further studies not reviewed here, such as transformers [98] and their vision equivalent [99]. These have been neglected because of their reliance of large datasets, in comparison to CNNs, and challenges with model convergence.

In this thesis, Chapter 4 develops an algorithm to assist an operator in the clinical decision-making process to assess placenta location. As shown in this literature review, none of the current algorithms [28, 29, 30] provide assistance to an operator to assess placenta location, instead preferring full automation. Chapters 5 and 6 address the challenge of multiple video sweep analysis, a step towards understanding US video sweep protocols holistically. As shown in this literature review, there is an opportunity to develop algorithms that combine information from complementary US video sweeps, especially when knowledge of the acquisition protocol is used.

Chapter 3

CALOPUS Ultrasound Video Dataset

This chapter describes real-world, obstetric US video data obtained as part of the CALOPUS project, an interdisciplinary, multi-site prospective cohort study which investigates automated prediction of pregnancy risk factors using US image analysis and machine learning-based algorithms. A simple US video sweep protocol has obtained real-world obstetric video data. This chapter is based on the following published work:

A. Self, Q. Chen, B. K. Desiraju, S. Dhariwal, **A. D. Gleed**, D. Mishra, R. Thiruvengadam et al. "Developing clinical artificial intelligence for obstetric ultrasound to improve access in underserved Regions: Protocol for a Computer-Assisted LOw-cost Point-of-care UltraSound (CALOPUS) Study". In: *JMIR Research Protocols* 11(9) (2022), p.e37374.

A. Self, **A. D. Gleed**, S. Bhatnagar, J. A. Noble, and A. T. Papageorghiou. "VP18. 01: Machine learning applied to the standardised six-step approach for placental localisation in basic obstetric ultrasound". In: *Ultrasound in Obstetrics and Gynecology* 58 (2021), pp.172 – 172.

3.1 CALOPUS Ultrasound Protocol

The CALOPUS US protocol is a modified version of the Abuhamad et al. [11] six-step approach for a focused basic obstetric US assessment. The protocol consists of steps A-E, where each step instructs an operator to take simple, 2-D US video sweeps by following a grid describing the transducer trajectory. The protocol is shown graphically in Figure 3.1, where rectangles show the transducer orientation and + the maternal umbilicus. Each step is paired with a clinical objective. The sweep paths use body landmarks for guidance, such as the umbilicus or pelvis, which makes the protocol simple to take for an US-inexperienced operator. Studies have shown that the protocol can be taught to an operator in a single day [14].

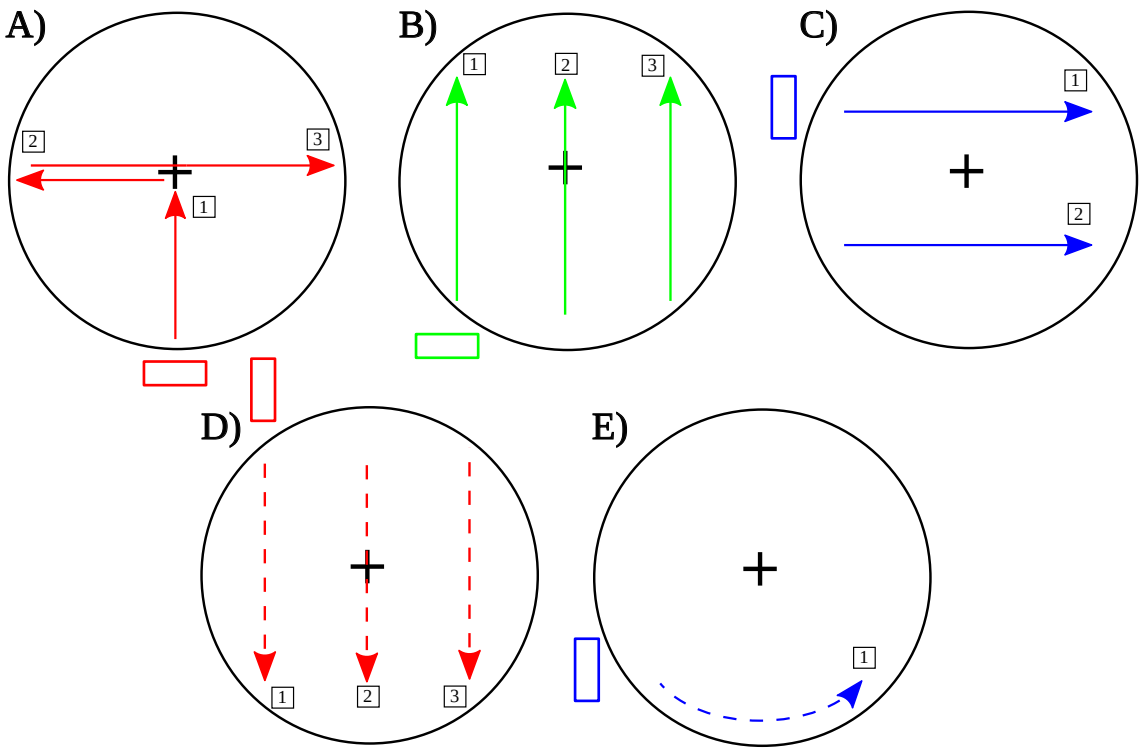

Figure 3.1: CALOPUS US protocol visualised on separate maternal abdomens. Rectangles show the transducer orientation, fixed in each step, and + the umbilicus.

3.1.1 Step A

The clinical objective of step A is to **assess the fetal presentation, lie, and cardiac activity**. Step A is an axial, T-shaped sweep, beginning above the pubic symphysis, scanning upwards towards the umbilicus, and laterally to the maternal right and then left. This sweep takes approximately 30s to complete.

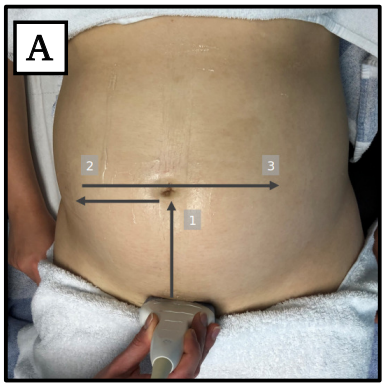

Figure 3.2: Step A shown on a real-world maternal abdomen.

3.1.2 Step B

The clinical objective of step B is to **assess the number of fetuses**. Step B consists of three axial video sweeps, beginning at the base of the maternal abdomen and scanning upwards over the maternal right, middle, and then left. Each sweep takes approximately 10s to complete.

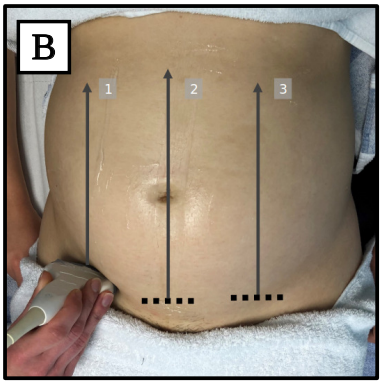

Figure 3.3: Step B shown on a real-world maternal abdomen.

3.1.3 Step C

The clinical objective of step C is to **assess the number of fetuses and quantity of amniotic fluid**. Step C consists of two sagittal video sweeps, beginning at the upper maternal right quadrant and scanning over to the maternal left. The sweep is repeated in the lower quadrant, below the umbilicus. Each sweep takes approximately 5s to complete.

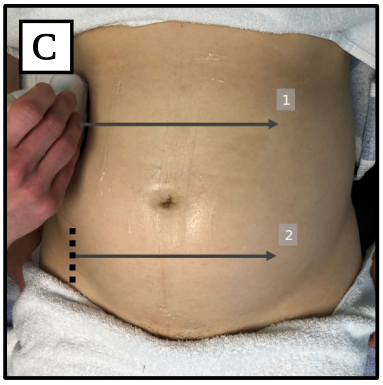

Figure 3.4: Step C shown on a real-world maternal abdomen.

3.1.4 Step D

The clinical objective of step D is to **assess the location of the placenta**. Step D consists of three sagittal video sweeps beginning at the upper maternal right and sweeping down to the lower quadrant. The sweep is repeated at the middle and maternal left. Each sweep takes approximately 10s to complete.

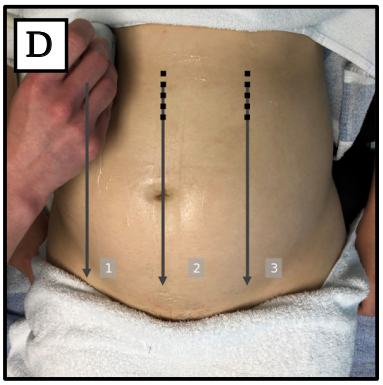

Figure 3.5: Step D shown on a real-world maternal abdomen.

3.1.5 Step E

Step E was introduced in the CALOPUS US protocol to improve visualisation of the maternal cervix, an important anatomical landmark needed to assess placenta location. Additionally, the lower uterine segment is scanned to aid identification of possible low-lying placentas. Thus, the clinical objective of step E is to **assess the location of the placenta and improve visualisation of the maternal cervix**. This sweep takes approximately 20s to complete.

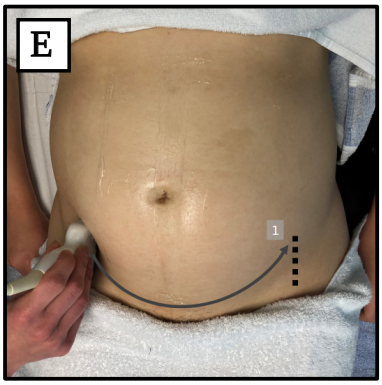

Figure 3.6: Step E shown on a real-world maternal abdomen.

3.1.6 Analysis of Bladder Fullness in Step E

In specialist placenta clinics, it is common practice to scan women with full bladders when assessing placenta location. The maternal bladder is anatomically adjacent to the cervix, thus expansion and contraction of the bladder during filling and voiding affects the sonographic appearance of the maternal cervix. It is considered easier to visualise the maternal cervix with a full bladder. We conducted a study to quantify the effect of bladder fullness on visualisation of the maternal cervix. Since step D and E were designed to assess placenta location and visualise the maternal cervix, we investigated clinical decision-making using step D only and steps D and E combined. A trained sonographer interpreted US video to make a clinical decision using the video content alone. Interpretation was performed offline to acquisition.

The hypotheses of this study were:

1. *Does the inclusion of step E in the CALOPUS US protocol help visualise the maternal cervix more often than step D only?*
2. *Is a filled maternal bladder more significant for visualisation of the maternal cervix than the addition of step E?*
3. *Is the addition of step E in the CALOPUS US protocol clinically significant?*

A total of 212 subjects were required using statistical sample size estimates with a type-1 error rate of 0.05 and 90% power. This was calculated by an external statistics team. Step D and E US videos were interpreted offline by a sonographer to reach a clinical decision, assessing placenta location and presence of the maternal cervix, at first using only step D videos followed by using steps D and E.

Table 3.1 shows the results of this study. A bladder was considered filled if the volume was greater than 200ml using a cuboid volumetric calculation and manual annotations of bladder length and width. Of the 212 recruited subjects, 65 and 147 had a filled and a poorly filled bladder respectively.

Table 3.1: Results of US video interpretation for assessment of placenta location and visualisation of maternal cervix.

	Bladder filled (n=65) Subjects, (%)	Bladder poorly filled (n=147) Subjects, (%)
Visualisation of maternal cervix		
Step D only		
Seen (n=29, 13.7%)	18 (28)	11 (7)
Unseen (n=183, 86.3%)	47 (72)	136 (93)
Steps D and E		
Seen (n=120, 56.6%)	59 (91)	61 (41)
Unseen (n=92, 43.3%)	6 (9)	86 (59)
Assessment of placenta location		
Step D only		
Correct (n=165, 77.8%)	53 (82)	112 (76)
Incorrect (n=47, 22.2%)	12 (18)	35 (24)
Steps D and E		
Correct (n=189, 89.2%)	61 (94)	128 (87)
Incorrect (n=23, 10.8%)	4 (6)	19 (13)

The inclusion of step E increased visualisation of the maternal cervix by fourfold (14% against 57%, $p \leq 0.0001$), regardless of bladder filling. Bladder filling increased visualisation of the maternal cervix, both without step E (7% to 28%) and with it (41% to 91%). Thus, for a filled bladder, adding step E increased visualisation of the maternal cervix (28% against 91%, $p \leq 0.0001$). Similarly, adding step E increased correct assessment of placenta location as low-lying (78% against 89%, $p \leq 0.0001$).

In conclusion, both bladder filling and inclusion of step E were found to independently improve placenta location assessment and visualisation of the maternal cervix. The results here are statistically significant. In the context of this thesis, Chapter 4 develops an US image analysis algorithm for assessment of placenta location using video data from step E. The results here show the clinical value of including bladder filling and step E in the CALOPUS US protocol.

3.2 Study Recruitment

Video data was recorded at two sites globally, the John Radcliffe Hospital, Oxford, UK, and the Civil Hospital, Gurugram, India. At each site, trained sonographers obtained video data with a GE Voluson E8 US machine with C2-9 or C1-5 curvilinear probes (GE Healthcare). All video data was anonymised prior to storage on secure servers at the Institute of Biomedical Engineering, Oxford, UK or the Translational Health Science and Technology Institute, Faridabad, India. Literate women provided written consent for participation in the study. Illiterate women were explained details of the study in the presence of a literate, impartial witness. In this case, verbal consent and a thumb print were taken, in addition to a signature of the witness. Table 3.2 shows the number of recruited subjects by gestational age. In addition to the total $5,140$ subjects in Table 3.2, 431 subjects were scanned with the six-step approach [11] in early stages of the project. As this was a prospective study, the CALOPUS protocol was developed concurrently with ongoing recruitment.

Table 3.2: Number of recruited subjects in CALOPUS project by gestational age (weeks$^{+\text{days}}$) and protocol.

Gestational age (weeks$^{+\text{days}}$)	Six-Step Approach [11] Subjects, (%) (n=431)	CALOPUS US Protocol [8] Subjects, (%) (n=5,140)
< 14	4 (0.9)	909 (17.7)
14^{+0} to 17^{+6}	1 (0.2)	435 (8.5)
18^{+0} to 24^{+6}	265 (61.5)	1,485 (28.9)
25^{+0} to 29^{+6}	7 (1.6)	354 (6.9)
30^{+0} to 34^{+6}	69 (16)	1,022 (19.9)
$\geq 35^{+0}$	85 (19.8)	935 (18.1)

3.3 Manual Annotations

Video frames were manually annotated at both sites for fetal and maternal anatomical structures, see list below. Annotations were generated by a single sonographer and four radiologists at UK and India sites respectively. A combination of bounding boxes and frame-level labels were annotated to reduce annotation labour. For steps A and E, bounding boxes were annotated, see Figure 3.7, and for the remaining steps, frame-level labels. Table 3.3 shows the number of unique videos and frames annotated across both sites. The following anatomical structures were annotated:

Fetal anatomical structures: *head, cerebellum, spine, abdomen, heart, stomach bubble, pelvis, and femur.*

Maternal anatomical structures: *cervix, vagina wall, bladder, placenta, and amniotic fluid.*

Table 3.3: Number of unique video and frame-level annotations.

CALOPUS step	Videos (n=1,057) Subjects, (%)	Frames (n=586,253) Subjects, (%)
Step A	446 (42.2)	254,095 (43.3)
Step B	161 (15.2)	114,958 (19.6)
Step C	163 (15.4)	67,633 (11.5)
Step D	164 (15.5)	103,269 (17.6)
Step E	123 (11.6)	46,298 (7.9)

Manual annotations were generated using a VATIC backend [100] and a web page-based tool hosted on a dedicated desktop machine. During the project, the annotation tool changed to CVAT [101] to increase functionality to include polygon- and point-based annotations.

To assess variation of manual annotations, Self et al. [102] conducted a study to quantify intra- and inter-annotator agreement using annotations from a single independent annotator at each site. Manual annotations were generated on $18,717$ frames from 20 unique videos by an independent sonographer and radiologist at UK and India sites respectively. All anatomical structures previously listed were annotated with rectangular bounding boxes, except the cervix and vagina wall, which were annotated later in the project. Manual annotations were repeated on on $18,717$ frames two weeks later to assess intra-annotator agreement. Intra- and inter-annotator agreement was quantified using four metrics given below:

1. **Exact match:** a frame was an exact match if the number and type of anatomical structures was identical between two different annotations.

2. **Partial match:** a frame was a partial match if the number and type of matched anatomical structures was $> 50\%$ between two different annotations.

3. **Match per structure:** for a single anatomical structure, the ratio of the total number of exact matches (in frames) and the total number of frames annotated with that anatomical structure.

4. **Intersection-over-union (IoU) $> 50\%$:** for a single anatomical structure, a frame was an IoU match if the IoU was $> 50\%$ between two different bounding box annotations. The IoU is a score between $0 - 1$ which represents the area similarity between two sets, A and B:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

Table 3.4: Intra- and inter-annotator agreement for two independent annotators.

	Intra-annotator agreement		Inter-annotator agreement
	Annotator 1	Annotator 2	
Exact match (%)	84	91.1	71.4
Partial match (%)	98.1	98.3	94.5
Match per structure (%)	91.8	95.7	83.1
IoU >50% (%)	94.8	97.1	93.4

Table 3.4 shows the intra- and inter-annotator agreement for two independent annotators on $18,717$ frames. Annotator 2 had higher intra-annotator agreement scores consistently compared to Annotator 1, suggesting more reproducible annotations. For both annotators, partial matches were higher than exact matches and intra-annotator agreement scores were higher than inter-annotator agreement scores.

In conclusion, Table 3.4 quantifies the variation in manual annotations generated in the CALOPUS project by assessing the intra- and inter-annotator agreement using four metrics. The results can be used as a baseline for future studies assessing the variability of manual annotations. Future studies should also investigate the variability of manual annotations in the context of clinical significance or the training of machine learning-based algorithms.

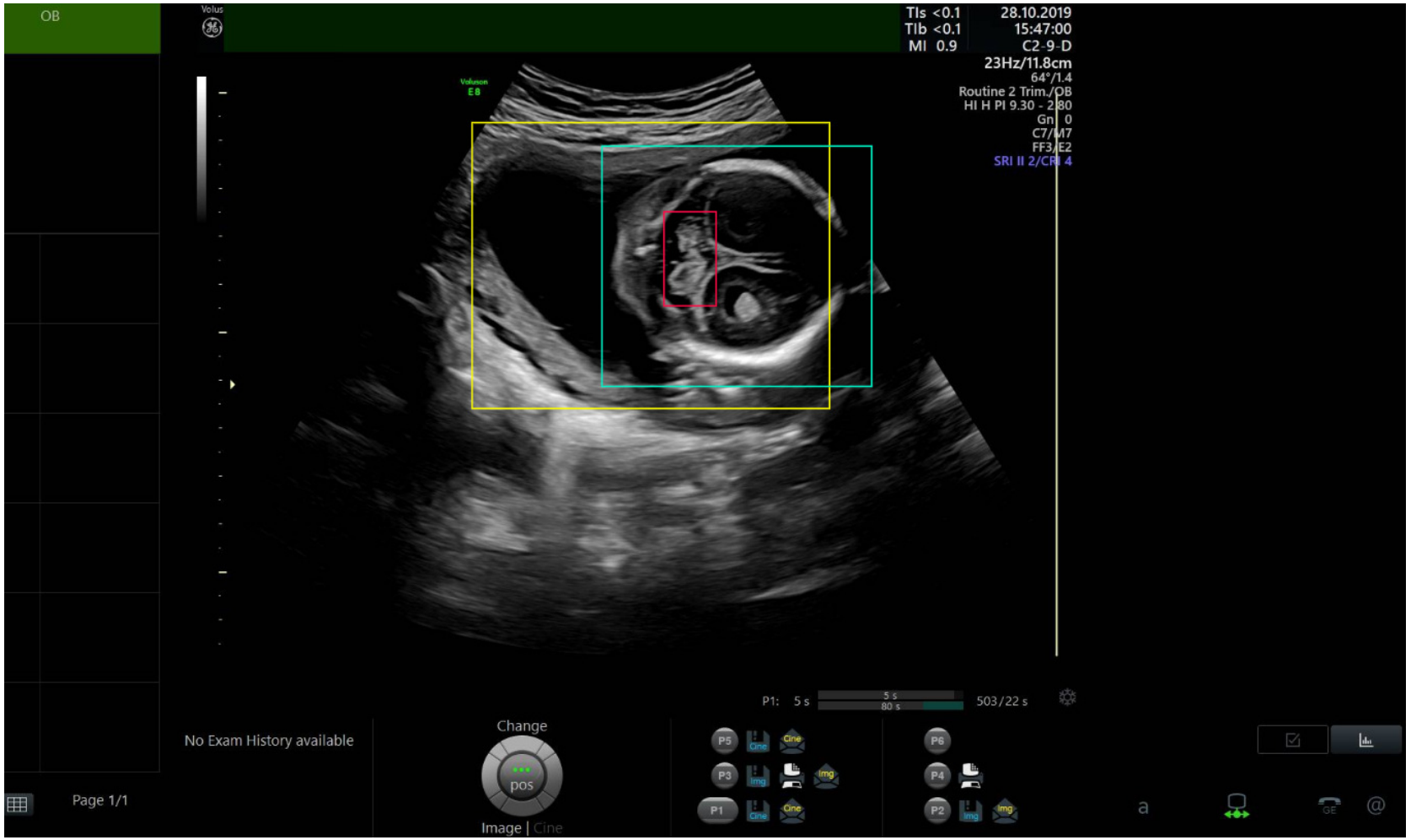

Figure 3.7: Example bounding box annotations. Yellow, cyan, and red show amniotic fluid, fetal head, and cerebellum respectively.

3.4 Sample US Video Frames

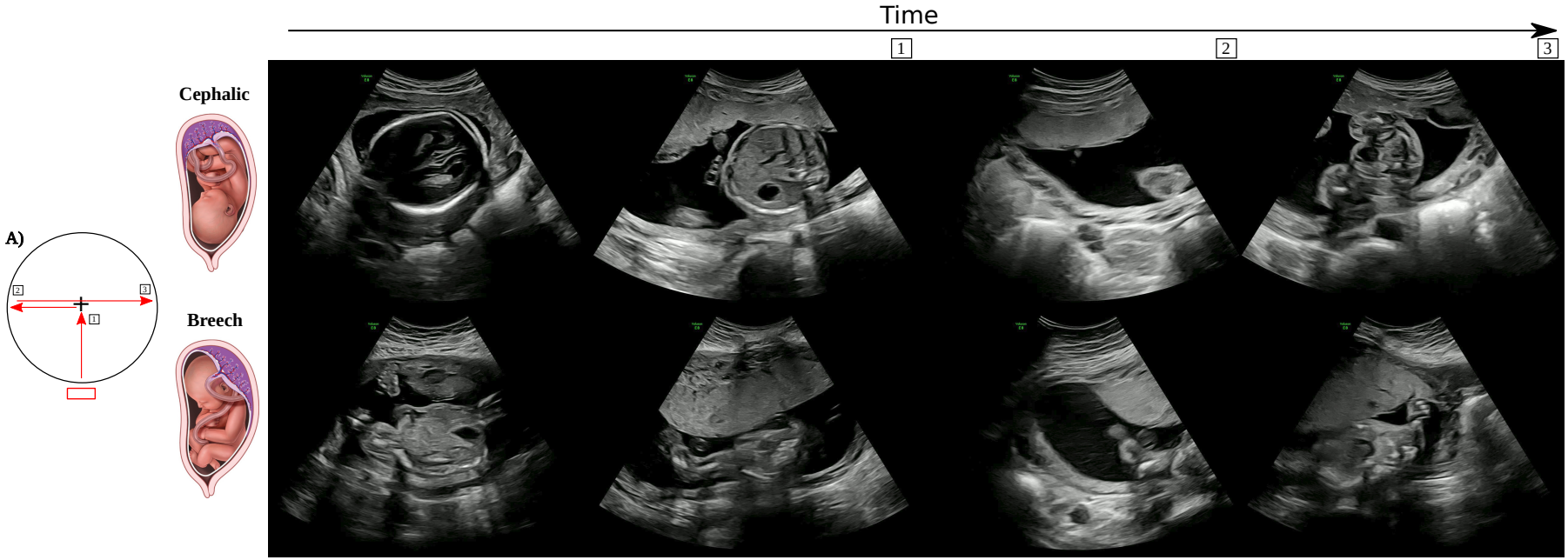

Figure 3.8: Sample step A video frames. Note the high variation in observed anatomies.

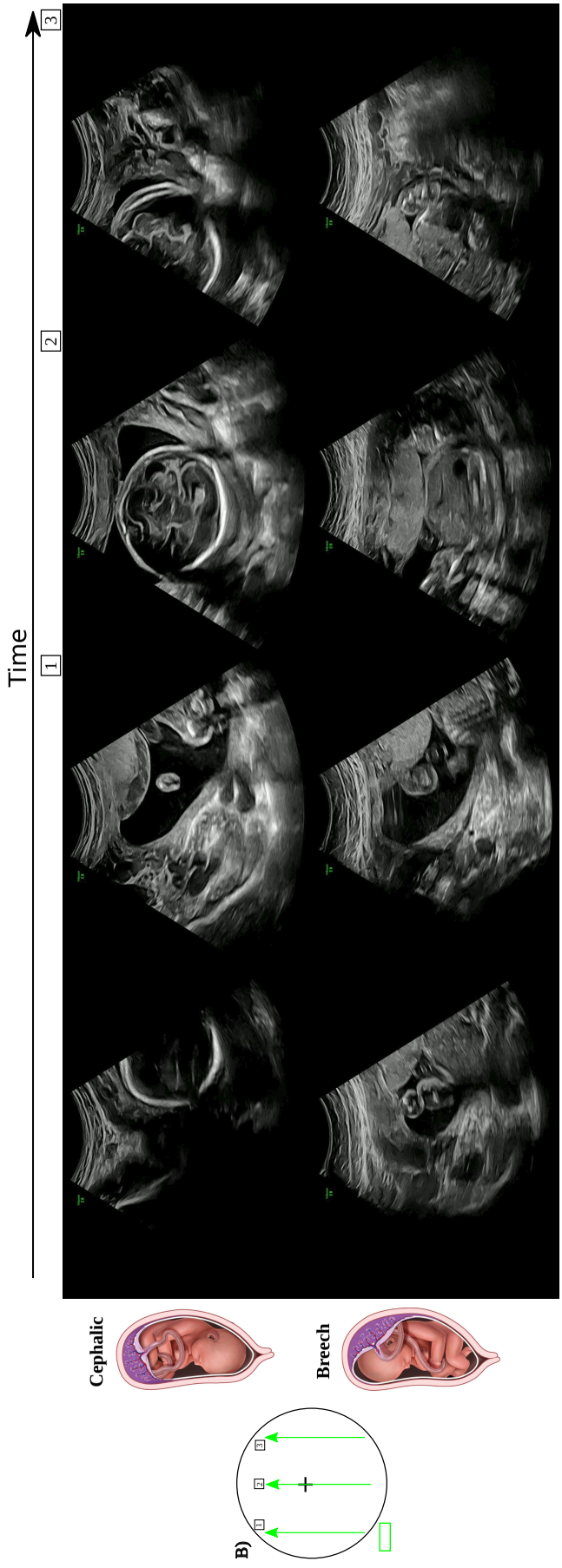

Figure 3.9: Sample step B video frames. Note the high variation in observed anatomies.

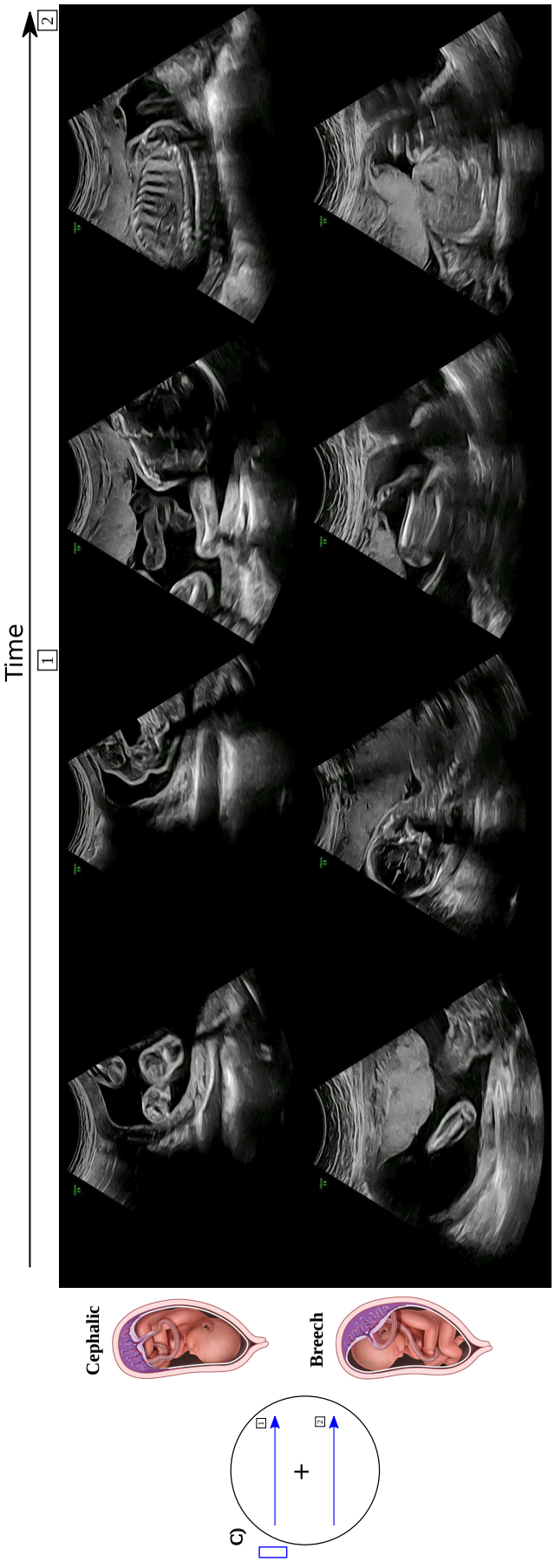

Figure 3.10: Sample step C video frames. Note the high variation in observed anatomies.

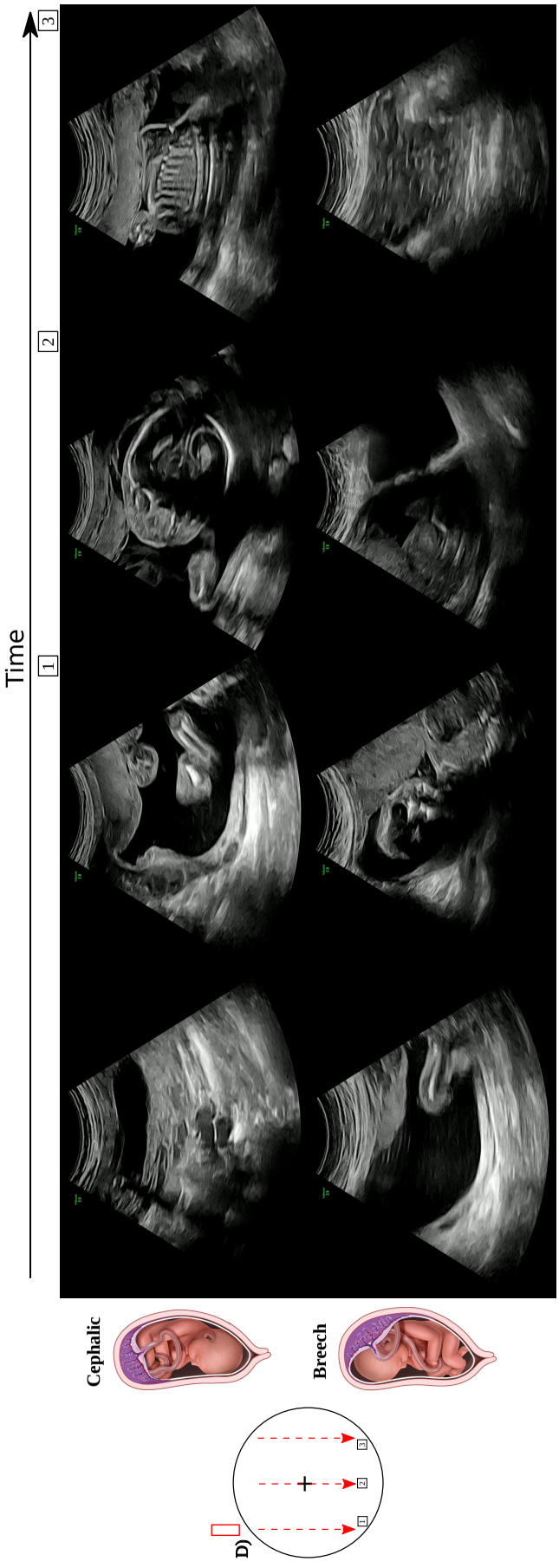

Figure 3.11: Sample step D video frames. Note the high variation in observed anatomies.

Figure 3.12: Sample step E video frames. Note the high variation in observed anatomies.

3.5 Multi-Sweep Analysis

In this thesis, we introduce a new approach to analysing simple sweep protocols. Consider all steps visualised on a common maternal abdomen space, see Figure 3.13. The video sweeps combine to form a complementary video dataset of a single subject, a type of sparse 3-D dataset. This approach is a step away from clinically motivated single step analysis, where each video is interpreted to assess a specific diagnostic criterion. The research challenge is how to build computational algorithms that can understand and use this type of sparse 3-D data for clinical and research purposes. In CALOPUS, we do not use position-tracked US transducers or set a pre-defined scan speed, which will affect how 2-D US video is transformed to 3-D.

In this thesis, Chapter 4 considers *single sweep analysis*, while Chapters 5 and 6 consider *multi-sweep analysis*. In Chapter 4, we develop an US image analysis algorithm to assess placenta location using a single input video sweep. In Chapter 5, we perform a graph-based analysis of three video sweeps, where nodes model unique video sweeps. Finally, in Chapter 6, we demonstrate feasibility of placenta 3-D reconstruction using three video sweeps which scan over parts of placenta volume necessary for 3-D reconstruction.

3.6 Conclusion

The CALOPUS project has acquired a dataset of real-world, obstetric US video data recorded at two sites globally. Tables 3.2 and 3.3 summarise the number of subjects recruited and video annotations generated to date. We have not separated the dataset by site, since the same protocol and US machine type were used. By pooling the data, we increase subject and operator scan diversity. The richness of the video data can be leveraged to understand simple US video sweep content and develop image analysis and machine learning-based algorithms.

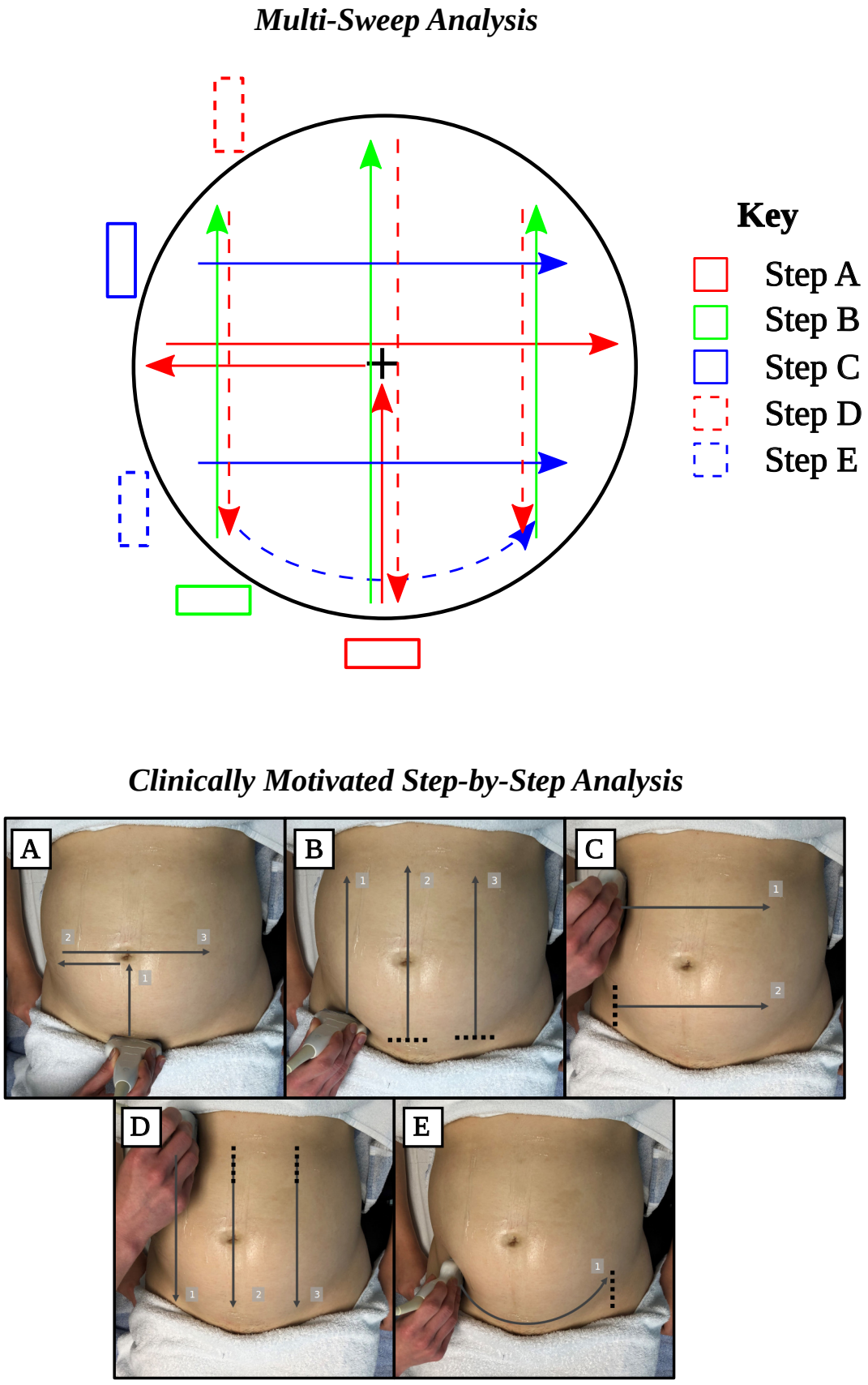

Figure 3.13: Multi-sweep analysis approach.

Chapter 4

Automatic Image Guidance for Assessment of Placenta Location

This chapter describes an US image analysis algorithm that automatically produces an assistive video overlay from a simple U-shaped video sweep. The overlay assists the interpretation of US video to assess placenta location. Assessment of placenta location is a non-trivial task for an operator. We describe the design and evaluation of a deep learning-based automatic segmentation model and a t-SNE analysis of 2-D placenta shapes. The t-SNE analysis reveals the spectrum of placenta shapes in the problem space. A probabilistic graphical model, in the form of a conditional random field as a recurrent neural network (CRF-RNN), is used to improve segmentations for the highly variable placenta shape. From the automatic segmentations, image guidance is created at the frame-level, translating the clinical criteria into assistive visual information. This chapter is based on the following published work:

A. D. Gleed, Q. Chen, J. Jackman, D. Mishra, V. Chandramohan, A. Self, S. Bhatnagar, A. T. Papageorghiou, J. A. Noble. "Automatic image guidance for assessment of placenta location in ultrasound video sweeps". In: *Ultrasound in Medicine and Biology* 49.1 (2023), pp. 106–121.

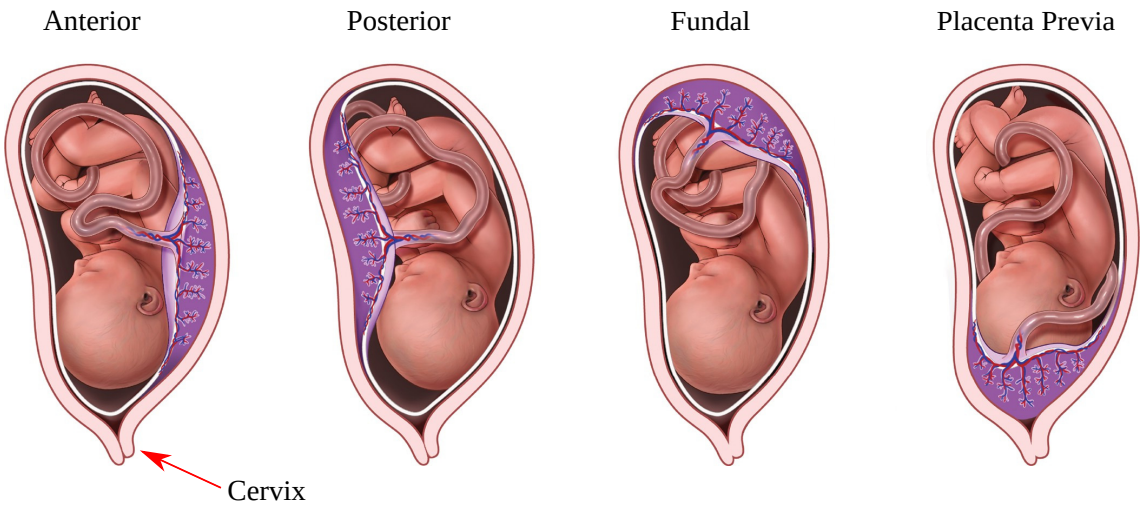

Figure 4.1: Common placenta locations defined with respect to standard anatomical directions. Adapted from [5].

4.1 Introduction

It is important to assess placenta location during pregnancy to avoid obstetric complications and inform the safest mode of delivery [103]. The human placenta is a transient, extracorporeal organ which provides the developing fetus with nutrients, in addition to facilitating gas and waste exchanges [31]. The placenta implants on the uterine wall to interface with the maternal circulation. A normal placenta location is categorised with respect to standard anatomical directions: anterior, posterior, and fundal. An abnormal placenta location consists of the placenta obstructing or located close to the maternal cervix. In the former, this is called placenta previa. In the latter, this is called a low-lying placenta and is usually given a further categorisation such as anterior low-lying. Figure 4.1 shows diagrammatically these placenta locations. In clinical practice, a sonographer will assess placenta location by confirming that the placenta is sufficiently apart from the cervix. If the sonographer suspects the placenta is close to the cervix, they will measure between the lower placenta edge and the internal os. The lower placenta edge is the lowest observed point of the placenta. The internal os is the terminating end of the cervix, located inside the uterus. Clinical guidelines define a 2-cm threshold to differentiate between

normal and abnormally located placentae [104]. This method of assessment through measurement is extremely challenging for an operator, trained or otherwise. They are required to navigate to a plane to obtain an optimal image and identify two anatomical landmarks. Although the placenta is a large anatomical structure and has a characteristic echogenicity, the same is not true for the internal os, which is small and lacks obvious image features.

The contribution of this chapter is to suggest a new approach to assess placenta location by using an assistive overlay. Since we require the identification of anatomical landmarks, it is logical to model the problem using image segmentation. The next section describes the relevant dataset.

4.2 Dataset

To capture relevant anatomical structures, step E, a U-shaped video sweep was selected from the CALOPUS US protocol, see §3.1. This is a simple U-shaped video sweep which scans from maternal right to left over the pelvis. The transducer is held in the sagittal orientation. Figure 4.2 shows the trajectory of the video sweep. The sweep takes approximately 20s and captures approximately 600 frames per video. To sufficiently constrain the problem, both placenta and maternal bladder were selected as relevant structures to be automatically segmented. The placenta was selected because we are interested in assessing its spatial location in the uterus. The bladder was selected for two reasons. First, the maternal bladder is anatomically located above the cervix and this is a constraint of human anatomy. Second, as we have previously described in §3.1.6 (*Analysis of Bladder Fullness in Step E*), the inclusion of a full bladder in step E improves visualisation of the cervix.

We selected a subset of 150 videos from the CALOPUS dataset based on two further inclusion criteria. These were anterior placenta position and placenta-bladder concurrency at the video level. The anterior placenta position is located closer to the US

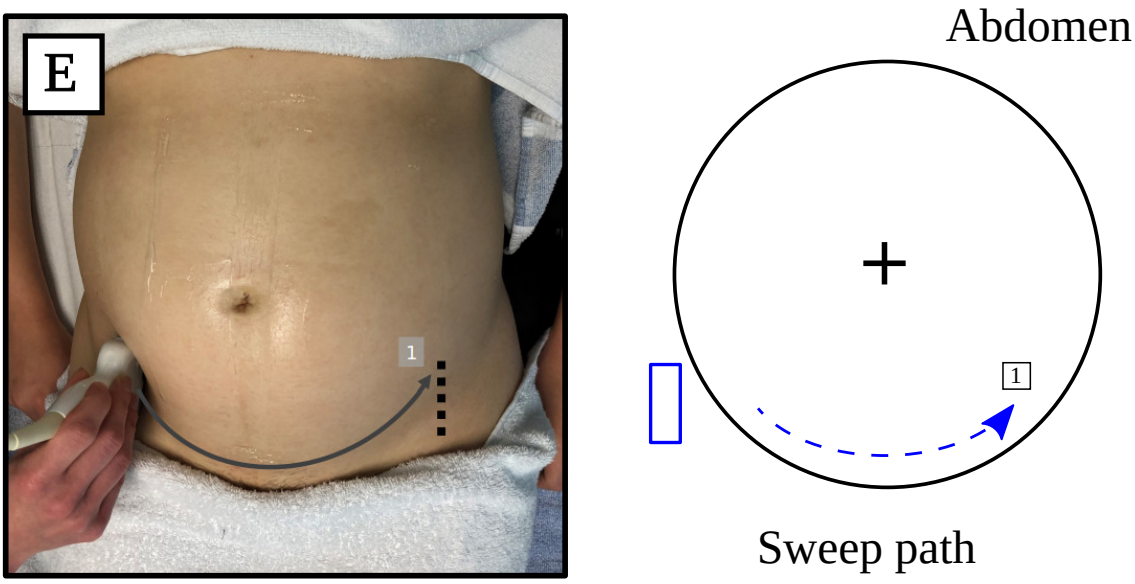

Figure 4.2: Step E trajectory. The rectangle shows the transducer orientation and + the umbilicus.

transducer and suffers from fewer imaging artefacts than other placenta positions, such as posterior. Placenta-bladder concurrency in a video was also required since we are interested in assessing video frames when these structures are co-located. The GA of the fetus was known, but for the purpose of this study ignored. This was because the objective was to assess the placenta location independent of GA. The gestational age range and mean of the 150 videos was $20 - 36$ and 22 weeks respectively.

To provide training data for a deep learning model, the placenta and maternal bladder were manually annotated at the pixel level using the computer vision annotation tool (CVAT) [101]. This was performed by two engineering researchers trained by a clinician to recognise and segment these structures in US video. In total, $3,114$ frames from 15 videos were manually segmented at the pixel level for the placenta and maternal bladder. Clinical supervision was provided throughout to ensure the quality of the manual segmentations. Table 4.1 summarises the number of manually segmented frames per video and corresponding fetal GA.

Table 4.1: Summary of manual annotations and data partitions. *x_y_Train* shows training data. Remaining videos have been used for testing.

ID	Manually Segmented Frames	GA (weeks$^{+\text{days}}$)	Total Frames
A_UK_Train	310	19^{+6}	987
B_UK_Train	158	36^{+1}	
C_UK_Train	182	20^{+3}	
D_UK_Train	168	20^{+0}	
E_UK_Train	169	20^{+0}	
A_UK	207	20^{+2}	1,808
B_UK	148	19^{+6}	
C_UK	190	20^{+6}	
D_UK	162	20^{+2}	
E_UK	270	30^{+5}	
F_UK	298	32^{+0}	
G_UK	299	35^{+6}	
H_UK	234	36^{+1}	
A_India	192	35^{+4}	319
B_India	127	35^{+4}	

4.2.1 t-SNE Analysis

We observed different categories of 2-D placenta shapes that appeared when manually segmenting frames in step E. Figure 4.3 shows three representative placenta shapes. We hypothesised that a spectrum of placenta shapes existed that describe how the US transducer obtains 2-D samples of the placenta, a naturally 3-D structure. We sought to quantify this spectrum of 2-D placenta shapes. Knowledge of this spectrum could be used to design an effective automatic segmentation model through prior understanding of shape [105]. Further, in this problem, we did not have access to 3-D data, such as MRI scans, which motivated the need to model the 2-D placenta shapes using a data-driven method, rather than develop an analytical model describing 3-D placenta geometry, such as in [106].

We used a statistical data visualisation technique, t-distributed stochastic neighbour embedding (t-SNE), which maps high-dimensional data to a 2-D or 3-D space [107]. t-SNE is used to reveal patterns in high-dimensional data which would otherwise be

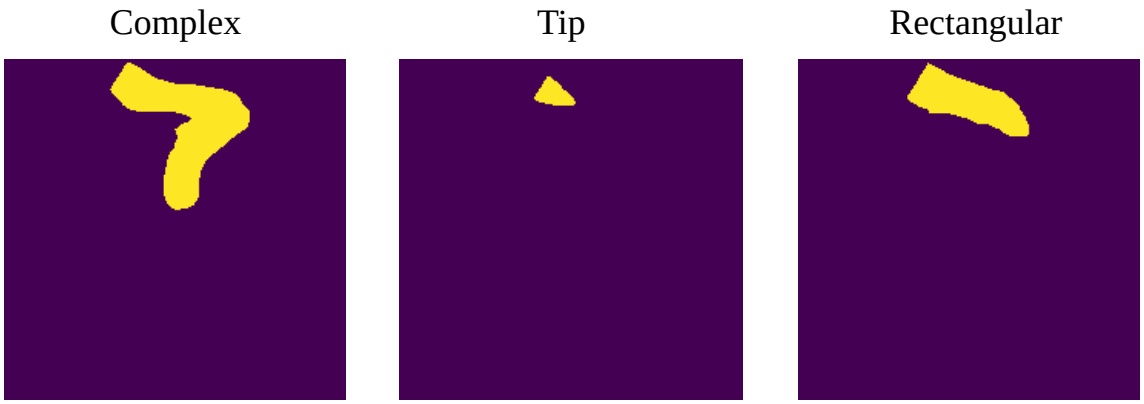

Figure 4.3: Three representative placenta shapes, shown as yellow on a purple background. Each shape is from a different subject.

difficult to visualise. t-SNE is a modified version of the original stochastic neighbour embedding (SNE) [108], but uses a Student t-distribution rather than a Gaussian, to compute the similarity between two points in the low-dimensional space. The similarity of points in the high-dimensional space are computed according to their conditional probabilities, as in the original SNE [108].

We performed a t-SNE analysis using $2,013$ manually segmented frames. These were binary masks that described 2-D placenta shapes. The binary masks were cropped, mirroring the same preprocessing of the US images, reducing the dimensions from $1920(x) \times 1080(y)$ to $1070(x) \times 820(y)$, before being resized to 224×224. The masks were flattened to produce vectors of length $50,176$. Principal component analysis (PCA) was used to first reduce the dimensionality to 30 components, as in [107]. This achieved a cumulative explained variation of 88%. t-SNE was then used with different values of perplexity and number of iterations. The perplexity can be considered as a smooth measure of the effective number of neighbours. We experimented with values of perplexity between 2 and 100 and iterations between $2,000$ and $10,000$. Each 2-D visualisation was assessed to find the optimal 2-D embedding, with respect to the parameters. By optimal, we mean observing consistent clusters that appeared only when the parameter selection matched well with the problem space. A perplexity of 30 and $1,000$ iterations matched this criterion.

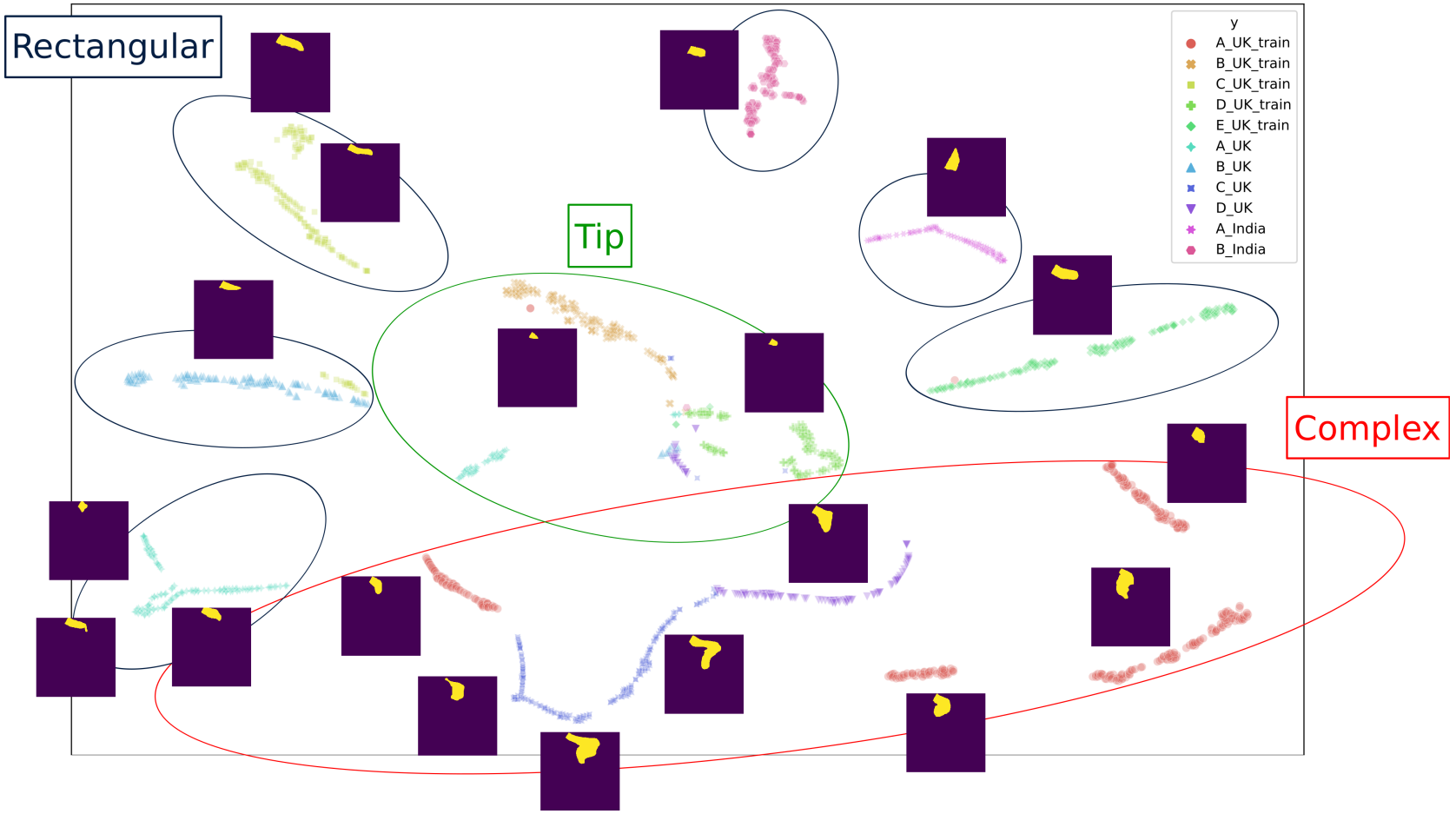

Figure 4.4: t-SNE visualisation of placenta shapes. We manually outline three clusters (shown in red, green, and blue) which represent placenta shapes of *complex*, *tip*, and *rectangular* respectively. A representative binary mask at different points of the embedding is also shown, to illustrate the observed placenta shape. The placenta is shown as a yellow object on a purple background. Each point is an embedding of a single video frame, colour-coded here by video instance.

Figure 4.4 shows the 2-D embedding from t-SNE analysis. Each point in the embedding is a binary mask from a manually segmented video frame, colour-coded by video instance. The placenta shape is directly shown in several key places as a yellow object on a purple background. We make several interesting observations of this analysis. We remark upon a cluster that appears, bounded by a red ellipse in Figure 4.4, which shows that the placenta shapes of three subjects (*A_UK_train*, *C_UK*, and *D_UK*) map to the same area in the 2-D embedding. This is interesting because these are the same subjects that we have visually observed to have a complex shaped placenta. In the entire CALOPUS dataset, we have observed other subjects with this placenta shape category. Similarly, we remark on two other clusters, one where the placenta shape is a tip (green ellipse), and second, where the placenta shape is approximately rectangular (blue ellipses). We show these shape categories of complex, tip, and rectangular in Figure 4.4 as clusters bounded by coloured ellipses of red, green, and blue respectively.

Figure 4.4 is interesting as it describes the spectrum of 2-D placenta shapes seen in step E. Visualisation of 2-D placenta shapes with t-SNE has not been explored before. The insight gained from this analysis motivated the design of the automatic segmentation model. In particular, we sought to model the highly variable placenta shape using a combination of deep learning and a probabilistic graphical model. We discuss the design of these next.

4.3 Deep Learning-Based Segmentation Model

In image segmentation, the goal is to learn a function, $\mathcal{F}$, that assigns probabilities to pixels in an image, $\mathbf{I} \in \mathbb{R}^{w \times h}$, from a set of N labels, $\mathcal{L} = \{x_1, x_2, ..., x_N\}$. Deep learning has achieved state-of-the-art performance on this computer vision task as it can learn features from the data itself [109]. Ronneberger, Fischer, and Brox [24] developed the seminal U-Net for automated segmentation of medical images using

deep learning. That study demonstrated the effectiveness of the proposed architecture on three datasets of different medical image modalities, including US. The use of skip connections allows fusion between coarse and fine features by concatenating inputs between corresponding layers of the encoder and decoder branches. The flexibility of the encoder-decoder design allows for simple modification of the standard architecture where arrangements of convolutional layers can be swapped out to address task-specific challenges.

In this regard, we used a modified U-Net architecture [110], combining the strengths of the encoder-decoder design with residual learning. The convolutional blocks are replaced with ResNet blocks [61]. These blocks, similar to [24], utilise a skip connection to map the original input (*i.e.* the identity), $\mathbf{x}$, to the convolved output, $\mathcal{F}(\mathbf{x})$. This improves both performance and convergence when training. In addition to the modified U-Net, we also compared several other architectures to provide a baseline. We used a fully convolutional network (FCN-8) [71] and an Attention U-Net [72].

4.3.1 Probabilistic Graphical Modelling

To effectively model the placenta shape, we combined our architecture with a conditional random field (CRF). This is a type of probabilistic graphical model that uses pixel labels to construct a Markov random field that is subsequently conditioned on a global observation (*i.e.* the image, $\mathbf{I}$). Krähenbühl and Koltun [111] investigated fully-connected CRFs, solving the otherwise intractable problem of many graph edges through a pairwise combination of Gaussians. Zheng et al. [6] formulated the conditional random field as a recurrent neural network (CRF-RNN), modelling the mean-field inference steps as common CNN operations. The energy (or cost) of assigning labels $\mathbf{x} \in \mathcal{L}^N$, to pixels in an image is given by:

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{4.1}$$

Here $\psi_u(x_i)$ is the unary energy components and represents the cost of assigning pixel i label x_i. Unary energy components are obtained as the output of the CNN. $\psi_p(x_i, x_j)$ is the pairwise energy components and represents the cost of assigning pixels i and j labels x_i and x_j simultaneously. The pairwise potentials are modelled as the sum of weighted Gaussians:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \underbrace{\sum_{m=i}^{M} w^{(m)} k_G^{(m)}(\mathbf{f}_i, \mathbf{f}_j)}_{k(\mathbf{f}_i, \mathbf{f}_j)} \tag{4.2}$$

Here $\mu(x_i, x_j)$ is the label compatibility function (of the Potts model) which assigns a fixed penalty if pixels of similar properties are assigned different labels. $\mathbf{f}_i$ represents the feature vector of pixel i and is derived from image features of spatial location p_i, p_j, and greyscale intensity (or RGB) values I_i, I_j, and is controlled by parameters θ_α, θ_β, and θ_γ:

$$k(\mathbf{f}_i, \mathbf{f}_j) = w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) \tag{4.3}$$

Subsequently, mean field inference that minimises the energy $E(\mathbf{x})$ is formulated as common CNN layers found in RNNs. We define the *negative* unary energy as $U_i(l) = -\psi_u(X_i = l)$ where X_i is the random variable of pixel i which represents the label assigned to pixel i. Finally, we define the partition function as $Z_i = \sum_l \exp(U_i(l))$ to obtain a valid probability distribution. Algorithm 1 provides the exact operations.

We used the implementation of the CRF-RNN to improve segmentation performance by relating the intensity and spatial location of pixels through a probabilistic graphical model. The placenta has a characteristic echogenicity in US and the CRF-RNN provided a structured way to model related pixel intensities. This information when combined with spatial location, provided constraints to model the highly variable placenta shape. The CRF-RNN parameters are optimised simultaneously during the

Algorithm 1: Mean-field inference in fully-connected CRFs, written here as common CNN operations [6].

input : negative unary energy $U_i(l)$)

output: refined segmentation mask $Q_i(l)$

$Q_i(l) \leftarrow \frac{1}{Z_i} \exp(U_i(l))$ for all i ; `// U from CNN`

while not converged **do**

- $\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l)$; `// Smoothing`
- $\check{Q}_i(l) \leftarrow \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l)$; `// Convolution`
- $\hat{Q}_i(l) \leftarrow \sum_{l' \in \mathcal{L}} \mu(l, l') \check{Q}_i(l)$; `// Convolution`
- $\breve{Q}_i(l) \leftarrow U_i(l) - \hat{Q}_i(l)$; `// Concatenation`
- $Q_i(l) \leftarrow \frac{1}{Z_i} \exp(\breve{Q}_i(l))$; `// Softmax`

end

normal training of CNN layers. This removes the need for offline post-processing, something common in other CRF implementations, and makes the CRF-RNN suitable for a point-of-care implementation. We experimented with combining the CRF-RNN with the U-Net and FCN-8.

4.3.2 Implementation

Training and test partitions were configured of 987 frames from 5 videos and $2,127$ frames from 10 videos respectively. Separation occurred at the video level to remove correlation between frames in the training and test sets. Since automatic segmentation was performed at the frame level, $2,127$ frames were tested. Table 4.1 summarises train and test partitions for all manually segmented frames. A further test partition of $73,308$ frames from 135 videos was also made. These frames did not have manual segmentations but were used to assess automatic segmentation performance qualitatively.

The models were implemented in the PyTorch framework (version 1.9.1) [112] and trained using a Nvidia RTX 2080 GPU (Nvidia Corporation). US frames were

extracted at a resolution of $1920(x) \times 1080(y)$ and cropped to remove any text or GUI information, reducing dimensions to $1070(x) \times 820(y)$. Frames were resized to 224×224. A preliminary analysis did not find noticeable differences when using different input sizes. Frames in the training set were selected at random and partitioned into batches of 10. A cross-entropy loss was used to update the convolutional parameters at each step. Data augmentations were applied to frames in a training batch to bolster the effective size and consisted of both spatial and intensity transformations. Our objective was to produce US realistic transformations that would be plausible as part of the acquisition protocol. This might include reversing the orientation of the transducer, small transducer rotations, and small perturbations in brightness or contrast. The spatial transformations consisted of random horizontal flipping, $\pm$ 6.25% translations, $\pm$ 10% scales, and $\pm$ 15° rotations, taken from the image centre. Grid distortions were also used, with 10 cells per frame. The intensity transformations consisted of random $\pm$ 20% perturbations to gamma, brightness, and contrast. Prior to training, all pixels in frames were normalised to zero mean and unit variance intensity.

Transfer learning was used to initialise convolutional and CRF-RNN parameters. This was justified since the number of manually segmented frames in the training set was relatively low. For the U-Net, parameters were initialised from a ResNet-18 ImageNet pretrained encoder [20] and for the FCN-8 and Attention U-Net, from their respective implementations [71, 72]. Training occurred for 250 epochs via the Adam optimizer, using a fixed learning rate of 1×10^{-4}. The CRF-RNN was added to both the U-Net and the FCN-8 and these underwent an additional 75 epochs of training, without data augmentations applied. The parameters of the CRF-RNN were initialised from its original implementation, which used a combined training dataset from Pascal VOC 2012 and MS COCO 2014 [6]. The number of iterations for mean-field inference of the CRF-RNN was set to 10, as in [6]. Both convolutional and CRF-RNN parameters were updated at the end of each training epoch.

4.3.3 Performance Metrics

We evaluated automatic segmentation performance using both area and shape metrics for $2,127$ manually segmented frames in the test set. The Dice coefficient is a score between $0-1$ which represents the area similarity between two sets, A and B:

$$Dice = \frac{2\,|A \cap B|}{|A| + |B|} \tag{4.4}$$

At evaluation, each frame was passed through the model under test and the Dice coefficient was computed by comparing the automatic segmentation to the manual segmentation. This allowed comparison of segmentation performance in terms of area. The mean of the Dice coefficient (mDice) was computed across all video frames in the test set to quantify the segmentation performance globally. To further quantify segmentation performance, $73,308$ frames were qualitatively assessed by checking each frame for correct segmentation of the anatomical outline. All frames were checked and this assessment was supervised by a clinical expert. All models were evaluated for both placenta and maternal bladder.

Hu moments are seven moment invariants calculated from a weighted average of the pixel intensities in an image. They are scale, rotation, and translation invariant and are used widely in computer vision tasks related to shape [113]. Subsequently, I_1, I_2, and I_3 are three shape similarity metrics which use aggregated Hu moments to compute a distance in a high-dimensional space between two shapes, A and B:

$$I_1(A,B) = \sum_{i=1}^{7} \left| \frac{1}{m_i^A} - \frac{1}{m_i^B} \right| \tag{4.5}$$

$$I_2(A,B) = \sum_{i=1}^{7} \left| m_i^A - m_i^B \right| \tag{4.6}$$

$$I_3(A,B) = \max_{i=1\ldots7} \left| \frac{m_i^A - m_i^B}{m_i^A} \right| \tag{4.7}$$

Here $m_i^A = sign(h_i^A) \log h_i^A$ and $m_i^B = sign(h_i^B) \log h_i^B$ where h_i^A and h_i^B are the i^{th} Hu moments from shapes A and B respectively. Taking logarithms is necessary to normalise all Hu moments to an identical scale. We used I_2 to compare segmentation performance in terms of shape. This is because I_2 has a useful property, that for small changes, the subtraction of two logarithms is approximately equal to the percentage change. In this context, this represents the percentage change in shape. I_2 is computed by summing the seven Hu moments. As these are scale, rotation, and translation invariant, this means that shapes of different spatial characteristics can be compared fairly. Area metrics, such as the Dice coefficient, only assess the degree of area match and are influenced by scale [114]. The Dice coefficient is not suitable for evaluating differences in shape, which motivated our use of the I_2 metric.

We evaluated automatic segmentation performance, in terms of shape, by comparing the U-Net + CRF-RNN over a baseline U-Net. We sought to quantify the contribution of the CRF-RNN in automatic placenta segmentation, in terms of shape. From the t-SNE analysis, in §4.2.1, we visualised the spectrum of 2-D placenta shapes seen in step E. This spectrum was interpreted to observe shape categories which were challenging to automatically segment. We observed that the U-Net + CRF-RNN generalised better to complex shaped placentas, compared to a baseline U-Net. To quantify the performance difference, we developed a novel metric, the percentage shape error, $\boldsymbol{\varepsilon}$.

For a single video, automatic placenta segmentations are generated. For a single segmented frame, the I_2 metric is computed by comparing the automatic segmentation to the manual segmentation. The I_2 value represents the distance between the automatic segmentation and the manual segmentation, in terms of shape. For small values, this is approximately equal to a percentage change. Assembling the I_2 frame-level values for every frame in a video produces a video-level vector, $\mathbf{I}_2 = [I_2^1, I_2^2, ..., I_2^N]$, where N is the number of frames in a video. $\mathbf{I}_2$ is computed

for both the U-Net + CRF-RNN ($\mathbf{I}_2^{unet+crf}$) and a baseline U-Net ($\mathbf{I}_2^{unet}$). Finally, the percentage shape error, $\boldsymbol{\varepsilon}$, is computed as the difference between the video-level vectors:

$$\boldsymbol{\varepsilon} = \mathbf{I}_2^{unet} - \mathbf{I}_2^{unet+crf} \tag{4.8}$$

The percentage shape error, $\boldsymbol{\varepsilon}$, is an N-dimensional video-level vector. The values represent the I_2 difference between a baseline U-Net and the U-Net + CRF-RNN. Positive values indicate that the U-Net + CRF-RNN segmentation is a better match in terms of placenta shape to the manual segmentation, compared to the U-Net. Negative values indicate the U-Net is a better match in terms of the placenta shape, compared to the U-Net + CRF-RNN. We visualised $\boldsymbol{\varepsilon}$ as video-level histograms for 6 videos, of which 2 have a complex shaped placenta. By interpreting positive or negative values on the histograms, we can quantify the contribution of the CRF-RNN with respect to the placenta shape.

4.4 Area Metrics

We report the mDice for the placenta and maternal bladder in Table 4.2. The results show that the U-Net + CRF-RNN scores the highest compared to all models tested. The placenta scores higher than the maternal bladder, an interesting result which we comment on later. The U-Net and U-Net + CRF-RNN score similarly, which highlights the importance of further evaluation with a shape metric. Figure 4.5 shows representative segmentation examples of the placenta and maternal bladder for the U-Net + CRF-RNN. A colormap (*viridis*) is used in Figure 4.5 to provide contrast to the overlaid segmentations. Video *C_UK*, shown with a red bounding box, is a more challenging example to segment due to a complex placenta shape. This is shown visually in Figure 4.5, where the placenta is partially segmented.

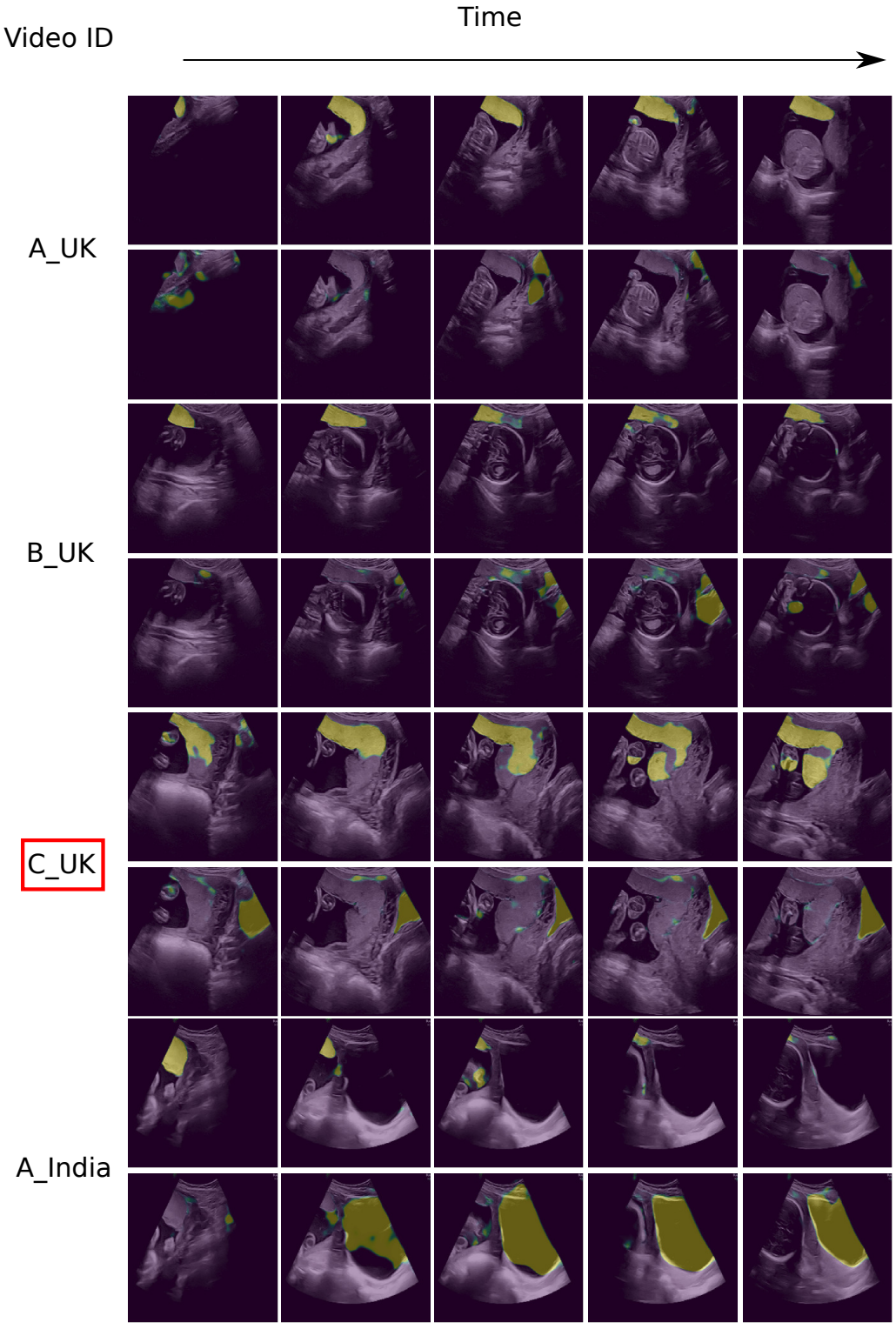

Figure 4.5: Representative automatic segmentation examples of placenta and maternal bladder by U-Net + CRF-RNN. The frames shown are from the manually segmented test set. The first row shows the placenta segmentation whilst the second shows the maternal bladder, repeated by video instance. Video *C_UK*, shown with a red bounding box, is a more challenging example to segment due to a complex placenta shape.

Table 4.2: Mean Dice coefficient (mDice) results on 2,013 manually segmented frames in test set. Bold and underlined entries are the first and second ranks respectively.

Architecture	**mDice**	
	Placenta	Bladder
FCN-8	0.71 ± 0.23	0.55 ± 0.30
U-Net	<u>0.82 ± 0.18</u>	<u>0.60 ± 0.29</u>
Attention U-Net	0.72 ± 0.23	0.51 ± 0.35
FCN-8 + CRF-RNN	0.56 ± 0.32	0.25 ± 0.37
U-Net + CRF-RNN	**0.83 ± 0.15**	**0.66 ± 0.24**

We report the results of the qualitative evaluation on automatic segmentation of the placenta and maternal bladder, for frames which did not have manual segmentations. The evaluation criterion was to ensure that the anatomical outline was correctly segmented. Figure 4.6 shows representative segmentation examples of the placenta and maternal bladder for the U-Net + CRF-RNN. As Figure 4.6 shows, for the placenta, all segmentations met the criteria of successfully segmenting the anatomical outline. There were frames where the placenta was not visible and a different anatomical structure was segmented. For the maternal bladder, the segmentations localised the structure but did not always segment the entire anatomical outline. We comment on both of these failures later.

4.4.1 Comparison to Prior Studies

We compared our U-Net + CRF-RNN model to three related studies in the literature which use a U-Net for automatic placenta segmentation to assess placenta location. We compared the mDice, in addition to other metrics reported in the respective studies. Table 4.3 shows a comparison of the mDice for our U-Net + CRF-RNN model against the conventional U-Nets of Saavedra et al. [28], Schilpzand et al. [29], and Arroyo et al. [5]. We report comparable results.

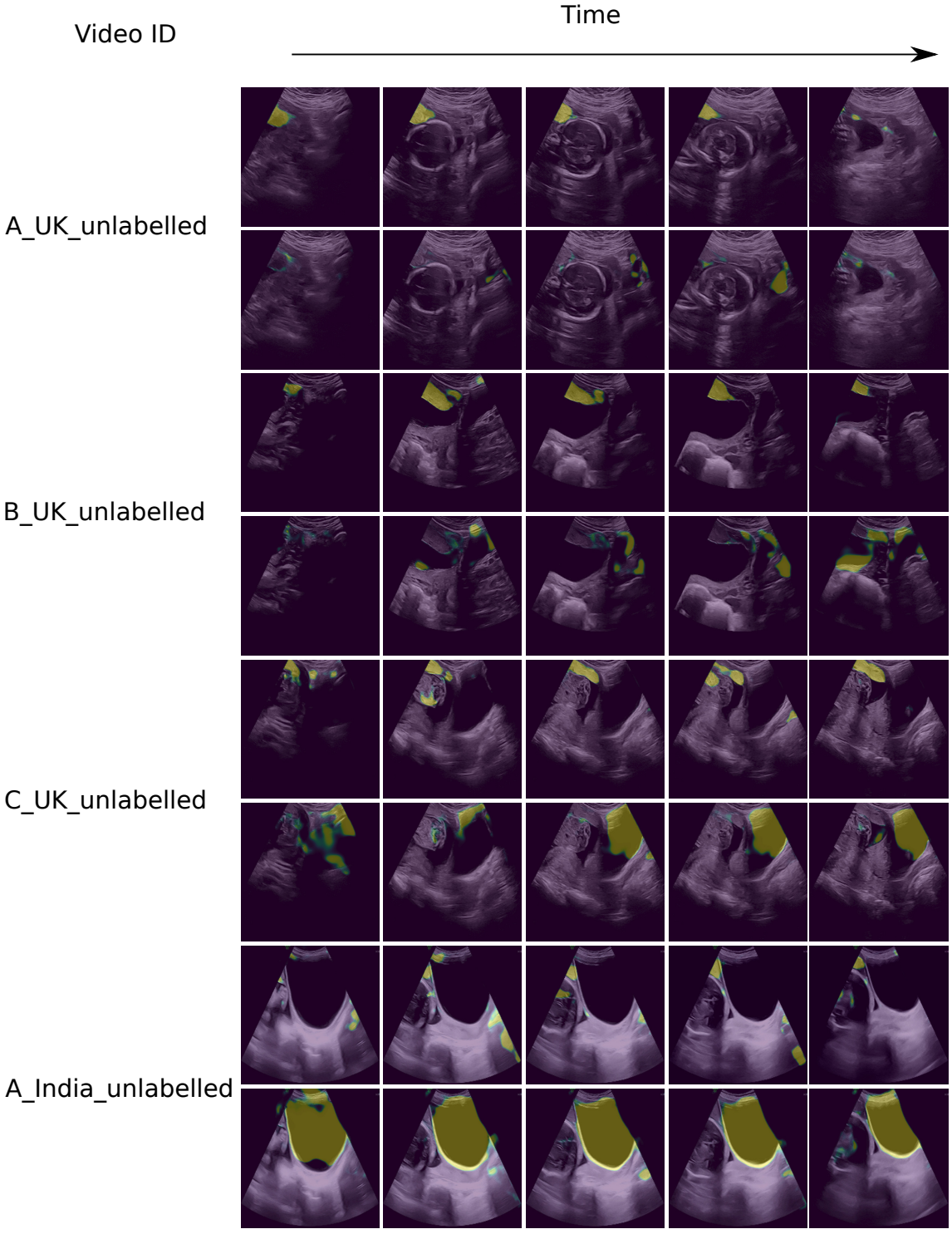

Figure 4.6: Representative automatic segmentation examples of placenta and maternal bladder by U-Net + CRF-RNN. Frames shown here did not have manual segmentations and were assessed qualitatively. First row shows placenta segmentation whilst the second shows maternal bladder segmentation, repeated by video instance.

Table 4.3: Metric comparison of our model to three related studies for automatic placenta segmentation to assess placenta location.

Architecture	**Metric**	
U-Net + CRF-RNN (Ours)	mDice	0.83 ± 0.15
U-Net (Schilpzand et al. [29])	mDice	0.84 ± 0.23
	True Detection Rate	0.95
U-Net (Saavedra et al. [28])	Sensitivity	75%
	Specificity	92%
U-Net (Arroyo et al. [5])	mIoU	0.86 ± 0.25

4.5 Shape Metrics

We report the percentage shape error, $\boldsymbol{\varepsilon}$, for 6 videos in the test set and visualise them as video-level histograms shown in Figure 4.7. The motivation was to quantify the performance benefit of the CRF-RNN, in terms of placenta shape. The histograms show the percentage shape error, $\boldsymbol{\varepsilon}$, calculated using Equation 4.8. Figure 4.7 is interesting as the histograms for *C_UK* and *D_UK* show a large positive skew. We illustrate this with red arrows. This means that the automatic placenta segmentation of the U-Net + CRF-RNN is a better match in terms of shape than the U-Net. Relating these findings to the t-SNE analysis in §4.2.1, we observe that videos *C_UK* and *D_UK* show a complex shaped placenta. We illustrate these videos with red bounding boxes. These are the only videos in Figure 4.7 which show a positive skew. This provides compelling evidence of the advantage of using the CRF-RNN, especially in the case of automatic segmentation of challenging placenta shapes. A maximum of 14% improvement in the percentage shape error is shown in Figure 4.7. No negative skew is observed in the remaining videos (*A_UK, B_UK, A_India, B_India*), which suggests that the CRF-RNN does not negatively affect segmentation, with respect to shape.

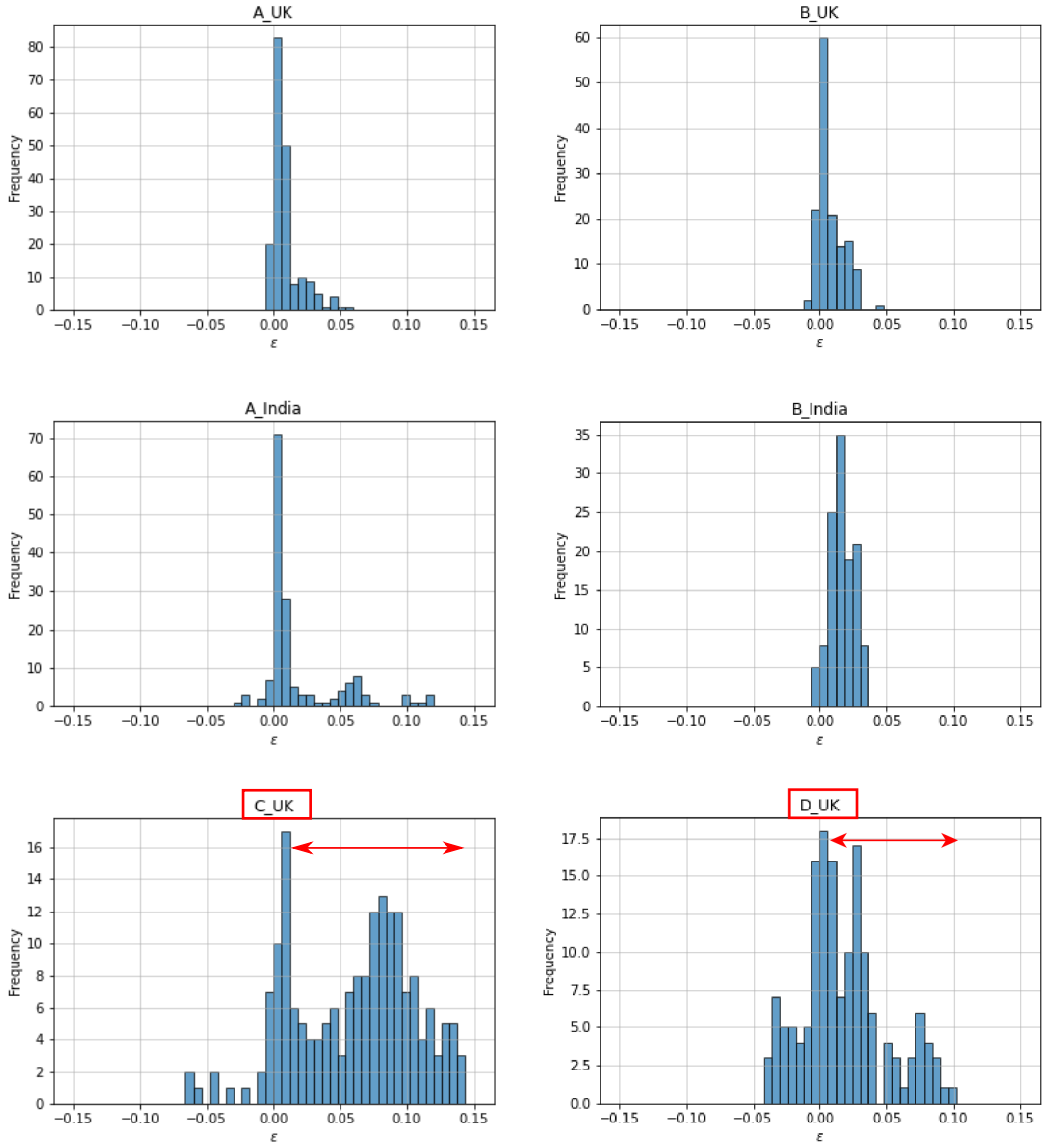

Figure 4.7: Percentage shape error histograms for 6 videos in test set. We consider *C_UK* and *D_UK* as complex shaped placentae (3rd row, shown in red bounding boxes) and *A_UK, B_UK, A_India* and *B_India* (1st and 2nd rows) as normal shaped placentae (*e.g.* tip and rectangular). The histograms show the percentage shape error, $\boldsymbol{\varepsilon}$, calculated usinmg Equation 4.8. We illustrate the positive skew in *C_UK* and *D_UK* (red arrows) which shows the U-Net + CRF-RNN has improved segmentation over a baseline U-Net, with respect to placenta shape.

4.6 Design of Assistive Overlay

We developed an approach to use the automatic placenta and maternal bladder segmentations to create an assistive video overlay. An operator can use the overlay to assist the interpretation of US video to assess placenta location. In clinical practice, a sonographer will assess placenta location by confirming that the placenta is sufficiently apart from the cervix. If the sonographer suspects the placenta is close to the cervix, they will navigate to a plane to obtain an optimal image, containing two anatomical landmarks. These are the lower placenta edge and the internal os. The lower placenta edge describes the lowest observed point of the placenta. The internal os is the terminating end of the cervix located inside the uterus. From the optimal image, they will manually identify each landmark and measure between them. Clinical guidelines define a 2-cm threshold to differentiate between normal and abnormally located placentae [104]. Measurements which are smaller than 2-cm diagnose the placenta as low-lying (*i.e.* abnormally located) and the pregnancy must receive further management [115, 116, 103]. Figure 4.8, 1, shows examples of US frames where this measurement has been performed, using standard US machine software that can generate an electronic calliper. Figure 4.8 illustrates the challenges in identifying the internal os; it is small and lacks obvious image features.

This method of assessment through measurement is extremely challenging for an operator, trained or otherwise. They are required to navigate to a plane to obtain an optimal image and identify two anatomical landmarks. We argue that it is simpler to visualise the allowable 2-cm clearance on the lower placenta edge directly. This removes the need to measure between the lower placenta edge and internal os. We further constrain the problem by proposing to use the bottom of the maternal bladder in place of the internal os, since anatomically the cervix is below the bladder. This reduces the complexity of the task, since the placenta and maternal bladder are naturally larger structures.

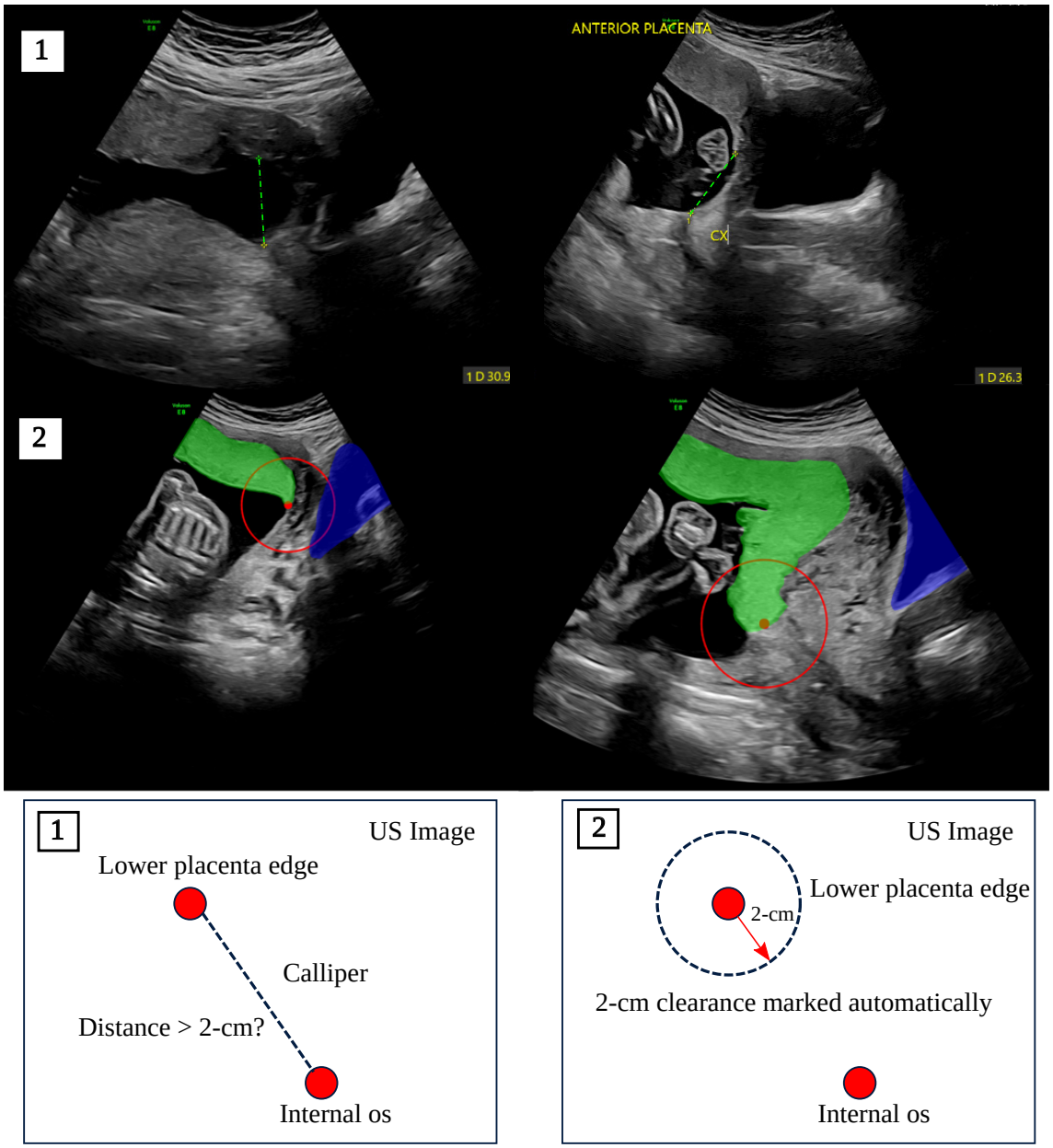

Figure 4.8: Comparison of current clinical practice and proposed overlay to assess placenta location. [1] shows the current method, with electronic calliper and measurement shown. The calliper has been manually highlighted for clarity. [2] shows the proposed overlay. The placenta and maternal bladder have been manually segmented as green and blue respectively, and the lower placenta edge is marked as a red point. The 2-cm clearance is showed as a red circle. Diagrammatic versions of both methods are shown in the third row.

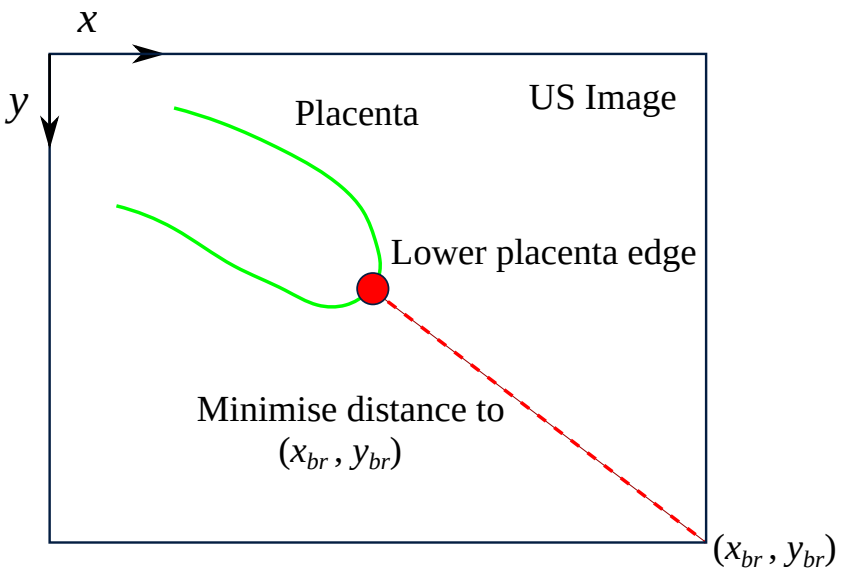

Figure 4.9: Diagram showing rule to automatically find the lower placenta edge. The pixel on the placenta segmentation (green) which minimises the distance to (x_{br}, y_{br}) is chosen.

We developed an image overlay to guide an operator in assessing placenta location this way. Figure 4.8, [2], shows the placenta and maternal bladder manually segmented in green and blue respectively, and the lower placenta edge manually marked as a red point with a circular 2-cm clearance. This information is overlaid on the original US image to provide image guidance on the identification of key anatomies and landmarks (*e.g.* placenta, maternal bladder, lower placenta edge), in addition to the allowable 2-cm clearance. It is possible to extend this approach to US video, where a single image becomes a video frame in a sequence. This creates an assistive video overlay. An operator interprets the information in the assistive video overlay to assess placenta location. Explicitly, they would check the 2-cm clearance does not intersect with the bottom of the maternal bladder.

The overlay is generated by post-processing automatic placenta and maternal bladders segmentations. Binary morphological operations are performed on each segmentation separately. An opening using a square 5×5 kernel is used. The largest segmentation connected component is selected and the rest discarded. The lower placenta edge is automatically found by searching for the closest pixel of the pla-

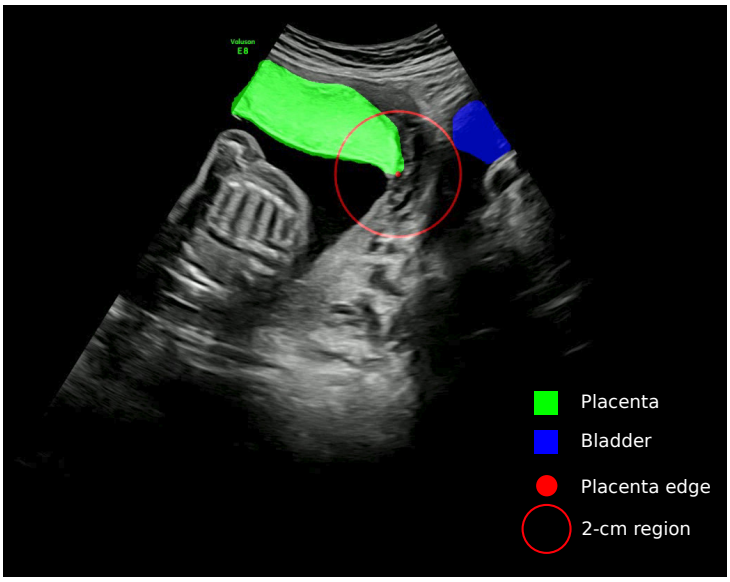

Figure 4.10: Example frame from US video with assistive video overlay. The placenta and maternal bladder are shown in green and blue respectively, the lower placenta edge and 2-cm clearance in red.

centa segmentation to the last pixel in the US frame, (x_{br}, y_{br}). Figure 4.9 shows this diagrammatically. The pixel which minimises the distance to (x_{br}, y_{br}) is chosen. This was a rule that we found empirically worked well in practice. This is likely because the internal os is commonly found near the bottom right of an US frame. US frames are combined with the colourised automatic segmentations, and the labelled lower placenta edge with 2-cm clearance. This is repeated for every frame in a video sequence to create an assistive video overlay. Figure 4.10 shows an example frame where the overlay has been generated automatically.

We generated an assistive video overlay for each of the 10 videos in the test set. Figure 4.11 shows representative frames from 4 videos, including a complex shaped placenta shown in a red bounding box. Videos *A_UK*, *A_India*, and *B_India* contain correct segmentations. In video *C_UK*, the segmentation has failed. This results in incorrect information in the overlay. The lower placenta edge is marked in the wrong location because the automatic placenta segmentation is incorrect. This failure is because of a complex shaped placenta in video *C_UK*.

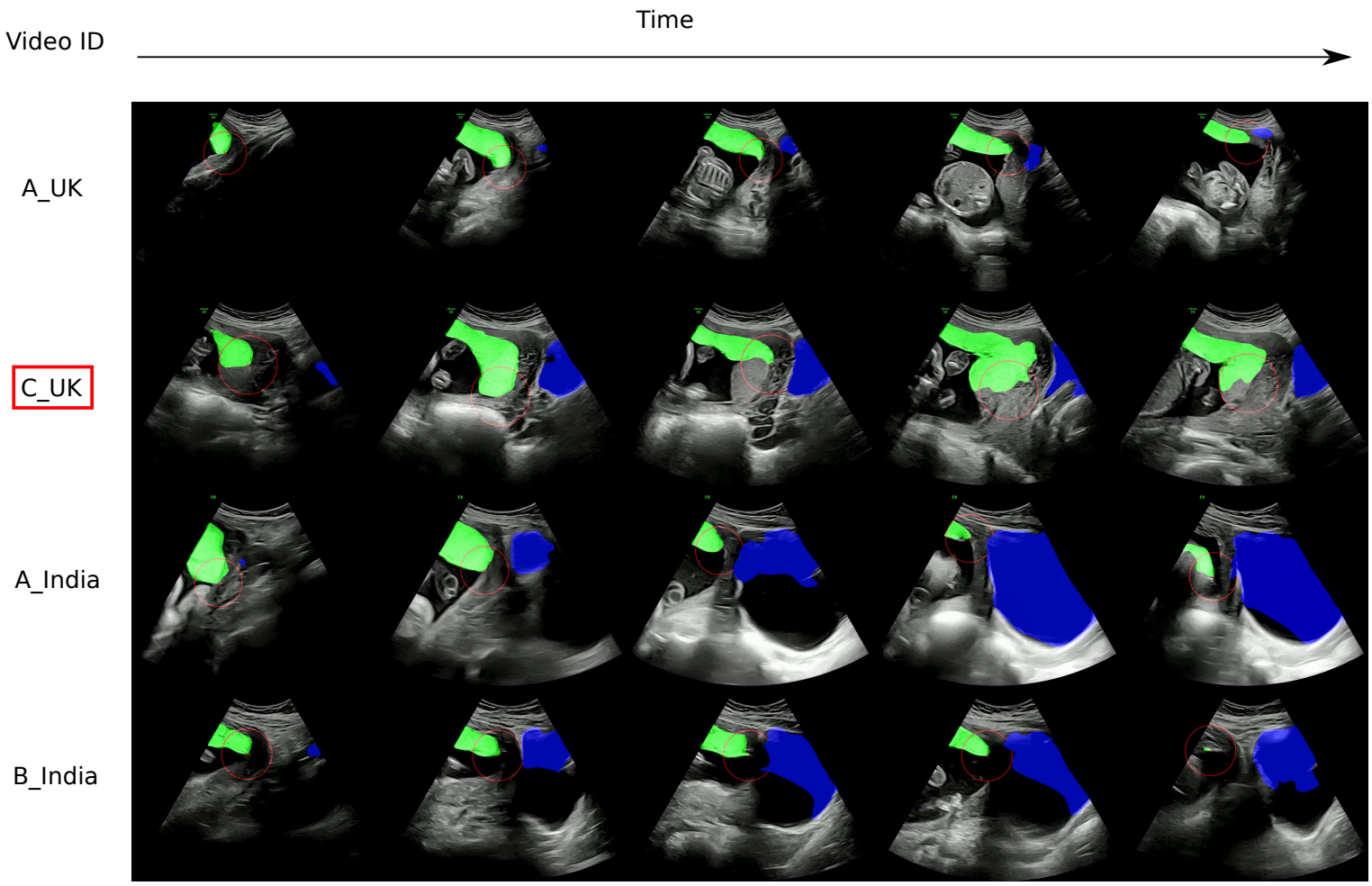

Figure 4.11: Representative frames of assistive overlays. In video *C_UK*, the segmentation has failed which leads to incorrect information in the overlay. *C_UK* shows a complex shaped placenta, shown in a red bounding box.

4.7 Discussion

We have described the major components of an US image analysis algorithm to assess placenta location. An overview of the algorithm is shown in Figure 4.12. We next discuss the significance of our results, in the context of this thesis, and the state-of-the-art.

We showed a spectrum of 2-D placenta shapes visualised by a t-SNE analysis in Figure 4.4. Interestingly, all complex shaped placentae (*A_UK_train*, *C_UK*, *D_UK*) map to approximately the same cluster. No supervision or further information is provided to the t-SNE algorithm except the binary masks describing placenta shape. Further inspection of the binary masks at other clusters confirms the validity of the three proposed shape categories (complex, tip, rectangular). We varied the hyperparameters of t-SNE but reached the same conclusion discussed here. Figure 4.4 is useful as it represents a 2-D spectrum of placenta shapes that are visualised in the U-shaped video sweep of step E. Analysis of placenta shapes within US video sweeps has not been explored with a visualisation technique such as t-SNE before.

Using the Dice coefficient, we showed that the U-Net + CRF-RNN is a superior architecture compared to other baselines (U-Net [24, 110], Attention U-Net [72], FCN-8 [71]). Table 4.2 shows the mDice for the U-Net + CRF-RNN evaluated on $2,127$ frames in the test set. For the placenta and maternal bladder, this was 0.83 $\pm$ 0.15 and 0.66 $\pm$ 0.24 respectively, the highest score observed for each structure. Table 4.3 shows the mDice for the U-Net + CRF-RNN compared to three related studies that use automatic placenta segmentation to assess placenta location [28, 29, 5]. We report comparable results, although we acknowledge different metrics have been reported in Saavedra et al. [28] and Arroyo et al. [5]. These studies do not evaluate beyond a single conventional U-Net, offering a clear advantage of our proposed U-Net + CRF-RNN. We employ the CRF-RNN to improve segmentation perfor-

mance by relating the intensity and spatial location of pixels through a probabilistic graphical model. We comment later on the clinical implications of our work in the context of previous studies for assessment of the placenta location using simple US video sweeps.

Using a shape metric, $\boldsymbol{\varepsilon}$, we showed the performance benefit of the CRF-RNN, in terms of placenta shape. Figure 4.7 shows that a maximum of 14% improvement in the percentage shape error, $\boldsymbol{\varepsilon}$, is achieved. Interestingly, the histograms of videos *C_UK* and *D_UK* show a positive skew. These are the same videos revealed in t-SNE analysis to have a complex shaped placenta. This suggests that the U-Net + CRF-RNN improves segmentation performance over a baseline U-Net, when faced with a complex shaped placenta. This is a valuable result as it justifies the inclusion of the CRF-RNN, something that area-based metrics (*e.g.* Dice coefficient) cannot do. We visually observed that the CRF-RNN improves segmentation performance on video frames of a complex shaped placenta, but produces similar segmentation performance for other shape categories (*e.g.* tip, rectangular). The absence of positive skew in the histograms of the remaining videos (*A_UK, B_UK, A_India, B_India*) in Figure 4.7 strengthens this argument. No negative skew is seen either which suggests that the U-Net + CRF-RNN does not produce any detriment over a baseline U-Net.

We showed in Figures 4.5 and 4.6 that the automatic segmentations correctly localise both structures, in addition to correctly segmenting the anatomical outline. There are two failure modes that we now discuss, see Figure 4.13. For the placenta, there were examples when the placenta was out of view and a different anatomical structure was segmented. This was particularly true for the fetal limbs. For the maternal bladder, there were examples when the entire anatomical outline was not segmented. We suggest that this is because of a lack of texture-based information. In US images, the bladder is sonolucent.

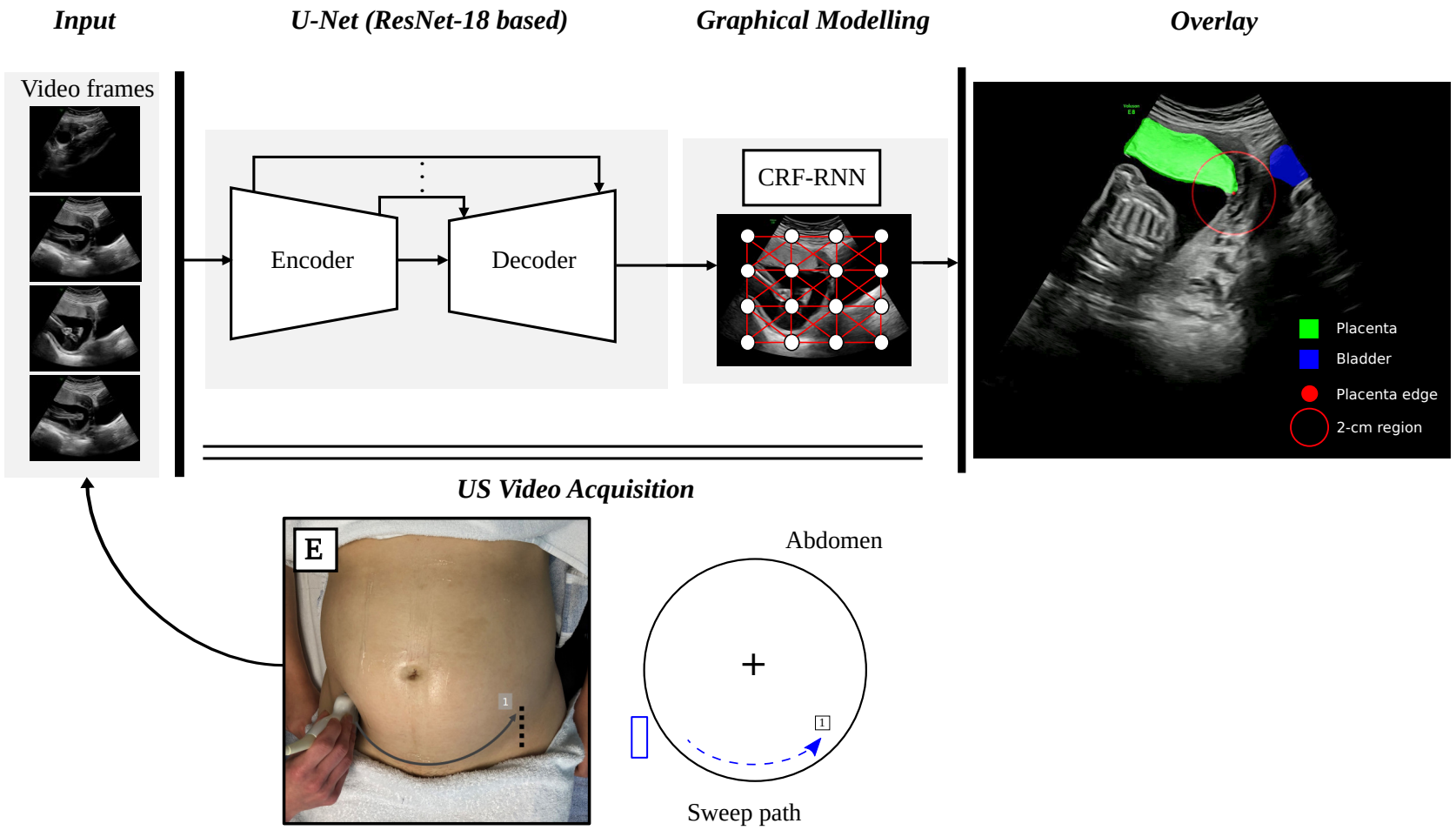

Figure 4.12: Overview of US image analysis algorithm. A simple U-shaped sweep (step E) obtains obstetric US video. The placenta and maternal bladder are automatically segmented using a U-Net + CRF-RNN. An assistive video overlay is constructed from automatic segmentations and US video frames.

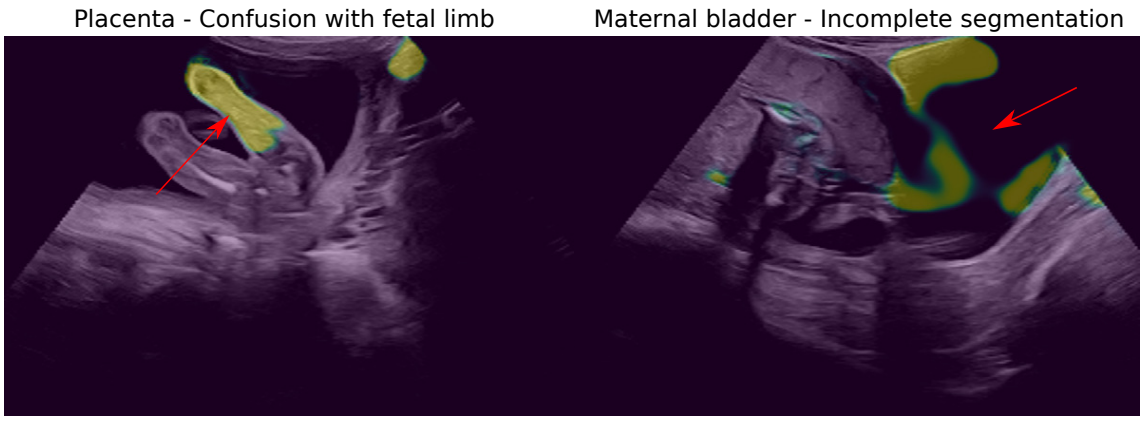

Figure 4.13: Representative frames from two failure modes of automatic placenta and maternal bladder segmentations. We use red arrows to show an incorrectly segmented fetal limb (L) and a partially segmented maternal bladder (R).

We showed in Figure 4.11 representative frames of the assistive video overlay, generated automatically from 4 videos. Although the addition of the CRF-RNN has clear performance benefit, the automatic segmentations are not perfect. There are 2 frames (from 1 video) shown in Figure 4.11 where the placenta is partially segmented. This results in incorrect image guidance information, since the whole structure is not highlighted. The lower placenta edge (*e.g.* centre of 2-cm clearance) is also marked in the wrong location. As before, we relate these findings to the t-SNE analysis. The failure is due to a complex shaped placenta in video *C_UK*. The remaining frames in Figure 4.11 show correct image guidance information. We extend the t-SNE plot in Figure 4.4 to give interpretability to the image analysis algorithm. Figure 4.14 shows the t-SNE plot where we have marked the location of 2 videos, one success (A) and one failure (B). For each, we show a representative frame of the assistive video overlay, to relate the image guidance information to the location on the t-SNE plot. Figure 4.14 provides an interpretable way to relate the spectrum of placenta shapes with the generated image guidance information.

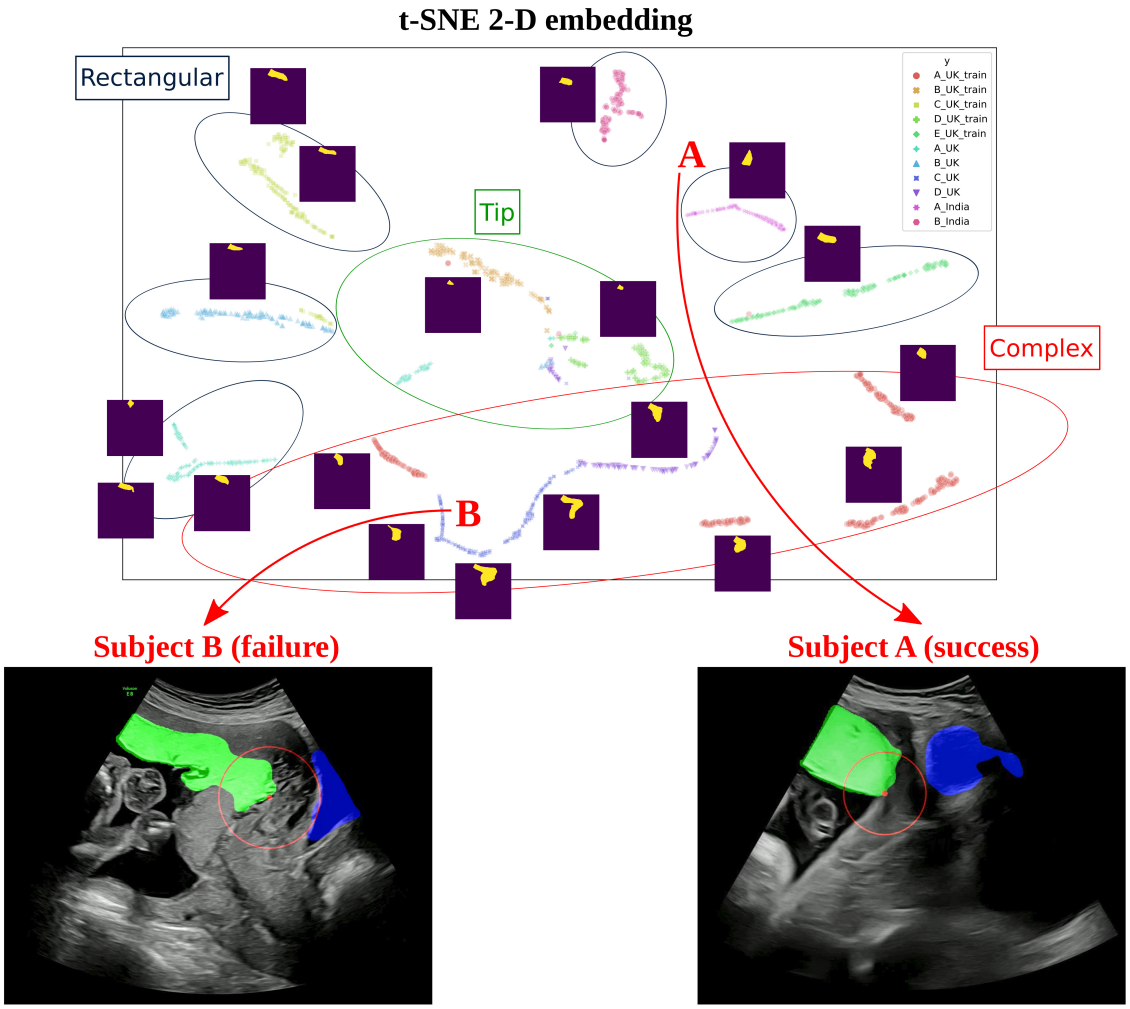

Figure 4.14: Extended t-SNE plot showing the location of a success (A) and a failure (B) subject in terms of the image guidance information. For each, a representative frame from the assistive video overlay is shown. Subject A (success) maps to the rectangular shape cluster, while subject B (failure) maps to the complex shape cluster.

We have shown proof-of-concept results for the assistive video overlay. The algorithm provides automatic image guidance to an operator by highlighting the relevant anatomical structures and landmarks. This does not require operator input or expertise. An operator with no prior experience in US could be trained to obtain obstetric US video by following step E. Sufficient training can be achieved in a single day [14]. An operator may be trained interpret the 2-cm clearance and the bottom of the maternal bladder to assess placenta location. In the context of this thesis, this may be sufficient for risk stratification when assessing placenta location.

A subject could then seek the appropriate level of care, such as that offered at a municipal hospital.

We now place our contribution in context of previous studies and comment on possible clinical implications. All previous studies [28, 29, 5] have used automatic placenta segmentation to fully automate assessment of placenta location. However, our algorithm requires an operator to review the assistive video overlay to assess placenta location (*i.e.* a human-in-the-loop solution). This means that an operator is likely more flexible than a machine when a catastrophic failure is encountered, a potential failure mode of fully automated algorithms. For example, if a low-lying placenta is misclassified as normal, this could cause harm, as the subject would not be risk stratified to receive the appropriate pregnancy management. Furthermore, long-term use of the overlay may contribute to an operator's training and familiarity in recognising structures in US. A final advantage is automatic maternal bladder segmentation. This constrains the problem by providing information on a second anatomical structure. We have shown in §3.1.6 (*Analysis of Bladder Fullness in Step E*), the value of a full maternal bladder in assessment of placenta location. All previous studies have only considered automatic placenta segmentation [28, 29, 5].

4.8 Conclusion

We have described an US image analysis algorithm to assess placenta location and a t-SNE analysis of 2-D placenta shapes. An overview of these contributions are shown in Figures 4.12 (algorithm) and 4.14 (t-SNE analysis of 2-D placenta shapes). The algorithm automatically creates an assistive video overlay that provides image guidance to an operator to assess placenta location. We show proof-of-concept results using real-world obstetric US video. The t-SNE analysis visualises the spectrum of placenta shapes that appear when the US transducer obtains 2-D samples of the placenta, a naturally 3-D structure.

In the context of this thesis, we have conducted an analysis of obstetric US video data from a simple U-shaped video sweep. It is natural to extend this reasoning to include multiple, complementary video sweeps. The elegance of this approach is that multiple sources of complementary video information exist. The research challenge is to determine how best to combine the video information. The remainder of this thesis addresses this valuable question.

Chapter 5

Graph-Based Representations of Multiple Ultrasound Sweeps

This chapter explores the analysis of multiple US video sweeps by creating graph-based representations. A three-node graph models three video sweeps, where each node encodes a binary sequence representing the fetal head frame-level detection across all video frames in a sweep. We first perform a statistical analysis of large-scale manual annotations from video sweeps in the CALOPUS US protocol. This reveals common frame-level patterns of anatomy occurrence for different video sweeps. Particular insight is gained in patterns that correspond to fetal pose. In this regard, we build a graph convolutional network (GCN) to automatically classify fetal presentation, creating graphs that combine complementary video sweep information relating to fetal pose. The edges of the graphs are weighted using three different metrics, in both a subject and template approach. This chapter is based on the following published work:

A. D. Gleed, D. Mishra, V. Chandramohan, Z. Fu, A. Self, S. Bhatnagar, A. T. Papageorghiou, and J. A. Noble. "Towards multi-sweep ultrasound video understanding: Application in detection of breech position using statistical priors". In: *International Symposium on Biomedical Imaging*. IEEE, 2023.

5.1 Introduction

Previously in this thesis, we have described our contribution in the analysis of single US video sweeps to assess placenta location. In §2.2 (*US Protocols in LMICs*), we reviewed the state-of-the-art for analysis of US video sweep protocols using machine learning-based algorithms. This has included assessment of fetal biometry [14, 5], fetal presentation [14, 21, 5], and placenta location [29, 5]. We argue that a significant limitation of these studies is that they do not combine complementary information obtained from multiple US video sweeps. This is something a trained sonographer can do well. A human operator can scan across the maternal abdomen following a prescribed path and observe an anatomical structure. With experience, they may anticipate when the structure will be observed again, perhaps in a different orientation, in relation to the US transducer position. This provides a two-fold clinical benefit. First, it can direct US transducer navigation, to seek certain structures relevant to the clinical task at hand. Second, more relevant in the context of this chapter, it can provide spatial understanding of multiple pieces of complementary information. This is essentially developing an intuition of the spatial location of fetal anatomical structures in relation to the US transducer position. This is something highly relevant in the context of analysis of US sweep protocols. Figure 5.1.1 illustrates this concept for the fetal head, an important structure for fetal biometry and presentation, with images obtained from both step B (axial sweep) and step C (sagittal sweep).

The contribution of this chapter is to propose a solution to the analysis of multiple US video sweeps. Specifically, we develop a graph-based representation that combines information from multiple US video sweeps. We begin by performing a statistical analysis of large-scale manual annotations from video sweeps in the CALOPUS US protocol. We describe this statistical analysis next.

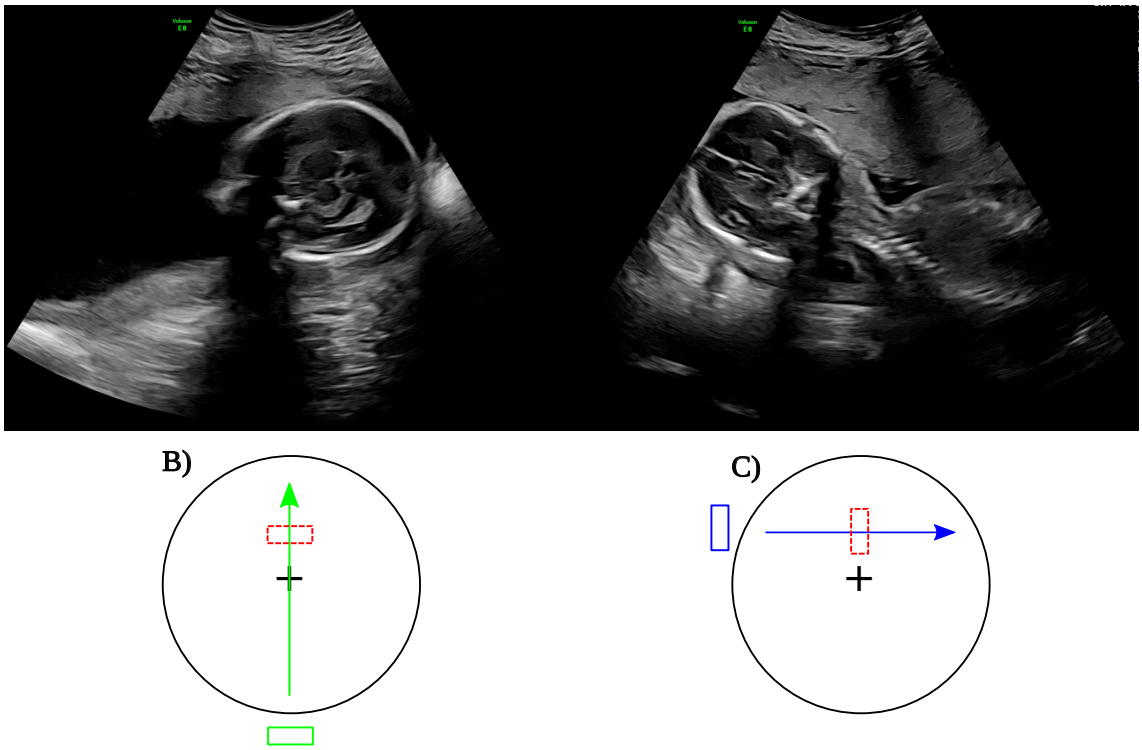

Figure 5.1: Fetal head in two frames obtained from step B (L) and step C (R) respectively. Rectangles show the transducer orientation and + the umbilicus. We show the approximate location of the US transducer in red where these frames are located.

5.2 Statistical Analysis of Video Sweeps

To characterise the video sweeps (*e.g.* steps A - E) in the CALOPUS protocol, we performed a statistical analysis of large-scale manual annotations generated by sonographers. As previously described in §3.3 (*Manual Annotations*), fetal and maternal anatomical structures were manually annotated at the frame-level, through a combination of rectangular bounding boxes and frame labels. Only steps A and E were annotated with bounding boxes. We hypothesised that the fetal pose is an important parameter in understanding how an US sweep protocol samples fetal anatomical structures. Fetal pose describes the spatial arrangement of fetal anatomy in the uterus. It is logical to suggest that the CALOPUS protocol obtains different information streams for different poses. We sought to understand this.

Table 5.1: Summary of manual annotations of steps A - E separated by fetal presentation.

Video Sweep	**Breech**	**Transverse**	**Cephalic**
Step A	84	95	149
Step B	33	16	42
Step C	32	18	43
Step D	42	39	71
Step E	27	29	40

We restricted our analysis to five fetal anatomical structures relevant to fetal pose: head, spine, abdomen, pelvis, femur. We collected frame-level manual annotations for steps A - E and separated them by fetal presentation. Fetal presentation is a simplification of fetal pose. It describes the spatial location of the head. This simplification was necessary since it is difficult to fully describe the fetal pose without additional information, such as a 3-D geometric model. We separated the annotations in each step by fetal presentation categories of breech, transverse, and cephalic. Table 5.1 shows the number of manual annotations for each step, separated by each presentation category. Each annotation corresponds to a unique video.

For each annotation, binary sequences were generated which encoded the presence (1) or absence (0) of a specific fetal anatomical structure. Five binary sequences were produced corresponding to the fetal head, spine, abdomen, pelvis, and femur. The length of each binary sequence was the number of frames in the corresponding video, N. Therefore, each binary sequence represented the frame-level detection of a specific fetal anatomical structure across all frames in a video. To compare binary sequences of different lengths, we normalised each sequence to a length of 100. Each sequence was divided into 100 segments of approximately equal length, n. The rounded mean was computed from all binary values in a segment. The rounded mean singularly replaced the multiple binary values in a segment. This reduced the length of each segment from n to 1, and overall the length of each binary sequence from N to 100. Algorithm 2 defines the normalisation of a binary sequence.

Algorithm 2: Normalisation of a binary sequence.

input : binary sequence of length N, where N is number of frames in video, x_i is the ith binary value in sequence

output: normalised binary sequence of length 100

divide binary sequence into 100 segments, approximately equal length n;

for each segment **do**

  compute $\tilde{x} = \text{round}(\frac{1}{n}\sum_{i=1}^{n} x_i)$;

  reduce dimensionality of segment from n to 1 ;

  set segment value to $\tilde{x}$;

end

The normalisation transformed each binary sequence to represent the frame-level detection of a specific fetal anatomical structure by the percentage of the corresponding video sweep, instead of the number of frames obtained. In the context of US video sweeps, this is useful as videos are obtained with an undefined number of frames, which is determined by the frame rate (fixed throughout this thesis) and the speed of the US transducer.

We used the normalised binary sequences to compute a sequence with values in the unit interval, $\{x_i \in \mathbb{R} \mid 0 < x_i < 1\}$, for each triad of fetal presentation, anatomical structure, and video sweep (*e.g.* breech presentation, fetal head, step A). For all normalised binary sequences in a triad, the arithmetic mean is taken at every ith position, from 1 to 100. This bounds the sequence to the unit interval $[0, 1]$, where each value is the probability of a positive frame-level detection of a fetal anatomical structure. Each sequence represented a statistical version of all normalised binary sequences in a triad. Figures 5.2 - 5.6 show the unit interval sequences for steps A - E. The sequences have been visualised as heatmaps using the *viridis* colormap. We hereafter refer to the sequences as statistical heatmaps to better describe the information they represent. We next discuss the significance of the statistical heatmaps in characterising the sweeps in the context of fetal presentation.

5.2.1 Step A

We remark upon different visual patterns in breech and cephalic presentations (BP, CP respectively) for the fetal head. We show these in Figure 5.2 with red bounding boxes. In BP, two distinct peaks are shown, yet in CP, there is only one. This is logical as the trajectory of the video sweep means that in CP, the head will be detected early (in 1) and not in the lateral arms (2, 3) of the video sweep. In BP, the head is detected later (in 1) and is also detected in the lateral arms (2, 3). These patterns show that step A is discriminative for the fetal presentation. We remark upon the location of peaks in each anatomical structure for BP and CP. An opposing step pattern is seen during 1. For BP, this begins with the femur and pelvis and steps upwards to abdomen, spine, and head. The opposite is true for CP. The pattern begins with the head and steps down with spine, abdomen, pelvis, and femur. This is a representation of the anatomical arrangement of fetal anatomy, as the US transducer begins at one structure and sweeps through the rest.

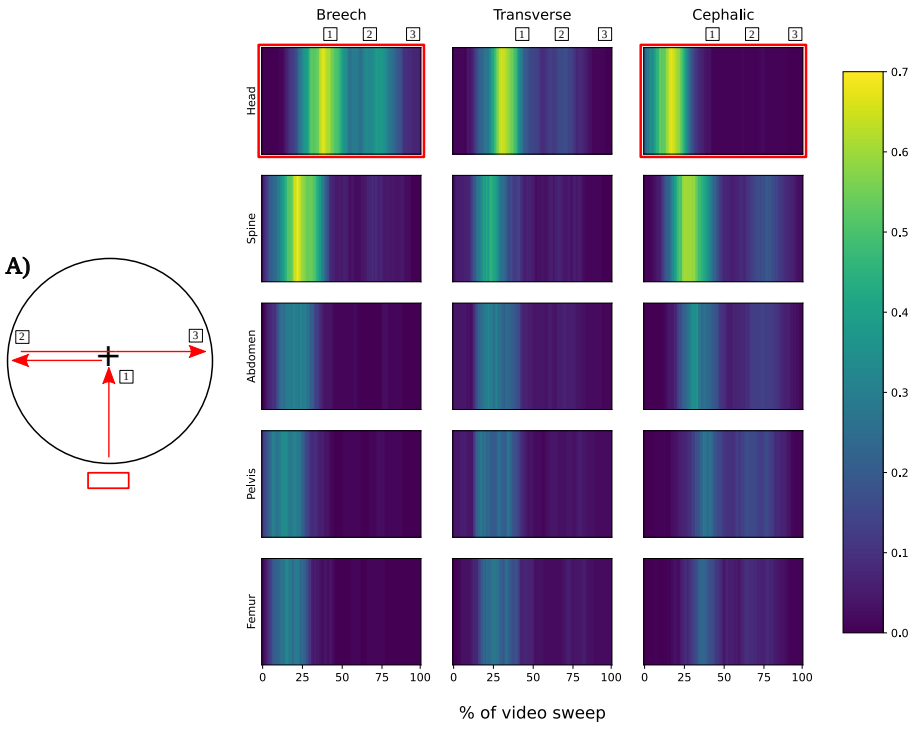

Figure 5.2: Statistical heatmaps of step A. Different visual patterns for the fetal head are shown in red bounding boxes.

5.2.2 Step B

We observe that the middle sweep (2) of step B has a higher probability for all fetal anatomical structures and presentations compared to the left and right lateral sweeps (1, 3). This suggests that the middle sweep (2) captures more video information on fetal anatomical structures than left or right lateral sweeps (1, 3). This is logical as the uterus is largest along the midline of the maternal abdomen. Interestingly, in transverse presentation (TP), the probability of detecting the spine along the middle sweep (2) is lower compared to BP and CP. This suggests that for a spine in a transverse orientation, an axial video sweep will obtain less video information compared to a spine in an axial orientation. We show this in Figure 5.3 with a red bounding box. We comment later on the opposite case for step D, when the video sweep obtains video information where the US transducer is in a sagittal orientation.

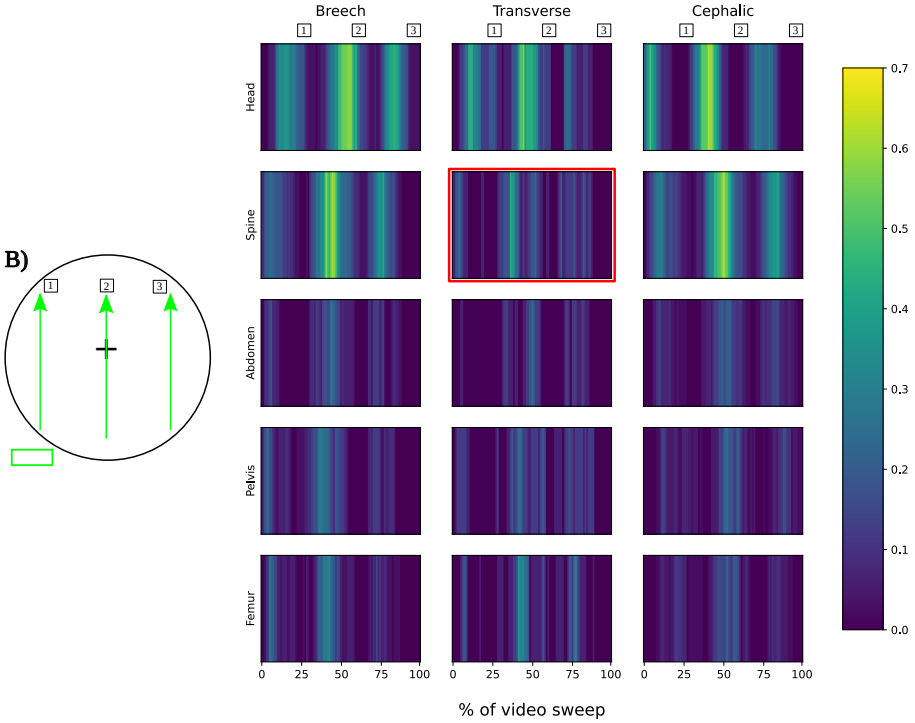

Figure 5.3: Statistical heatmaps of step B. A higher probability is observed for the middle sweep (2) compared to right or left lateral sweeps (1, 3).

5.2.3 Step C

We comment on different visual patterns seen in step C for BP and CP for the fetal head. We show these in Figure 5.4 with red bounding boxes. Two peaks are seen in BP (1, 2), whereas a single peak is seen in CP (2). Again, this is logical as in CP, the fetal head would not be located in the first video sweep (1), which scans above the umbilicus. This is a discriminative pattern for step C, similar to the pattern previously seen in step A, which describes the fetal presentation in terms of BP and CP. The probability of detecting the fetal head in CP in the second video sweep (2) is higher compared to the second video sweep in BP. This is likely because in CP, the fetal head is at the bottom of the uterus so the first video sweep (1) will not capture any video information about this anatomical structure. However, in BP, the fetal head is located around the umbilicus such that both video sweeps (1, 2) will obtain video information about this anatomical structure.

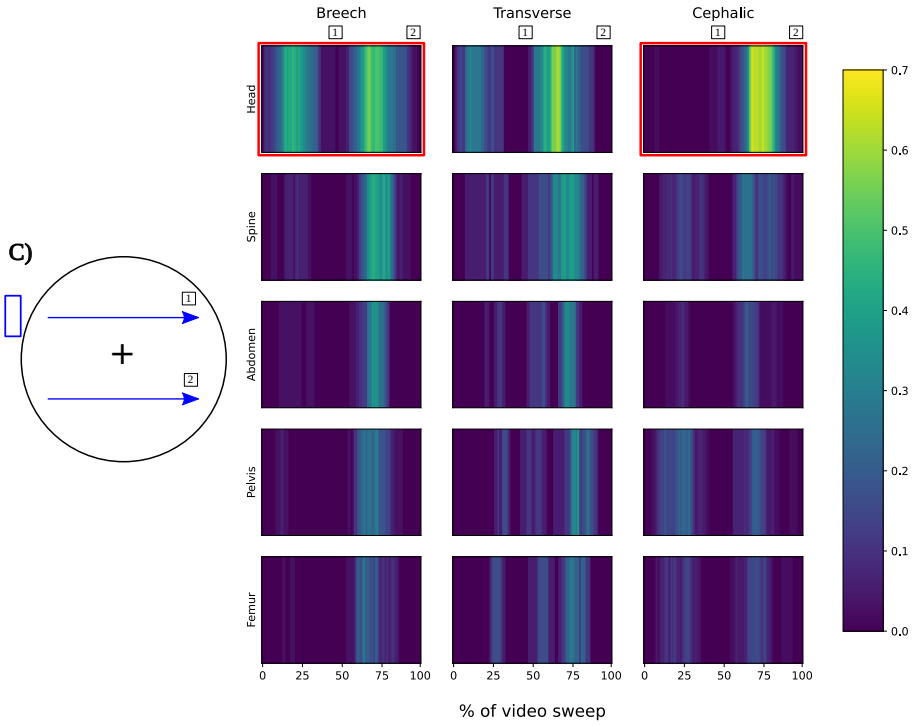

Figure 5.4: Statistical heatmaps of step C. Different visual patterns for the fetal head are shown in red bounding boxes.

5.2.4 Step D

We comment on the probability of detecting the fetal spine in TP. In step D, the US transducer is in a sagittal orientation. Interestingly, we see the reverse situation discussed previously in step B, where an axial video sweep obtains lower probabilities for the fetal spine in TP, compared to BP and CP. Here, we observe higher probabilities for detecting the fetal spine in TP, compared to BP and CP. We show this in Figure 5.5 with a red bounding box. We suggest that this is because more video information is captured when the US transducer scans over the spine perpendicular to its long axis. This is equivalent to the standard axial view of the spine which obtains cross-sectional views of the vertebrae. Higher probabilities for the fetal head are seen in the left and right arms of the video sweep ([1], [3]) for BP compared to CP. We suggest that this is because in CP, the fetal head will lie along the maternal midline. In BP, as the uterus is wider around the umbilicus, the fetal head can tend to the left or right as there is more space.

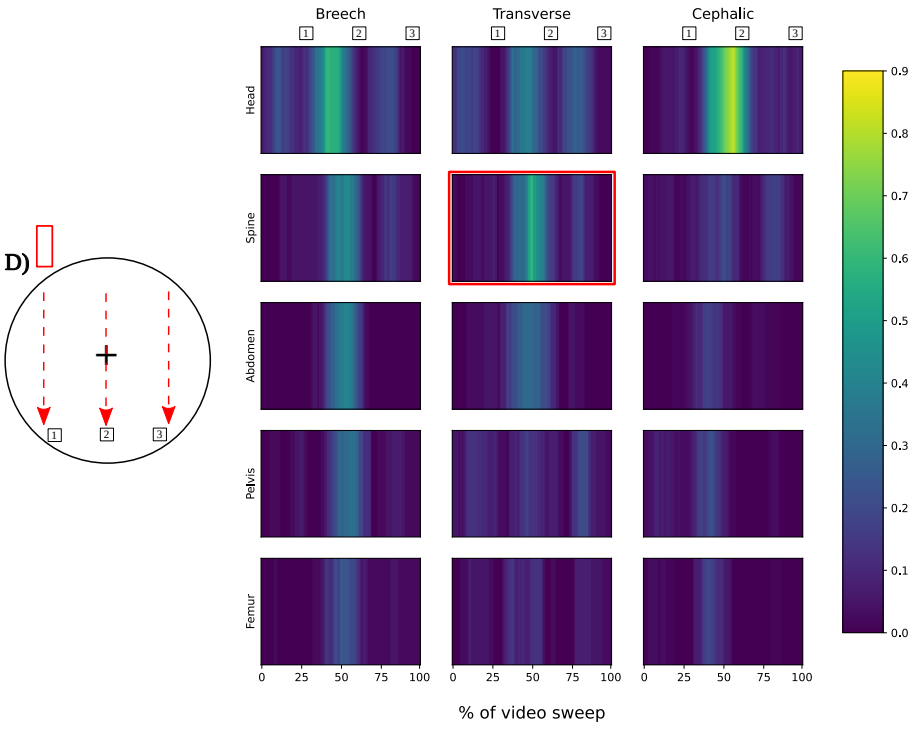

Figure 5.5: Statistical heatmaps of step D. A different visual pattern for the spine is shown in a red bounding box.

5.2.5 Step E

We observe different visual patterns in step E for BP and CP for the fetal head. We show these in Figure 5.6 with red bounding boxes. In CP, there is a distinct peak with a notably higher probability than in BP. Again, this is logical as in CP, the fetal head would be located low in the uterus such that step E would capture relevant video information on this fetal anatomical structure. Comparing each column of anatomical structures for BP and CP, we observe that they appear to follow opposite patterns. In CP, there is a peak for the fetal head and no peaks for the remaining anatomical structures (spine, abdomen, pelvis, femur). In BP, there are peaks for all anatomical structures except the head. In BP, the highest probability is for the abdomen. This is likely because it is the largest of the anatomical structures, ignoring the head, and would be positioned low in the uterus in BP. The abdomen does not have the same degrees of freedom of the femur, or the small size of the pelvis.

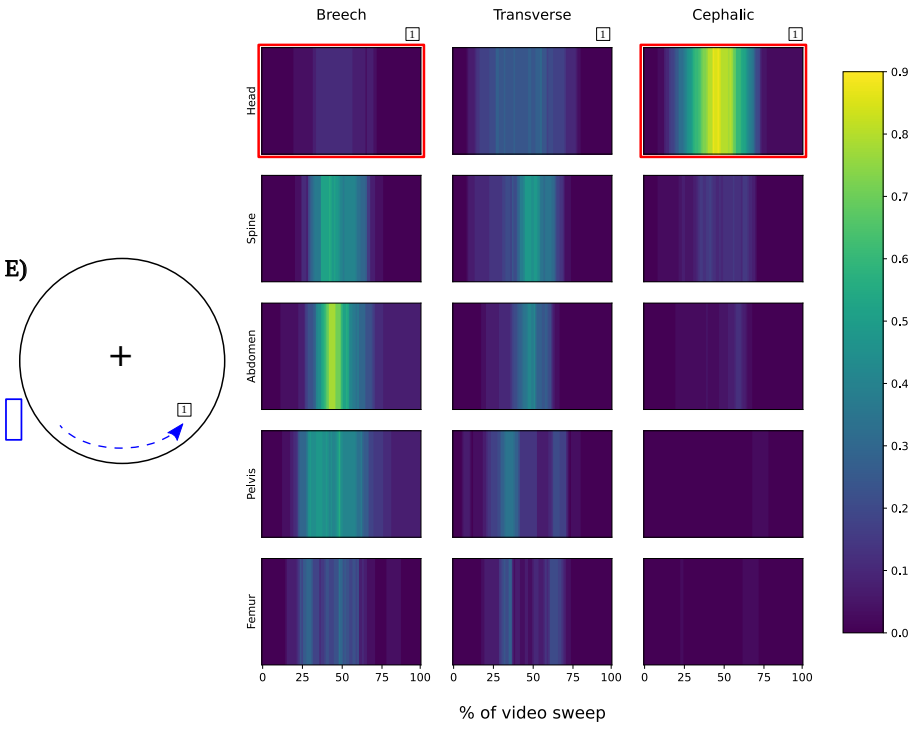

Figure 5.6: Statistical heatmaps of step E. Different visual patterns for the fetal head are shown in red bounding boxes.

5.2.6 Connecting Multiple Video Sweeps

The qualitative analysis of the statistical heatmaps of steps A - E is interesting as it characterises the frame-level anatomy detection patterns across US video sweeps of different trajectories. This analysis used large-scale manual annotations on real-world data recorded in hospitals. We can use this knowledge to select US video sweeps which show discriminative patterns that can be combined to create multi-sweep models to provide clinical benefit. Steps A, C, E show discriminative patterns of the fetal head for BP and CP. We observed a different number of peaks in steps A, C, E and a presence-absence pattern in step E. Figure 5.7 shows a summary of these fetal head discriminative patterns. To develop multi-sweep video understanding, we hypothesised that it would be interesting to combine the information in each video sweep with a relational model. Geometric deep learning has shown potential for machine learning using relational data, such as graphs. We describe the design of a geometric deep learning-based model next.

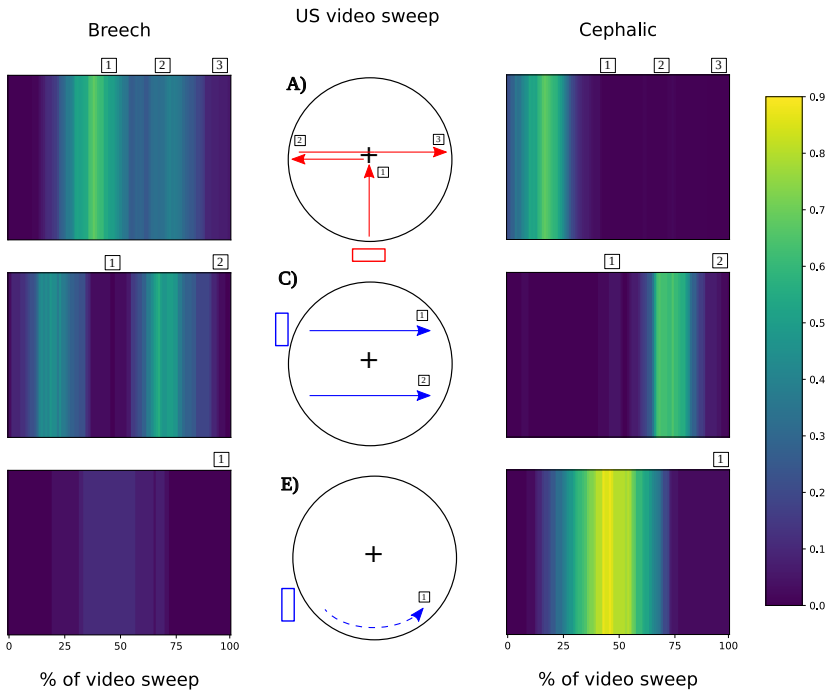

Figure 5.7: Fetal head statistical heatmaps of steps A, C, E, for BP (L) and CP (R).

5.3 Graph Convolutional Network

Graph neural networks (GNNs) are a class of deep learning-based models that operate on relational data, such as graphs. This is known as geometric deep learning. The earliest example is of Gori, Monfardini, and Scarselli [117], who proposed GNNs as an extension to recursive neural networks. Next, Hammond, Vandergheynst, and Gribonval [118] constructed wavelet transforms on graphs using spectral graph theory. This was developed further by Defferrard, Bresson, and Vandergheynst [119] who formulated Chebyshev approximations to polynomial filters. That study showed the equivalence of polynomial filters and standard convolutional layers in CNNs, thus enabling deep learning on graphs.

A graph, $\mathcal{G}$, is defined by a set of nodes, $\mathcal{N}$, and edges, $\mathcal{E}$.

$$\mathcal{G} = (\mathcal{N}, \mathcal{E}) \tag{5.1}$$

A graph's connectivity is defined by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where A_{ij} encodes the connectivity between nodes n_i and n_j, and N is the number of nodes in a graph.

$$A_{ij} = \begin{cases} 1, & \text{if an edge exists between nodes } n_i \text{ and } n_j. \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

Each node has a feature descriptor, $\mathbf{x}_i \in \mathbb{R}^D$, where D is the number of features. The node feature descriptors are summarised in a feature matrix, $\mathbf{X} \in \mathbb{R}^{N \times D}$. Each edge may be assigned an edge weight, w_{ij}. For the purposes of this chapter, we assume that the node feature descriptors are vectors and the edge weights are scalars. In practice, they can be either.

Writing a generalised neural network layer which takes a graph as an input, where $\mathbf{H}^{(0)} = \mathbf{X}$.

$$\mathbf{H}^{(l+1)} = f(\mathbf{H}^{(l)}, \mathbf{A}) \tag{5.3}$$

We can parameterise the neural network layer with a learnable weight matrix $\mathbf{\Theta} \in \mathbb{R}^{C \times D}$, where C is the number of channels. A non-linear activation function, such as a rectified linear unit (ReLU), allows multiple layers to be stacked.

$$f(\mathbf{H}^{(l)}, \mathbf{A}) = \mathrm{ReLU}(\mathbf{A}\mathbf{H}^{(l)}\mathbf{\Theta}^{(l)}) \tag{5.4}$$

There are two limitations to this model. Firstly, multiplication with $\mathbf{A}$ means that, for every node, we sum up the feature descriptors of all neighbouring nodes but not the node itself. This is solved by adding the identity matrix to $\mathbf{A}$, which is equivalent to adding self-loops in the graph, thus we define the identity-normalised adjacency matrix, $\hat{\mathbf{A}}$.

$$\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I} \tag{5.5}$$

Secondly, multiplication by $\mathbf{A}$ will change the scale of the feature descriptors based on the node degree. We can row normalise $\mathbf{A}$ so that every row sums to one, by multiplying by the inverse degree matrix, $\mathbf{D}^{-1}$, which simply encodes along the diagonal, the number of terminating edges at a node.

$$D_{ij} = \begin{cases} \mathrm{degree}(n_i), & \text{if } i = j. \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

This updates the definition of a neural network layer, where $\hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}$ is the row normalised adjacency matrix of $\hat{\mathbf{A}}$ and $\hat{\mathbf{D}}$ is the degree matrix of $\hat{\mathbf{A}}$.

$$f(\mathbf{H}^{(l)}, \mathbf{A}) = \mathrm{ReLU}(\hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{\Theta}^{(l)}) \tag{5.7}$$

Equation 5.7 is equivalent to taking the arithmetic mean of both a node feature descriptor and neighbouring node feature descriptors. This is mathematically simple, but it does not take into account the connectivity of neighbouring nodes. Kipf and Welling [89] proposed the graph convolutional network (GCN), which uses symmetric normalisation, replacing $\hat{\mathbf{D}}^{-\mathbf{1}}\hat{\mathbf{A}}$ with $\hat{\mathbf{D}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{\mathbf{1}}{\mathbf{2}}}$.

$$f(\mathbf{H}^{(l)}, \mathbf{A}) = \mathrm{ReLU}(\hat{\mathbf{D}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{H}^{(l)}\boldsymbol{\Theta}^{(l)}) \tag{5.8}$$

Symmetric normalisation considers the connectivity of both a node and its neighbours, which is useful when trying to model graph topologies. In the context of geometric deep learning, the largest eigenvalue of $\hat{\mathbf{D}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{\mathbf{1}}{\mathbf{2}}}$ is $\lambda_1 \leq 1$. This prevents exploding gradients which improves model convergence and performance during training. This allows stacking of GCN layers, similar to standard CNN layers. We used the GCN implementation of Kipf and Welling [89] for these reasons.

We constructed three-node graphs which encode information from three video sweeps. We used the insight gained in §5.2.6 (*Connecting Multiple Video Sweeps*) to select video sweeps steps A, C, E. We built subject-level graphs, where the node feature descriptors encode normalised binary sequences, representing the fetal head frame-level detection across all video frames in a sweep. Each node is connected to every other node. Figure 5.8 shows the graph construction process for a representative subject. The edges of the graph encode scalar weights.

A GCN was trained to automatically detect breech presentation from subject-level graphs, a task related to fetal pose. We used a two-layer GCN of 32 channels each, following Equation 5.8. Each layer is followed by a ReLU activation function. At the end of the GCN, global mean pooling is applied which obtains graph-level (arithmetic) mean node feature descriptors, followed by a fully-connected layer of 2 neurons which maps to classes BP and CP. The graph edge weights are encoded in $\hat{\mathbf{A}}$, which removes the need for a separate matrix.

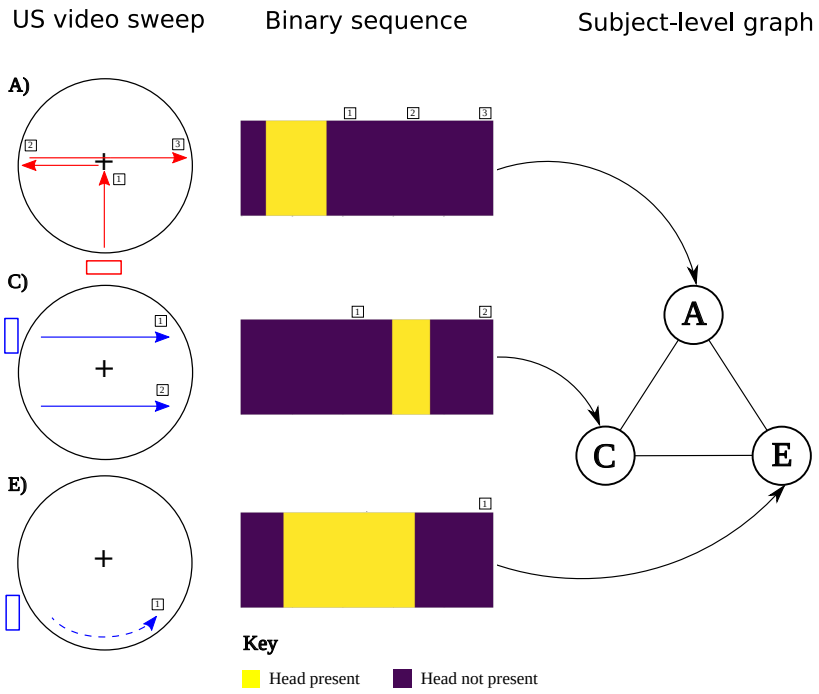

Figure 5.8: Graph construction using normalised binary sequences representing the fetal head frame-level detection across all video frames. Normalised binary sequences from a representative cephalic presenting subject are shown.

To provide a baseline, we also trained a three-layer multi-layer perceptron (MLP) using neuron counts of 32, 32, 2 for each layer respectively. Each layer is followed by a ReLU activation function. Therefore, the architecture of the MLP is identical to the GCN but learns from data where the graph's structure is not considered. This was used as a baseline to evaluate the benefit of using a GCN on relational data.

5.3.1 Metrics for Graph Edges

We experimented with graph edge weights of unity and defined by three metrics from signal analysis methods. We reasoned using descriptive edge weights would model the inter-sweep relationship for different fetal presentations, similar to how a sonographer may reason using multiple sources of complementary information in different scenarios. We compared three metrics: 1) discrete Fréchet distance, 2) dynamic time warping, and 3) Pearson correlation coefficient. We explored computing edge weights in a subject-specific approach and discuss this next.

Unity: Edge weights of unity reflect a binary graph, where each node is weighted equally compared to its own feature descriptor and its neighbours. Therefore, $\hat{\mathbf{A}}$ only encodes binary values. This was a baseline to compare to other metrics used as edge weights.

Discrete Fréchet distance: The discrete Fréchet distance (DFD), d_{dF}, is an approximation of the Fréchet distance for polygonal curves [120, 121]. For two curves, P and Q, the Fréchet distance, d_F can be described simply by imagining a man walking a dog on a leash. The man may move along one path (P), the dog on another (Q), both may vary their speed, but backtracking is not allowed. The Fréchet distance is the length of the shortest leash that can traverse both paths.

$$d_F(P,Q) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \left\{ d\Big(P\left(\alpha(t)\right), Q\left(\beta(t)\right) \Big) \right\} \tag{5.9}$$

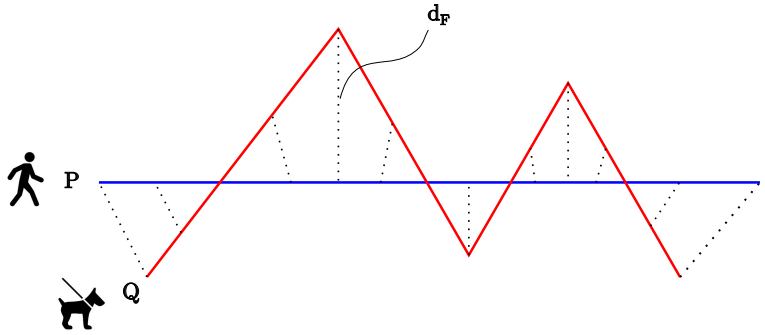

Figure 5.9: Fréchet distance for a man (curve P) walking a dog on a leash (curve Q).

Here, $P(\alpha(t))$ and $Q(\beta(t))$ are points on each respective curve, where $d(\cdot)$ computes the distance between them. Informally, t represents time. The discrete Fréchet distance, d_{dF}, is computed from a coupling, L, which is a sequence of pairs of points describing polygon line segments: $(p_1, q_1), ..., (p_n, q_n)$. The sequence of points is constrained such that the first point of P must match the first point of Q and similarly for the last points of both P and Q. The implementation of Eiter and

Mannila [120] reduces computational complexity by only considering four possible combinations. The length of a coupling $\|L\|$, is the maximum Euclidean distance between all pairs in a coupling.

$$\|L\| = \max_{i=1,...,n} \Big\{ d(p_i, q_i) \Big\} \tag{5.10}$$

The DFD is then the minimum of the length of all couplings in P and Q. The couplings describe the pairwise combinations of points according to the constraints above.

$$d_{dF} = \min_{\forall L \in P,Q} \Big\{ \|L\| \text{where } L \text{ is a coupling between } P \text{ and } Q \Big\} \tag{5.11}$$

The DFD is a similarity metric between two curves and models differences in curve shape. We used the DFD to represent the similarity between node feature descriptors, something that would be useful when modelling inter-sweep relationships. For a subject-level graph, we computed d_{dF} for each pair of nodes: A-C, A-E, C-E, using the node feature descriptors (normalised binary sequences) as P and Q. The edges of the graph were weighted by the scalar value of d_{dF} for each respective pair of nodes.

Dynamic time warping: Dynamic time warping (DTW), d_{tw}, computes the similarity between two curves that may vary in speed [122]. The key idea is that similar features that occur in different temporal scales can be detected by DTW. For two sequences of points, $\mathbf{p} = [p_1, ..., p_i, ..., p_n]$ and $\mathbf{q} = [q_1, ..., q_j, ..., q_m]$, arranged in an $n \times m$ matrix, i, j describes the alignment between points p_i and q_j. A warping path $\mathbf{w} = [(i_1, j_1), ..., (i_k, j_k)]$ maps points of p_i and q_j to minimise the Euclidean distance, $d(\cdot)$, between them.

$$D_{min}(i_k, j_k) = \min_{i_k, j_k} D_{min}(i_{k-1}, j_{k-1}) + d(i_k, j_k | i_{k-1}, j_{k-1}) \tag{5.12}$$

The cost of the warping path is computed by summing the Euclidean distances in **w**.

$$d_{tw} = \sum_{i=1}^{k} d(i_k, j_k) \tag{5.13}$$

Finding every combination of points is inefficient and sometimes intractable. Therefore, the first and last points of **p** and **q** must match and no backtracking is allowed. Also, we only permit path transitions to a single step forward in time.

For two similar curves, the cost of DTW is low. We used DTW to model the similarity between pairs of node feature descriptors. For a subject-level graph, we computed d_{tw} for each pair of nodes: A-C, A-E, C-E, using the node feature descriptors (normalised binary sequences) as **p** and **q**. The edges of the graph were weighted by the scalar value of d_{tw} for each respective pair of nodes. We used DTW as an additional metric, similar to the DFD. Since we had already normalised the binary sequences, DTW provided a way to model curve similarity in a way unaffected by the normalisation algorithm.

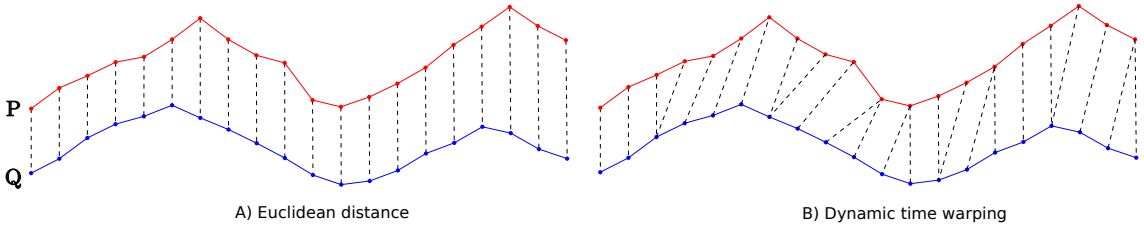

Figure 5.10: Comparison between A) Euclidean distance and B) dynamic time warping. Dynamic time warping returns the sum of the length of the warping path between curves P and Q (*e.g.* dotted lines in B)).

Pearson correlation coefficient: The Pearson correlation coefficient (PCC), p_{cc}, computes a score between -1 and $+1$ which quantifies the linear correlation between two curves. It is a measure of global synchronicity between two curves. The PCC is computed on a sample of a population, from pairs of points: $(p_1, q_1), ..., (p_n, q_n)$.

$$p_{cc} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2}\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}} \tag{5.14}$$

Here, $\bar{p}$, $\bar{q}$ are the (arithmetic) means of the points describing curves P and Q respectively. Unlike the DFD and DTW, it is possible for $p_{cc} < 0$, which represents negative correlation. This transforms our graph into a signed graph. We hypothesised that this would be useful in modelling negative inter-sweep relationships. For example, in BP, Figure 5.7 shows that it is likely that the fetal head will be detected at least twice in steps A and C each. However, in step E, it is unlikely the fetal head will be detected. In this case, A-E and C-E edge weights may be appropriately modelled with negative edge weights, since a presence-absence relationship is observed. Of all three metrics used in this study, only the PCC computes both negative and positive values. For a subject-level graph, we computed p_{cc} for each pair of nodes: A-C, A-E, C-E, using the node feature descriptors (normalised binary sequences) as P and Q. The edges of the graph were weighted by the scalar value of p_{cc} for each respective pair of nodes.

All metrics described thus far were computed in a subject-specific approach, where the node feature descriptors were unique to each subject. We sought to explore the possibility of using the statistical heatmaps in a way that would form a useful prior in a graph structure. In this regard, we repeated the statistical analysis described in §5.2 (*Statistical Analysis of Video Sweeps*), on a separate, independent set of subjects. The resulting statistical heatmaps can be considered statistical versions or priors of the node feature descriptors. We considered the feasibility of using these statistical priors to model edge weights which we named template edge weights.

5.3.2 Graph Edges from Statistical Priors

We used a separate, independent set of subjects to compute fetal head statistical heatmaps for steps A, C, E for BP and CP. Table 5.2 shows the number of manual annotations used for steps A, C, E separated by fetal presentation. Figure 5.11 shows the resulting statistical heatmaps. Figure 5.11 is virtually identical to Figure 5.7, which used the larger amount of manual annotations.

Table 5.2: Summary of manual annotations on holdout subjects for steps A, C, E separated by presentation.

	Breech	Cephalic
Step A	25	43
Step C	19	31
Step E	14	28

It is trivial to consider each statistical heatmap in Figure 5.11 as a node feature descriptor in a graph. Therefore, there are two ACE graphs each representing BP and CP. We named these graphs statistical priors, as they are essentially the average graph seen in each category of fetal presentation. To model graph edges, we computed edge weights between nodes A-C, A-E, C-E using three metrics described previously. We named these template edge weights, as they were computed from the statistical priors representing average graphs for BP and CP. Figure 5.12 shows numerical values of the template edge weights visualised as a distance matrix.

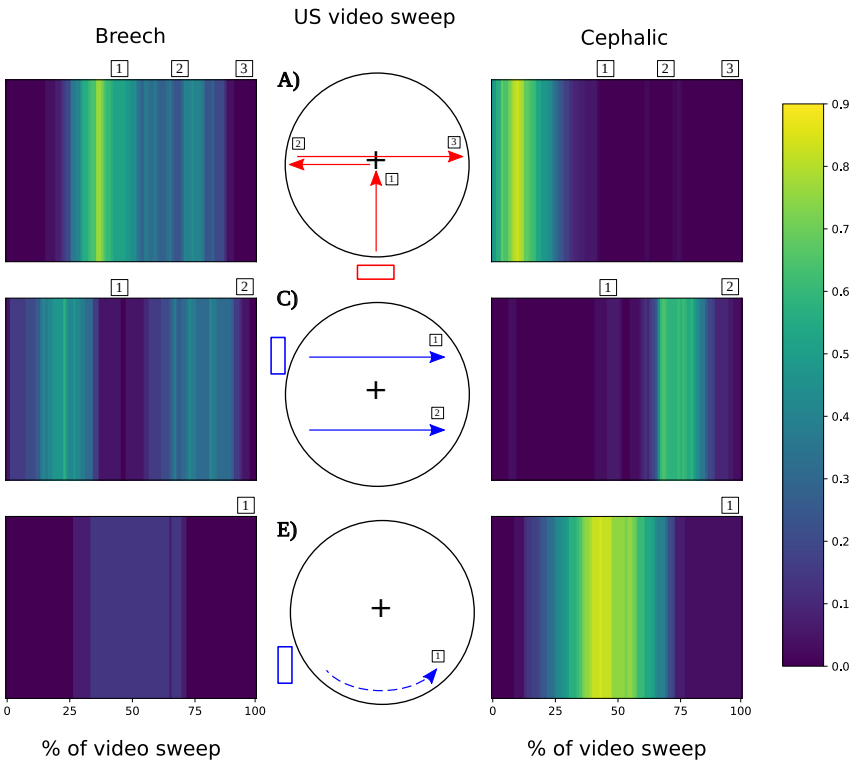

Figure 5.11: Fetal head statistical priors of steps A, C, E for BP (L) and CP (R).

Discrete Frèchet distance: The DFD only computes positive (distance) values between two curves. For breech, the maximum value is 0.62 between nodes A-E. This makes intuitive sense as A and E have the maximum and minimum peaks respectively in the ACE set. For cephalic, the maximum value is 0.72 between nodes A-C. This is the same scenario as before. Interestingly, if we compare values for nodes A-C between presentations, they are the minimum and maximum in each respective ACE set (*c.f.* 0.29 in BP to 0.72 in CP). This suggests that DFD edge weights between nodes A-C may be highly discriminative for classifying between breech and cephalic presentations, something that could be learned during a training scheme.

Dynamic time warping: Similarly, DTW only computes positive (distance) values between two curves, but values are usually > 1. This is because DTW sums up distance components along the warping path. For breech, the maximum value is 20.31 between nodes C-E. This is explainable by looking at the twin peaks of node C compared to the single peak of node E. Where there is a peak in C, there is a trough in E. For cephalic, the maximum value is 33.99 between nodes A-E. Again, this is explainable by comparing the location of single peaks in nodes A and E.

Pearson correlation coefficient: The PCC computes both positive and negative values as linear correlation can model either. For breech, the maximum and minimum values are $+0.72$ and -0.46 for nodes A-E and C-E respectively. Although the maximum probability for node E is lower than node A (*c.f.* 0.14 to 0.76), the overall shape is similar to node A, which leads to positive correlation. We discuss implications of this later. For cephalic, all values are negative with a minimum of -0.41 for nodes A-E. This means that all node feature descriptors are negatively correlated with each other, a property of their different shape. Comparing PCC values for breech and cephalic, a set of negatively correlated values describing CP may be useful in the context of geometric deep learning.

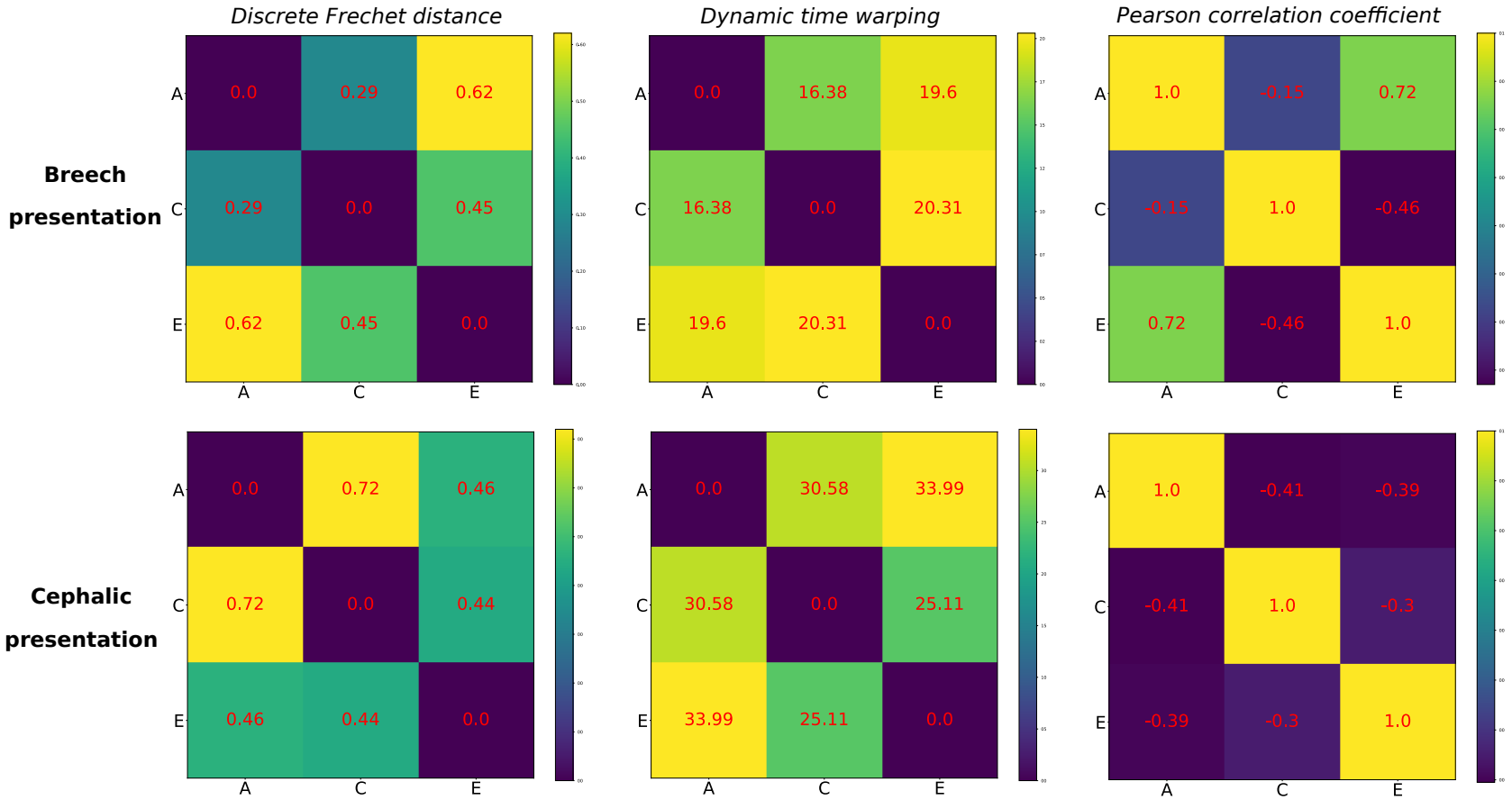

Figure 5.12: Numeric values of discrete Fréchet distance, dynamic time warping, and Pearson correlation coefficient computed between node feature descriptors A, C, E. Values are visualised as a distance matrix.

We explored combining template edge weights with subject-level graphs that had subject-specific node feature descriptors. As template edge weights existed for both BP and CP, the fetal presentation of a subject was used to select the correct template prior to any model training. We sought to explore whether template edge weights corresponding to a category of fetal presentation would help propagate relevant inter-sweep information in the context of geometric deep learning. We discuss the implications of this method later.

5.3.3 Automatic Annotation of US Video

To increase the number of annotated subjects available, a pre-existing VGG-16 model [123] was used to automatically produce frame-level annotations of the fetal head on US videos which did not have manual annotations. The VGG-16 model was trained to perform frame-level anatomy detection using an independent dataset of 123 second-trimester fetal US scans, acquired as part of the Perception Ultrasound by Learning Sonographic Experience (PULSE) project [124]. The anatomies consisted of: heart (inc. Doppler), 3-D mode, fetal head, maternal anatomy (inc. Doppler), spine, abdomen, nose and lips, kidneys, face-side profile, full-body-side profile, fetal bladder (inc. Doppler), femur, and a background class. For the purposes of this study, the fetal head frame-level detection consisted of all predictions made for the pooled classes of fetal head, face-side profile, and full-body-side profile. We used the VGG-16 model as it was an US-pretrained model capable of automatically labelling fetal head frames. Annotations generated by the model were used to construct further subject-level graphs.

We used the VGG-16 model to automatically generate annotations for 800 subjects, 400 in each presentation category. Visual inspection of the automatic frame-level annotations revealed some erroneous labels. This is not surprising, considering that the model was trained on an external dataset of second-trimester fetal US scans obtained by a sonographer. The video content of second-trimester US scans is very different

compared to simple US video sweeps, as a sonographer navigates the transducer to obtain optimal images of standard planes. We manually pruned the automatic annotations by visual assessment. The fetal presentation of a subject was used to select the corresponding statistical prior to provide an expected signal shape. Automatic annotations which did not follow the expected shape were removed. Figures 5.13 and 5.14 show this process for breech and cephalic respectively. To explain this further, consider Figure 5.13. The left hand side shows acceptable annotations for a breech presenting subject where the frame-level labels agree with the statistical prior (*c.f.* acceptable A, C, E frame-level labels (L) and corresponding statistical prior (M)). The right hand side shows erroneous annotations for a different breech presenting subject, where the automatic frame-level labels are implausible and disagree with the statistical prior (*c.f.* erroneous A, C, E frame-level labels (R) and corresponding statistical prior (M)). Figure 5.14 shows the same but for two different cephalic presenting subjects. After pruning, 317 subject-level annotations remained, 156 and 161 for breech and cephalic respectively.

5.3.4 Dataset

We used a total of 865 subjects in the CALOPUS dataset to create subject-level graphs. Since three video sweeps, steps A, C, E, were required per graph, this was a total of $2,595$ unique videos. A combination of manual and automatic annotations were used, in addition to video data from both sites. Figure 5.15 shows the dataset graphically. We balanced the number of breech and cephalic presenting subjects throughout to avoid a minority class. This is important since breech presentation accounts for approximately $3-4\%$ of fetuses clinically [125]. No subject in the 865 cohort was used to generate statistical priors or template edge weights. This maintained total separation between the two subsets of data, an important point we revisit later.

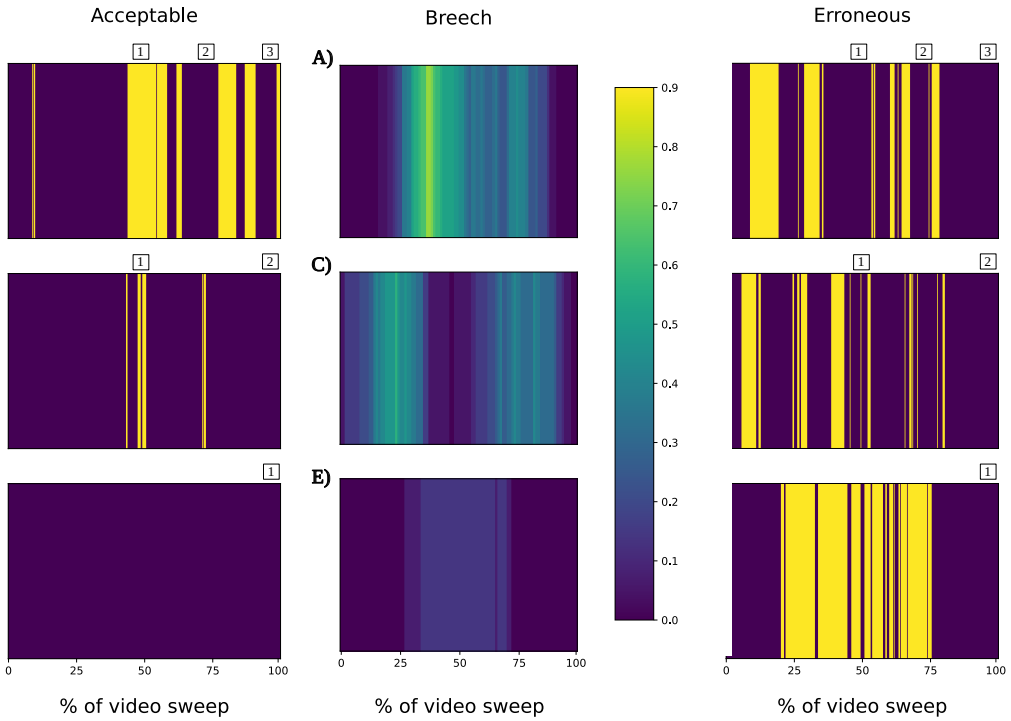

Figure 5.13: Comparison of acceptable (L) and erroneous (R) automatic annotations for two unique BP subjects to the corresponding statistical prior (M).

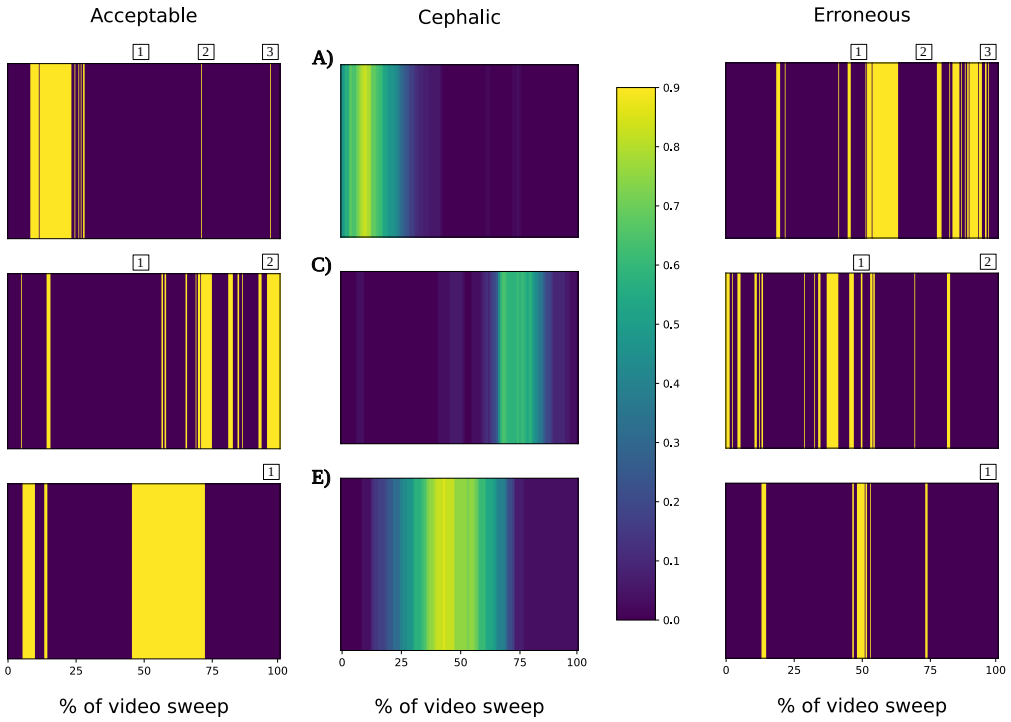

Figure 5.14: Comparison of acceptable (L) and erroneous (R) automatic annotations for two unique CP subjects to the corresponding statistical prior (M).

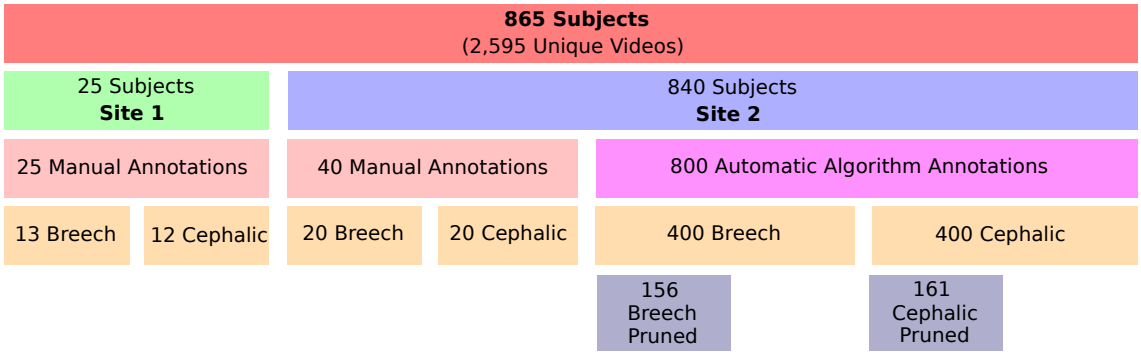

Figure 5.15: Summary of dataset showing the division of subjects by site, type of annotation (manual and automatic), and fetal presentation.

5.3.5 Implementation

To build a GCN to classify subject-level graphs, we implemented two train-test partitions, shown graphically in Figure 5.16. Each partition was constructed with a specific objective in mind. In partition (1), we sought to evaluate the performance of subject-level graphs using only high-quality, manually annotated data. A total of 65 subjects were used. The train and test sets were comprised of 40 and 25 subjects respectively. In partition (2), we sought to evaluate performance of subject-level graphs when training on manually annotated data, but testing on automatically annotated data. A further motivation was to avoid training on automatically annotated data, as the quality of these annotations had only been qualitatively assessed.

For both partitions, we explored assigning unity edge weights and subject-specific edge weights to all train and test subject-level graphs using three computed metrics. Furthermore, for each test set, we assigned template edge weights from the corresponding statistical prior. The presentation of a subject was used to select the correct breech or cephalic template of edge weights and this was performed offline to the training process. We only used the template edge weights in either test set at the point of evaluation when training had completed. This was to ensure the GCN did not learn from this indirect representation of the class label (*e.g.* breech or cephalic presentation) in an inappropriate way.

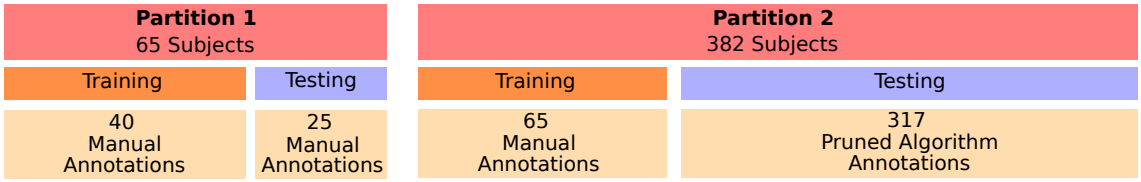

Figure 5.16: Train and test data partitions showing the division of subjects.

For all experiments, we trained a GCN for 100 epochs via the Adam optimizer, using a fixed learning rate of 1×10^{-4}. GCN layers were implemented using the PyTorch Geometric library (version 2.0.1) [9] and trained using a Nvidia RTX 2080 GPU (Nvidia Corporation). All parameters were randomly initialised and train and test data were randomly shuffled. A cross-entropy loss was used to update GCN parameters at each step. As each test partition was balanced in the number of breech and cephalic subjects (*e.g.* partition (1): 25 subjects, 13 breech and 12 cephalic, partition (2): 317 subjects, 156 breech and 161 cephalic), we evaluated model performance using the arithmetic mean classification accuracy on 100 iterations on each test set.

5.4 Classification Accuracy

We report the mean classification accuracy in Table 5.3. For both partitions, the baseline MLP and GCN (with unity edge weights) report similar classification accuracy. Using subject-specific or template edge weights of d_{dF} and d_{tw} does not give an increase in accuracy for either partition. We discuss possible limitations of these curve similarity metrics later and why these results show that they do not strengthen the graph representation over the baseline MLP. For partition (1), using subject-specific edge weights of p_{cc} leads to a striking increase in accuracy of 28%, over the baseline MLP. This suggests the strength in modelling graph edges by computing (Pearson) linear correlation between node feature descriptors, on a subject-by-subject case. For partition (2), using subject-specific edge weights of p_{cc} decreases accuracy by 9%, relative to the baseline MLP. We suggest that this is

Table 5.3: Mean classification accuracy results for GCN with subject-specific and template edge weights, in comparison to a baseline MLP. Bold and underlined entries are the first and second ranks respectively.

Method		**Edge Weights**	**Partition 1 (25 Test Subjects)**	**Partition 2 (317 Test Subjects)**
			Mean Test Accuracy [%]	
MLP			64 ± 4.2	70 ± 2.2
GCN (ours)	Subject-specific	Unity	60 ± 7.3	<u>74 ± 1.6</u>
		d_{dF}	60 ± 6.6	70 ± 1.5
		d_{tw}	64 ± 5.5	71 ± 1.5
		p_{cc}	**92 ± 5.0**	61 ± 2.8
	Template	d_{dF}	64 ± 5.8	71 ± 1.5
		d_{tw}	68 ± 5.5	72 ± 1.7
		p_{cc}	<u>84 ± 6.5</u>	**84 ± 2.4**

related to the quality of the automatic algorithm generated annotations and Pearson correlation modelling assumptions. For both partitions, using template edge weights of p_{cc} leads to a notable increase in accuracy of 20% and 14% respectively, over the baseline MLP. The template edge weights are not seen during training, only at testing. The template edge weights are computed from holdout subjects that are entirely separate of training subjects.

5.5 Discussion

We have proposed a graph-based representation of multiple US video sweeps. An overview of this chapter's contributions are shown in Figure 5.17. We now discuss the significance of our results, in the context of this thesis, and the state-of-the-art.

We performed a statistical characterisation of simple US video sweeps. Specifically, we computed statistical heatmaps for steps A - E for three categories of fetal presentation and provided qualitative interpretation. This analysis shows the average frame-level anatomy detection patterns across simple US video sweeps, using large-scale manual annotations of real-world data. The heatmaps offer insight to describe how fetal pose creates patterns of different magnitude and shape for different video sweep trajectories.

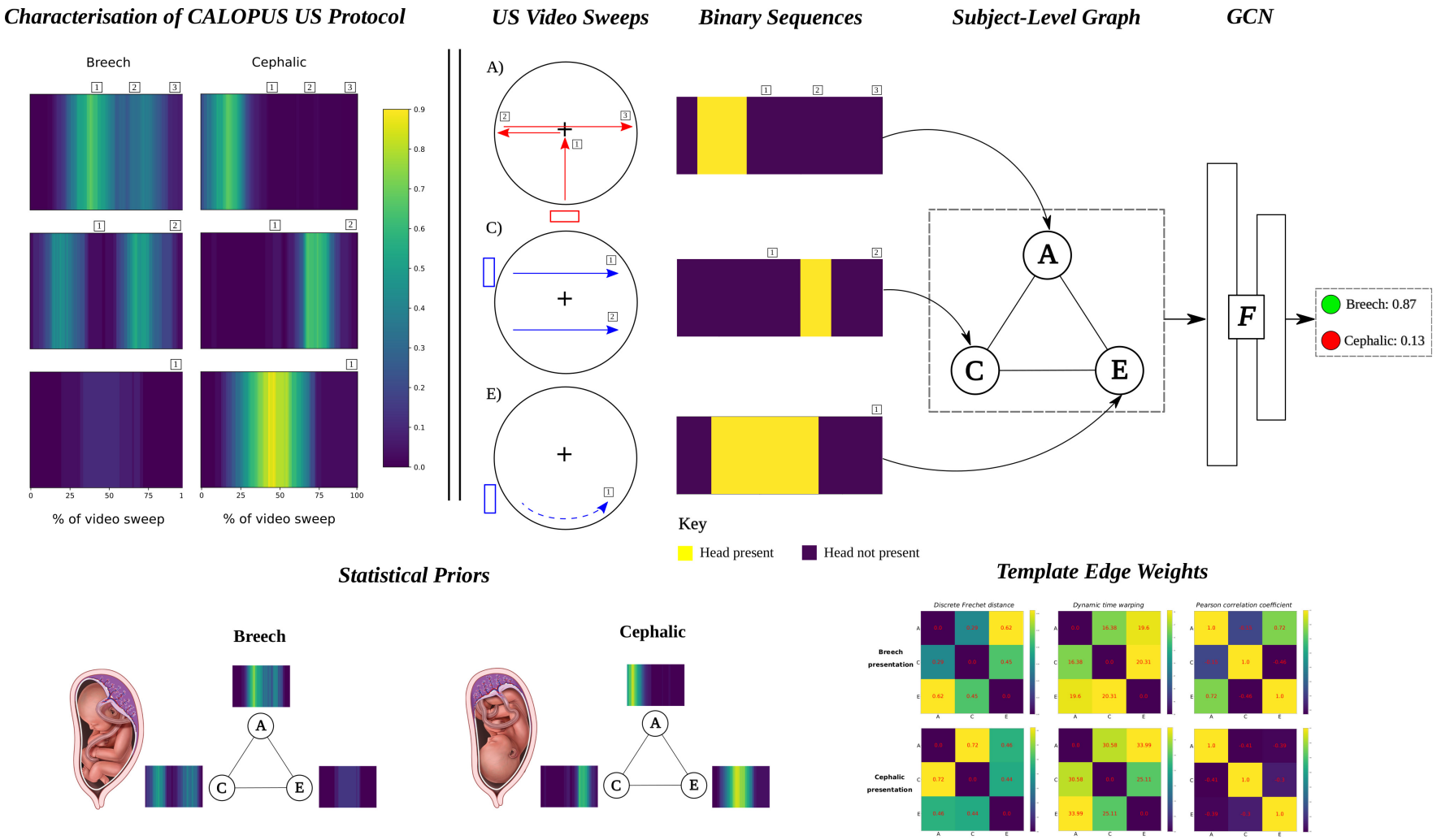

Figure 5.17: Overview of graph-based analysis. We acknowledge illustrations from [5].

Of note, the A, C, E heatmaps in Figure 5.7 highlight interesting differences between breech and cephalic presentation. Comparing breech (L) and cephalic (R) in Figure 5.7, each respective heatmap (*c.f.* A-breech to A-cephalic, *etc.*) differs in both magnitude and shape. This suggests that constructing subject-level graphs using these specific video sweeps can provide valuable information to classify fetal pose. Comparing only breech (L) or only cephalic (R) in Figure 5.7, each heatmap (*c.f.* A-breech to B- or C-breech, *etc.*) also differs in appearance to the remaining heatmaps, again in magnitude and shape. This suggests that subject-level graphs fitted with suitable edge weights are useful in modelling relationships between nodes, in a way which imparts multi-sweep video information relevant to the classification task in this chapter. We suggest that suitable edge weights should model both positive and negative relationships between nodes, as the results here show discriminative patterns of both inclusion and absence of signal in regions along the temporal axis between different nodes (*e.g.* A-breech and C-breech).

Table 5.3 shows that using subject-specific or template edge weights of d_{dF} or d_{tw} does not improve accuracy over the baseline MLP. We suggest that this is because these metrics are unable to model negative relationships, as they only compute positive values. For d_{dF}, this is a single distance value and for d_{tw}, it is a summation of distances describing the warping path. As discussed earlier, we suspect that modelling positive or negative relationships, between video sweeps or nodes, is likely to be important in modelling discriminative graphs of breech and cephalic presentation. In contrast, p_{cc} edge weights can model positive or negative relationships since Pearson linear correlation is a model that can fit either. This is reflected numerically in our results in partition (1), which uses subject-specific edge weights of p_{cc}, giving a striking improvement in classification accuracy of 28% over the baseline. In partition (2), using subject-specific edge weights of p_{cc} decreases classification accuracy by 9%. We suggest that this is because of the quality of the automatic annotations, which differ from manual annotations generated by a trained human. Furthermore,

Pearson linear correlation requires five modelling assumptions. We suspect that two of these assumptions do not hold true for the automatic annotations, which would explain the decrease in accuracy seen in partition (2). Specifically, the node-level distributions of the automatic frame-level annotations are no longer normally distributed and contain significant outliers. This is obvious by inspection of Figures 5.13 and 5.14 which show representative examples of automatic annotations.

Using template edge weights of p_{cc} leads to a notable increase in classification accuracy of 20% and 14% for both partitions respectively. This is a promising result, since these metrics are computed from data independent from any data used during training, and from an independent set of holdout subjects. The template edge weights are not seen during training so it is impossible for the GCN to learn from their representation. Overall, this shows the value of using a template, which in this context, describes relationships between different video sweeps for a particular configuration of fetal pose.

5.6 Conclusion

In conclusion, we have presented a graph-based analysis of multiple US video sweeps. We have characterised the video sweeps to reveal common anatomy detection patterns and used these to create statistical priors describing fetal pose. Our results show promise towards computational understanding of obstetric US video sweep data.

This chapter has considered one approach to modelling relationships between complementary US video sweeps using graphs. Next, we consider an entirely different approach; the challenging task of multi-sweep 3-D reconstruction.

Chapter 6

Placenta 3-D Reconstruction

This chapter describes early work on placenta 3-D reconstruction using multiple US video sweeps. The aim is to register complementary, untracked 2-D US videos to accurately reconstruct 3-D placenta geometries. We pose this challenging problem as spatio-temporal alignment of video. We first explore temporal alignment of video sweeps to represent video content at the same temporal scale. Then, we use affine transformations to spatially align images in temporally aligned video. The results in this chapter demonstrate the feasibility of placenta 3-D reconstruction in an untracked US sweep system.

6.1 Introduction

In §3.5 (*Multi-Sweep Analysis*), we discussed how complementary video sweeps form a sparse 3-D dataset of fetal and maternal anatomy. The research challenge is how to combine video information in the absence of external US transducer tracking (*e.g.* sensors or tracking cameras), which is a commonly taken approach [126, 127]. In our case, knowledge of the CALOPUS US protocol provides constraints. The protocol is visualised on a common space in Figure 6.1 (repeated from earlier in Chapter 3 for brevity). Each step in the protocol, with the exception of step E, follows a linear path or combination thereof, which simplifies this challenging problem.

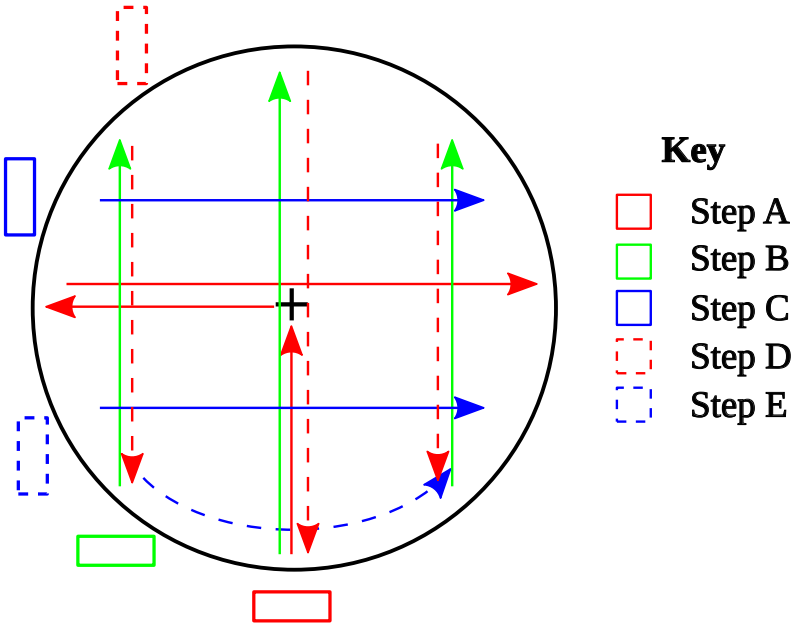

Figure 6.1: CALOPUS US protocol visualised on a common maternal abdomen space. + represents the umbilicus.

Consider step B subdivided into three linear sweeps. Each sweep follows an approximately straight line path beginning at the relative base of the maternal abdomen and sweeping upwards. It is likely that complementary anatomical information is captured as each sweep obtains a different 'viewpoint' of anatomies. Spatio-temporal alignment of the sweeps would achieve US video with extended field-of-view that could visualise large anatomical structures that do not completely fit in the sector width of conventional 2-D US transducers. Large candidate structures are the fetal head or placenta, especially in the third-trimester, when the fetus and placenta have grown. Building on research presented earlier in this thesis, we studied combining three linear sweeps of step B to obtain whole placenta geometries.

The human placenta is an interesting anatomy to study in this problem. At later GA, it does not completely fit into the transducer field-of-view which limits whole placenta imaging beyond the first-trimester. At later trimesters, whole placenta imaging is achieved using magnetic resonance imaging [128], multiple US transducers [37], or manual alignment of placental volumes [129], which limits practicality. The placenta does not move, unlike the fetus, which is a useful constraint in this problem.

The contribution of this chapter is to explore placenta 3-D reconstruction through spatio-temporal alignment of US video sweeps. We first assess video sweeps in the CALOPUS dataset to better characterise the problem.

6.2 Dataset

We used the statistical heatmaps generated previously in §5.2 (*Statistical Analysis of Video Sweeps*) to select the optimal video sweep in the CALOPUS US protocol for placenta 3-D reconstruction. We assessed a combination of maximum probability and qualitative factors of each video sweep.

Figure 6.2 shows statistical heatmaps of the probability of placenta frame-level detection. Calculation of these was identical to the method defined earlier in Chapter 5. A key difference now is that we no longer separate subjects by fetal presentation, as we are interested in the placenta and not the fetus. Figure 6.2 shows that the maximum probabilities are high (> 0.5) for all video sweeps. We ignored steps A and E because of the complex sweep paths (*e.g.* T-shaped and U-shaped respectively). We used qualitative factors to select step B over steps C and D. Step D obtains a sagittal video sweep and is unlikely to sample the placenta volume necessary for 3-D reconstruction. Step C obtains two sweeps compared to three in step B. The placenta is a large anatomical structure, so three sweeps may be required. This point is specifically mentioned by other researchers; for example in liver imaging, three sweeps are required for a conventional 2-D US transducer to capture left, middle, and right lobes [126]. Three video sweeps also provides more constraints for (non-placenta) anatomical correspondence between sweeps. Visual observation of videos confirmed that three sweeps were preferable to two because of greater placenta volume capture.

Three subjects (A, B, C) were selected from the CALOPUS video dataset because of the high quality placenta appearance in video frames of step B. Each subject

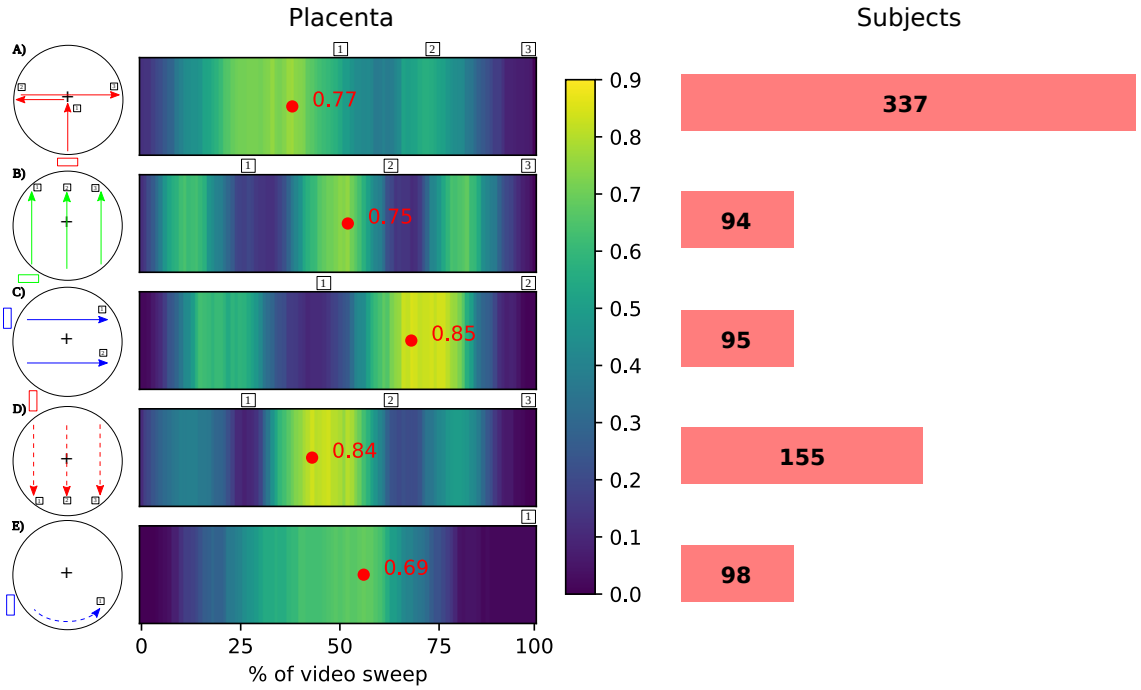

Figure 6.2: Placenta statistical heatmaps showing the probability of placenta frame-level detection. The maximum probability has been labelled as a red dot on each heatmap.

had an anterior placenta. Figure 6.3 shows sample video frames of the three linear sweeps of step B. Each frame contains a cross-sectional view of the placenta. The gestational ages of subjects A, B, C were 22^{+5}, 22^{+0}, 20^{+5} respectively.

In particular, subject C is interesting as the placenta can entirely fit in the middle sweep of step B. In this study, we do not have data from external tracking, so additional constraints may make the problem tractable. At the very least, if the scan speed can be normalised, there is a placenta ground truth volume, as the entire placenta is swept through in the middle sweep. Subjects A and B are different as each sweep only captures a section of the placenta. There is partial correspondence between the left and right sweeps compared to the middle. These subjects are more challenging cases as we do not have the placenta ground truth volume for comparison and correspondences between sweeps may be fewer or be non-obvious. We expand on this point later.

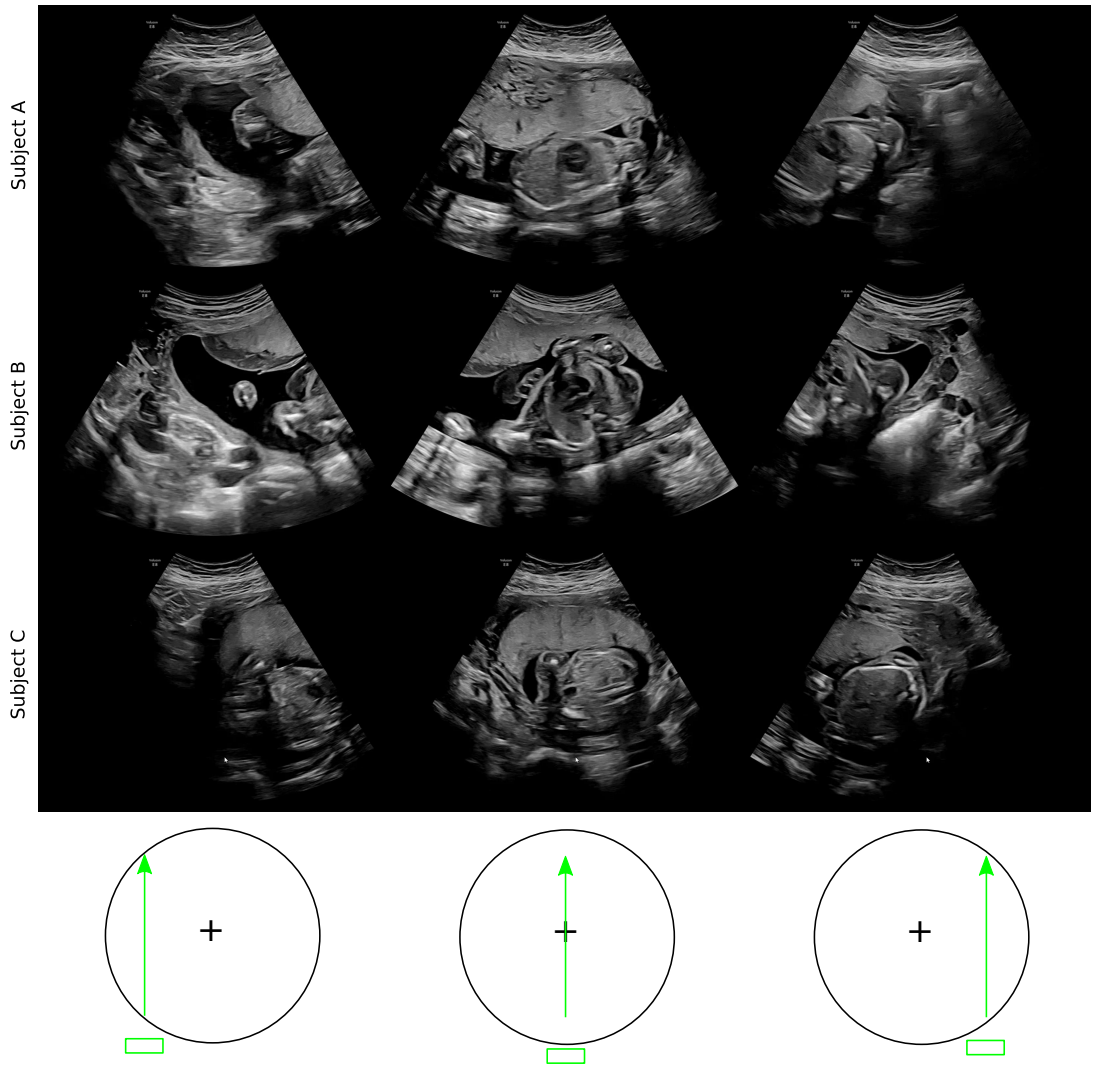

Figure 6.3: Sample video frames of three linear sweeps of step B for subjects A, B, C. Each frame contains a cross-sectional view of the placenta.

6.3 Spatio-Temporal Alignment of Video Sweeps

We posed registration of three linear video sweeps of step B as a spatio-temporal alignment problem. Caspi and Irani [130] investigated spatio-temporal alignment of video sequences of natural scenes, using two stationary cameras of different orientations. That study used a 1-D affine transformation and a homography matrix applied globally to temporally and spatially align sequences respectively. We followed a similar framework but implemented several key approximations relevant for US transducer movement in the context of US video sweeps.

Let $\mathbf{S}$ and $\mathbf{S}'$ be two input video sequences or sweeps, where $\mathbf{S}$ is the reference sequence and $\mathbf{S}'$ is the sequence to be spatio-temporally aligned (*e.g.* fixed and moving respectively). Let $\mathbf{x} = (x, y, t)$ be a space-time point in the reference sequence. We seek to find $\mathbf{x}' = (x', y', t') = (x+u, y+v, t+w)$, the corresponding space-time point in $\mathbf{S}'$. The spatio-temporal displacement is therefore:

$$\mathbf{u} = (u, v, w) \tag{6.1}$$

Where u and v describe the spatial displacement, and w the temporal displacement. The values of $\mathbf{u}$ need not be integers, u and v can be sub-pixel displacements, and w sub-frame displacements. We assume that all space-time points are globally constrained by the following model, where $\mathbf{P}$ is a global transform applied to every video frame in a moving sequence:

$$\mathbf{P} = (\mathbf{P}_{spatial}, \mathbf{P}_{temporal}) \tag{6.2}$$

We used affine transformations to estimate both $\mathbf{P}_{spatial}$ and $\mathbf{P}_{temporal}$ and review the theory next.

6.3.1 Affine Transformations

An affine transformation is a composition of a linear operation, $\mathbf{A}$, and a translation, $\mathbf{b}$. In general, affine transformations are a combination of scales, rotations, translations, dilations, and shears. Collinearity and ratios of distances are preserved. In 2-D, the goal is to find (x', y') by transforming (x, y).

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{A} \begin{bmatrix} x \\ y \end{bmatrix} + \mathbf{b} \tag{6.3}$$

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_{00} \\ b_{10} \end{bmatrix} \tag{6.4}$$

In [130], a 1-D affine transformation was used for temporal alignment: $t' = st + \Delta t$, where s is the ratio of the frame rate of each camera (*e.g.* $\mathbf{S}$ and $\mathbf{S}'$). In our case, the input sequences do not come from stationary cameras with different frame rates, but a moving US transducer with a fixed frame rate. To estimate $\mathbf{P}_{temporal}$, we introduced a temporal approximation inspired by the trajectory of the transducer through a stationary anatomy.

6.3.2 Temporal Alignment

We inspected the placenta frame-level annotations across video sweeps to assess the temporal misalignment. Figure 6.4 shows the placenta frame-level detections across all video frames in linear sweeps of step B. Yellow and purple segments show the placenta presence and absence respectively. White and red numbers show the total placenta frames and the total video frames in a sweep respectively.

Each row of Figure 6.4 shows that the placenta frame-level detections are already reasonably aligned. Specifically, we are comparing the width of each yellow segment on a row-by-row basis (*e.g.* for a single subject). No transformations have been applied yet. The x-axis of each detection plot shows the number of video frames. As each x-axis is visualised with the same length, this is analogous to the percentage of each video sweep, despite a different number of frames obtained. It is interesting to visually compare each linear sweep with this in mind, as it shows that the placenta frame-level detections are already reasonably aligned. The video sweeps are taken without reference to the output video stream, by following the CALOPUS protocol, and no feedback or guidance is given on US transducer speed.

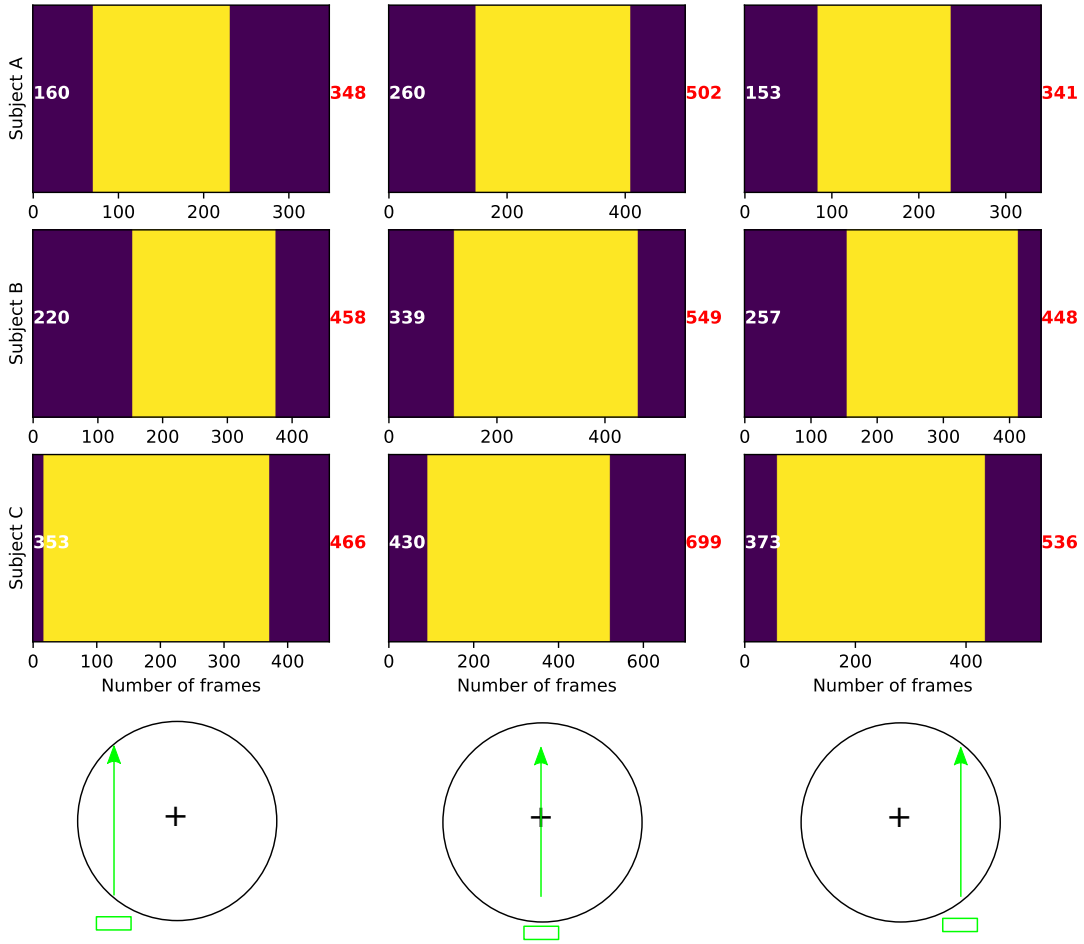

Figure 6.4: Placenta frame-level annotations across three linear sweeps of step B. Yellow and purple segments show the placenta presence and absence respectively. White and red numbers show the total placenta frames and the total video frames in a sweep respectively.

The middle column of Figure 6.4 shows that the total number of video frames of the middle sweep is higher compared to the left and right columns, which show left and right sweeps respectively. This is physically realistic as the US transducer may travel further along the maternal midline than the left and right sides. The midline is the widest part of the maternal abdomen.

To calculate $\mathbf{P}_{temporal}$, we introduced a temporal approximation inspired by the trajectory of the US transducer through a stationary placenta. Although we do not know the starting position of the transducer in each linear sweep, we do know the temporal location where the placenta is first detected. From this, we aligned each

placenta frame-level detection segment using a 1-D affine transformation to ensure that the first and last placenta frame matched. We treated the middle sweep as $\mathbf{S}$ (fixed) and the left and right sweeps separately as $\mathbf{S}'$ (moving). We used a 1-D affine transformation as the placenta frame-level detection segments were already reasonably aligned. A 1-D transformation would be sufficient to fully temporally align the placenta frame-level detection segments and create US video sweeps of the same length.

Recall in [130], $\mathbf{P}_{temporal}$ was computed from $t' = st + \Delta t$. We aligned the first placenta frames of $\mathbf{S}'$ and $\mathbf{S}$, by translating $\mathbf{S}'$ with Δt. Then, given that $\mathbf{S}$ obtains a longer video sweep, we normalised the sampling rate of $\mathbf{S}'$ using a new variable s'.

$$s' = FPS \times \frac{S'_p}{S_p} \tag{6.5}$$

Where S'_p and S_p are the number of placenta frames of $\mathbf{S}'$ and $\mathbf{S}$ respectively, and FPS is the fixed sampling rate of $\mathbf{S}$, the same as the US transducer, 30 FPS. The normalisation of the sampling rate is equivalent to a linear transformation of $s't$. Therefore, temporal alignment is computed as a 1-D affine transformation:

$$t' = s't + \Delta t \tag{6.6}$$

From this temporal approximation, we created temporally aligned video sequences containing only placenta frames. Figures 6.5, 6.6, 6.7 show sample frames from temporally aligned video sequences of subjects A, B, C respectively.

Qualitatively, the temporal alignment succeeded in creating video sequences of the same length where placental anatomy frame content matched, so far as we could observe by eye. We discuss the matching of fetal anatomy later. To complete the spatio-temporal alignment of sweeps, we next computed $\mathbf{P}_{spatial}$.

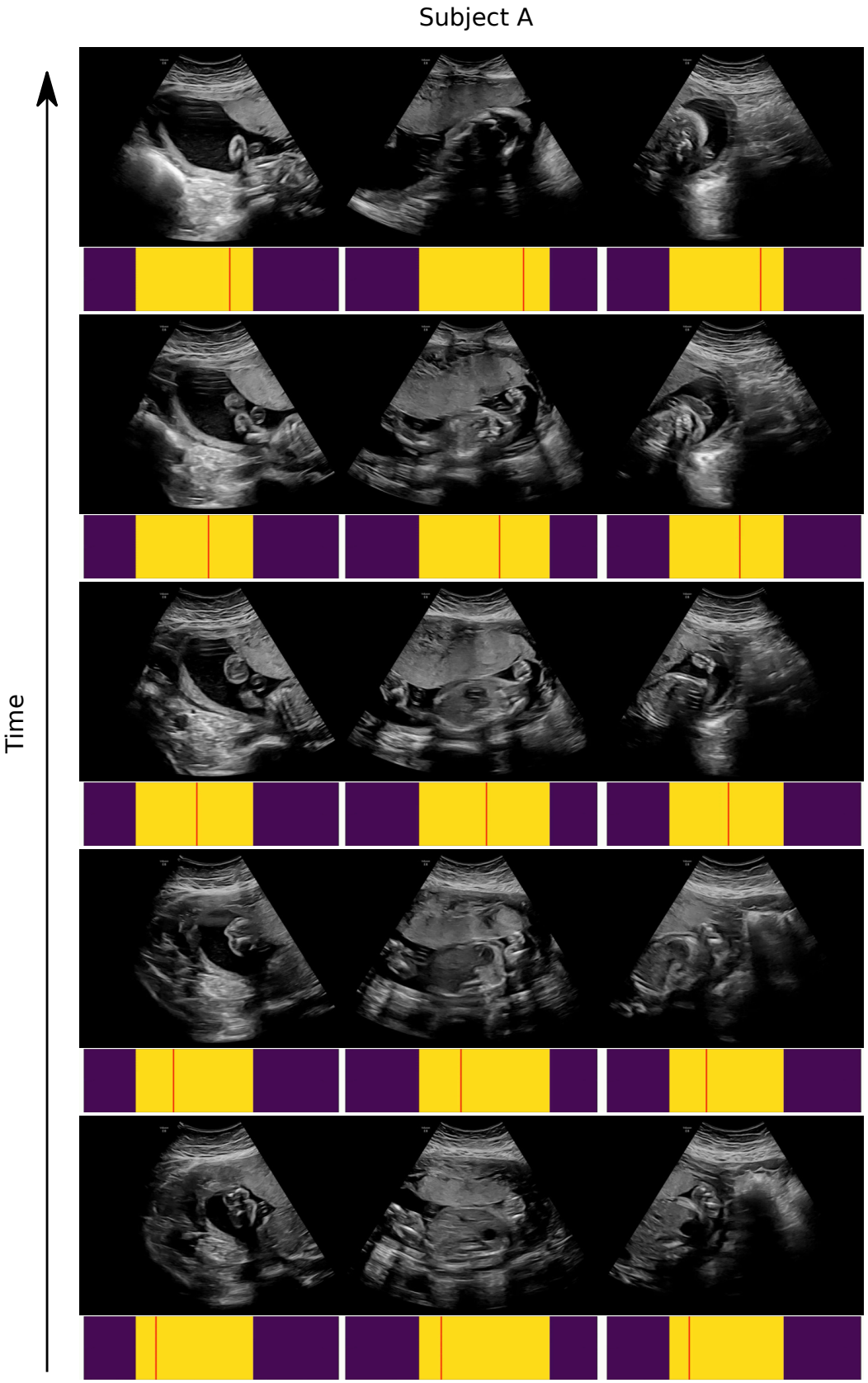

Figure 6.5: Sample temporally aligned frames of subject A. Each frame has a frame position indicator shown by a red line on a yellow and purple detection segment.

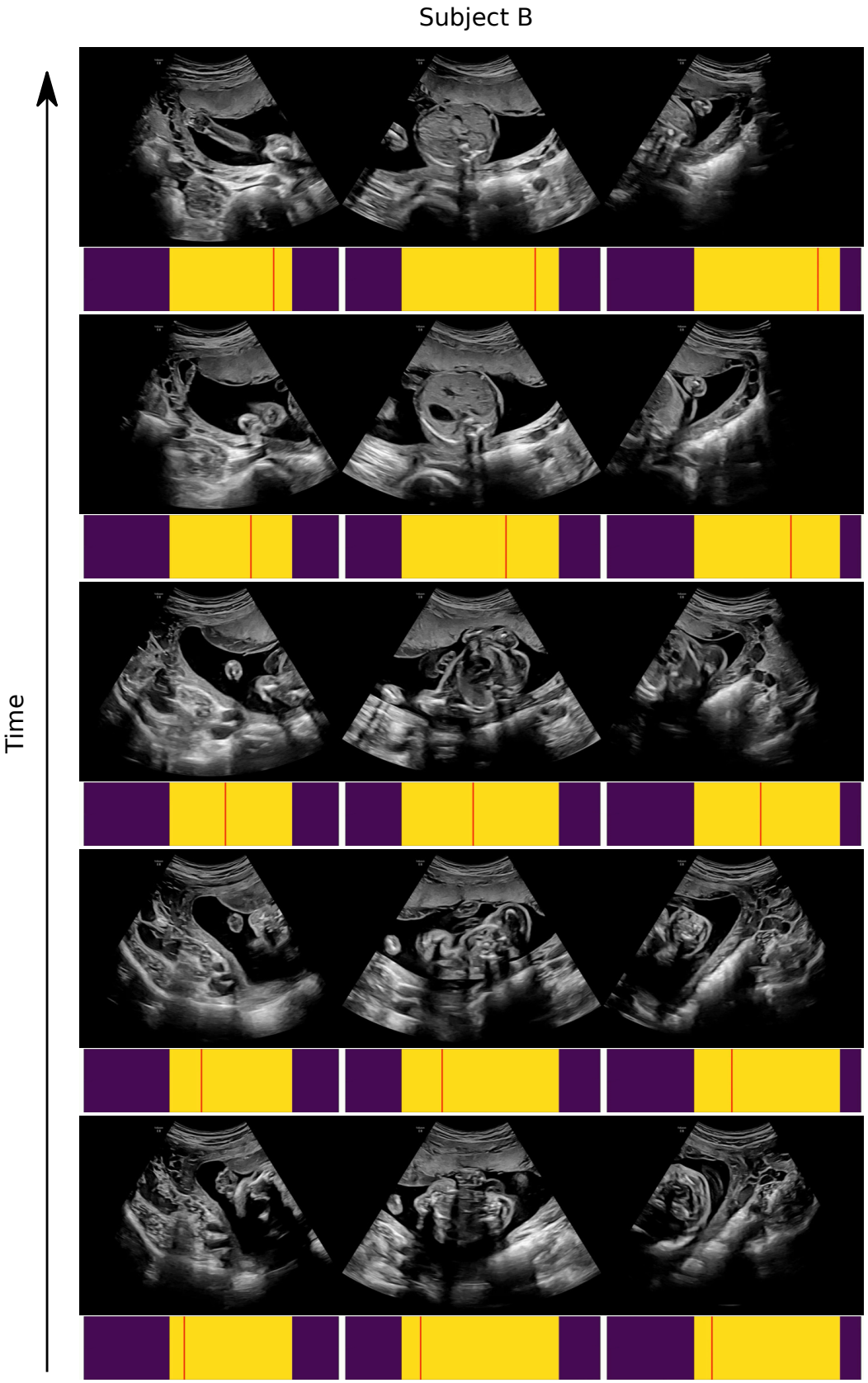

Figure 6.6: Sample temporally aligned frames of subject B. Each frame has a frame position indicator shown by a red line on a yellow and purple detection segment.

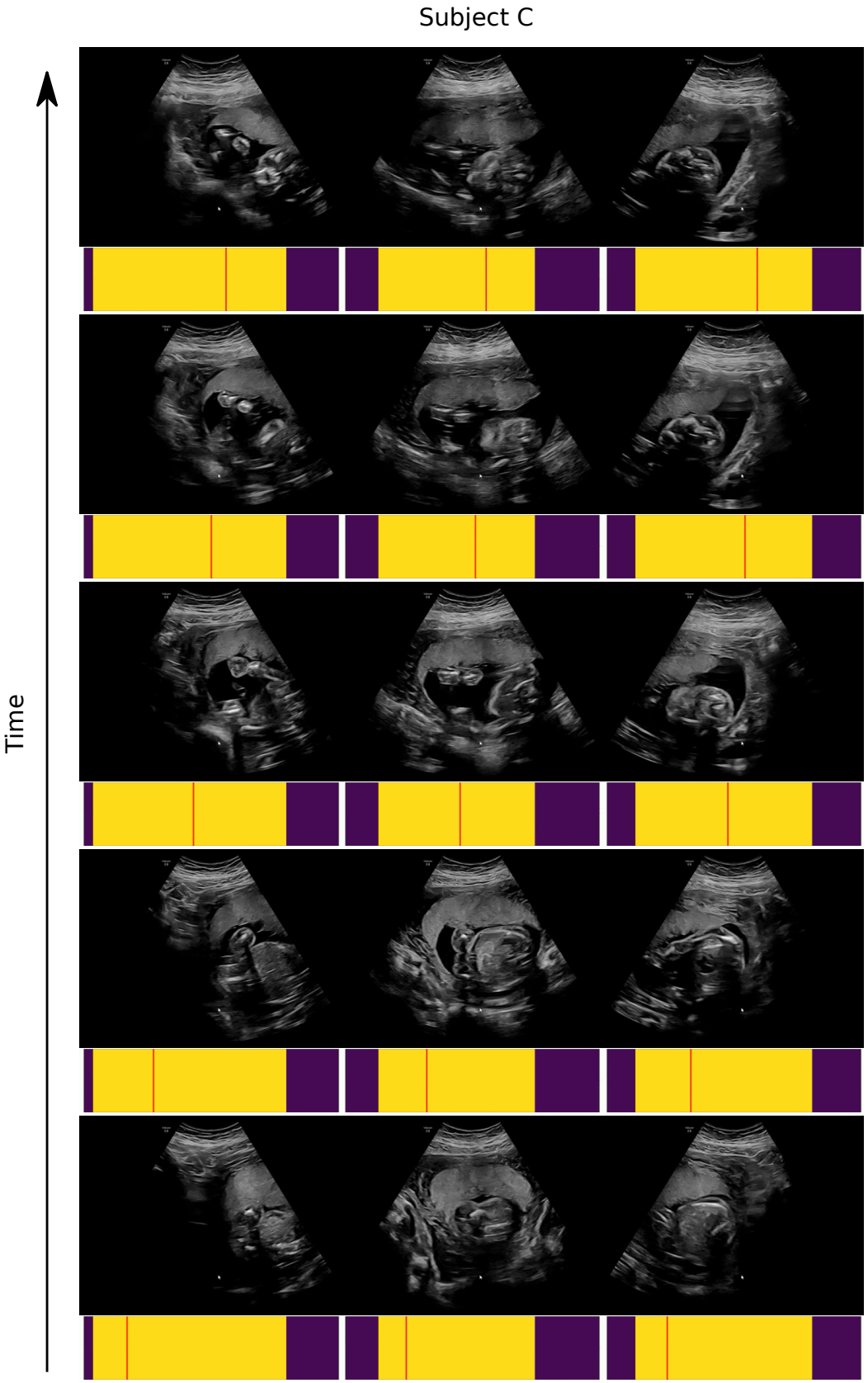

Figure 6.7: Sample temporally aligned frames of subject C. Each frame has a frame position indicator shown by a red line on a yellow and purple detection segment.

6.3.3 Spatial Alignment

We modelled $\mathbf{P}_{spatial}$ as a single 2-D affine transformation applied to all video frames in a sweep, mapping $\mathbf{S}'$ onto $\mathbf{S}$. An affine transformation is physically realistic to map the position of one transducer to another for this type of video sweep, as we are approximating the transducer trajectory as a straight line. A fixed affine transformation also assumes that each transducer moves at a fixed speed in each video sweep. As there are six unknowns in Equation 6.3, two sets of three points matched from an image pair (fixed, moving) are required to find the coefficients of $\mathbf{A}$ and $\mathbf{b}$ by solving the resulting system of linear equations.

We manually identified two frame pairs in temporally aligned sweeps to compute affine transformation coefficients. A frame pair was found by manually searching for temporally aligned frames which had corresponding keypoints easily identifiable by eye. Then, two sets of three points were manually labelled and used to compute affine transformation coefficients. Figures 6.8 and 6.9 show manually labelled points as a white triangle on each frame pair of subject C.

In this task, we did not choose keypoints which had specific anatomical or clinical significance, only that obvious correspondences existed between each frame view of the sweep pair. For example, we found that the left and right placenta edges, large placental arteries, or hyperechoic parts of the fetus were simple to match in each frame pair. Similarly, only subject C was used, as more anatomical information was present in the middle sweep, which could be matched by eye, with the left and right sweeps. As this was an exploratory study, we sought to understand whether a fixed, 2-D global affine transformation could sufficiently map $\mathbf{S}'$ onto $\mathbf{S}$, rather than to automate estimation and subsequent alignment.

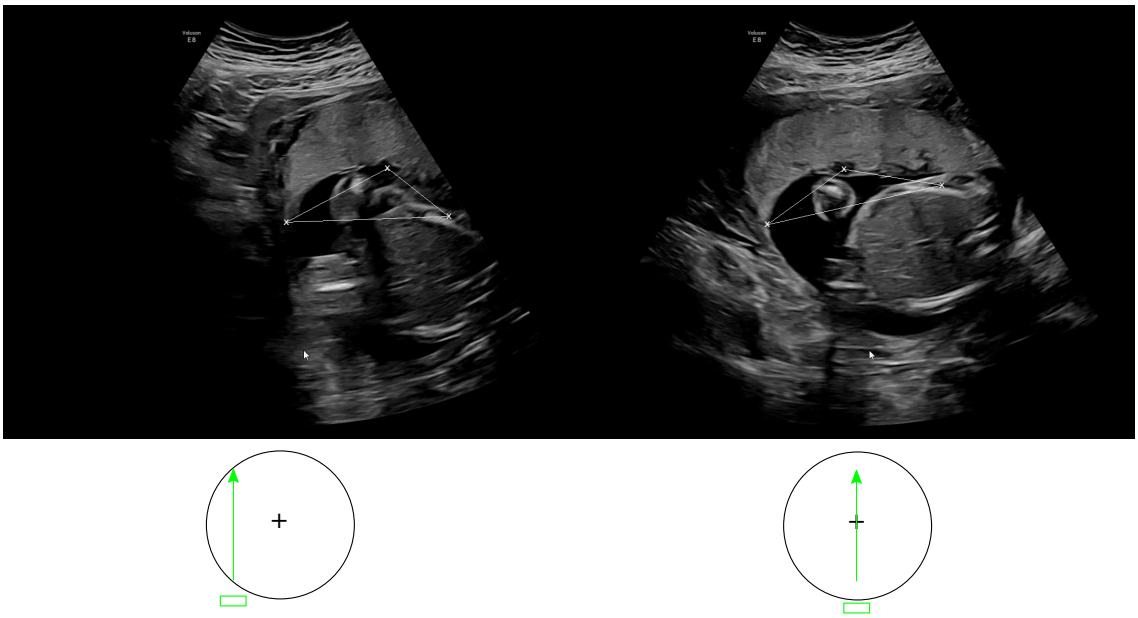

Figure 6.8: Keypoints of subject C frame pair for left and middle sweeps, shown as white crosses in a triangle.

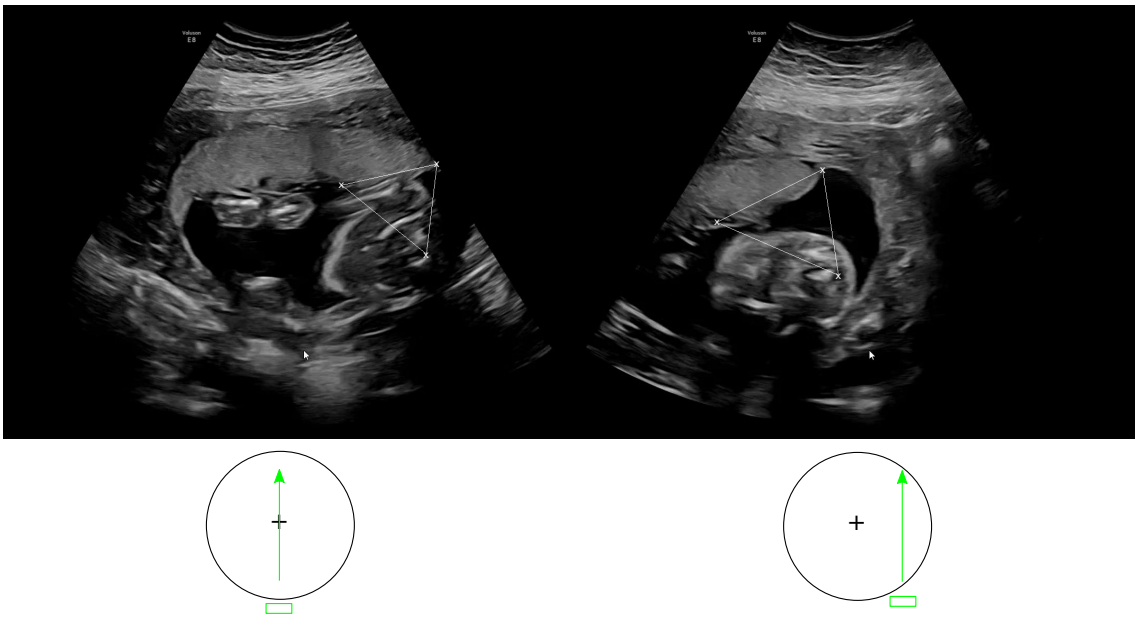

Figure 6.9: Keypoints of subject C frame pair for middle and right sweeps, shown as white crosses in a triangle.

We used the frame pairs and corresponding keypoints to compute two 2-D affine transformations for subject C. The middle sweep was treated as the fixed sequence ($\mathbf{S}$) always, and the left and right sweeps were modelled separately as the moving sequence ($\mathbf{S}'$). We transformed all frames of each moving video sweep, with the respective 2-D global affine transformation, keeping the middle sweep stationary. To create frames which fused multi-view image information, we used a simple rule where the maximum pixel is taken from overlapping pixel values.

6.4 Results

Figure 6.10 shows sample spatio-temporally aligned frames of subject C. Qualitatively, the fused placenta shape and texture appear visually reasonable when compared to the original whole placenta seen in the middle sweep. There are minor inconsistent regions at the placenta tips, shown with red arrows. The outline of the uterus appears reasonable. The height of the left and right US sectors are different. Physically, this could show different US transducer pressures on the maternal abdomen. The alignment of the fetal anatomy is poorer compared to the placenta, shown with a white arrow. Although fetal alignment is not the goal in this study, this is something of interest, as sometimes fetal movement is substantial between video sweeps which makes fetal keypoints unreliable.

Figures 6.11 and 6.12 show volume renderings of spatio-temporally aligned video sweeps of subject C. The placenta was manually segmented to create these visualisations. The volume renderings show visually in 3-D, the volume match between the moving sweeps $\mathbf{S}'$ (red, blue) and the fixed sweep $\mathbf{S}$ (green). Qualitatively, there is reasonable agreement between the moving sweeps to the fixed sweep. Misaligned regions appear mostly at the edges, shown with red arrows, something that was observed in the fused 2-D video frames in Figure 6.10. This is noticeably larger for the right moving sweep (blue) in comparison to the left moving sweep (red).

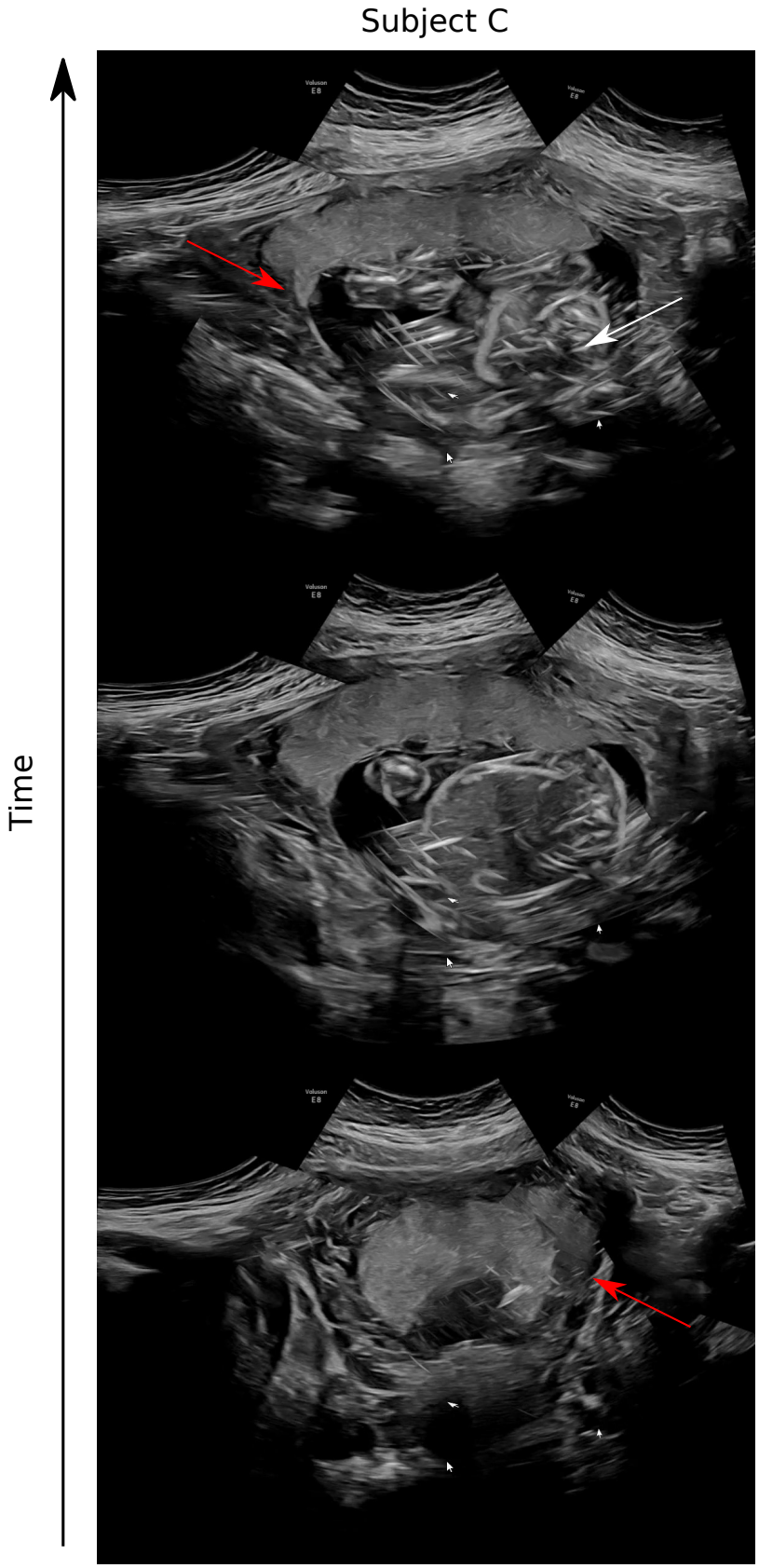

Figure 6.10: Sample spatio-temporally aligned frames of subject C. Red and white arrows show placental and fetal misalignment respectively.

Figure 6.11: Volume rendering of spatio-temporally aligned video sweeps of subject C. Red, green, and blue colours are left, right, and middle sweeps respectively. Red arrows show misaligned regions.

Figure 6.12: Volume rendering of spatio-temporally aligned video sweeps of subject C. Red, green, and blue colour are left, right, and middle sweeps respectively. A volume slice of the middle sweep is shown.

6.5 Discussion

We have described early work on placenta 3-D reconstruction using spatio-temporal alignment to register multiple, untracked 2-D US videos. The results suggest the feasibility of anatomical 3-D reconstruction from 2-D US video obtained by following a simple video sweep protocol. We now discuss the significance of our results in the context of this thesis.

Capitalising on earlier work presented in this thesis, we evaluated placenta statistical heatmaps to select the optimal video sweeps for 3-D reconstruction. Figure 6.2 shows that the maximum probabilities are high (> 0.5) for all video sweeps, an indication of the large placenta size. From this, we selected three linear video sweeps of step B to maximise placenta coverage in video frames.

Following the framework of Caspi and Irani [130], we temporally aligned video sweeps using a 1-D affine transformation of the form $t' = s't + \Delta t$. To find t', we introduced a temporal approximation that the first and last placenta frames of sweeps were temporally identical. Figures 6.5, 6.6, 6.7 show sample temporally aligned frames of subjects A, B, C respectively. Qualitatively, the placental anatomy frame content matched, so far as we could observe by eye. For fetal anatomy, there were frame pairs in temporally aligned sweeps which did not contain the same fetal structures, for example, the heart or stomach bubble, see Figure 6.13. It is possible that the fetus moves between sweeps which would explain this source of error and thus the placenta is unaffected, based on the temporal approximation introduced here. There is a motivating need to better quantify the temporal alignment error, something challenging with this type of dataset, which does not contain position-tracked US video or whole placenta MRI volumes.

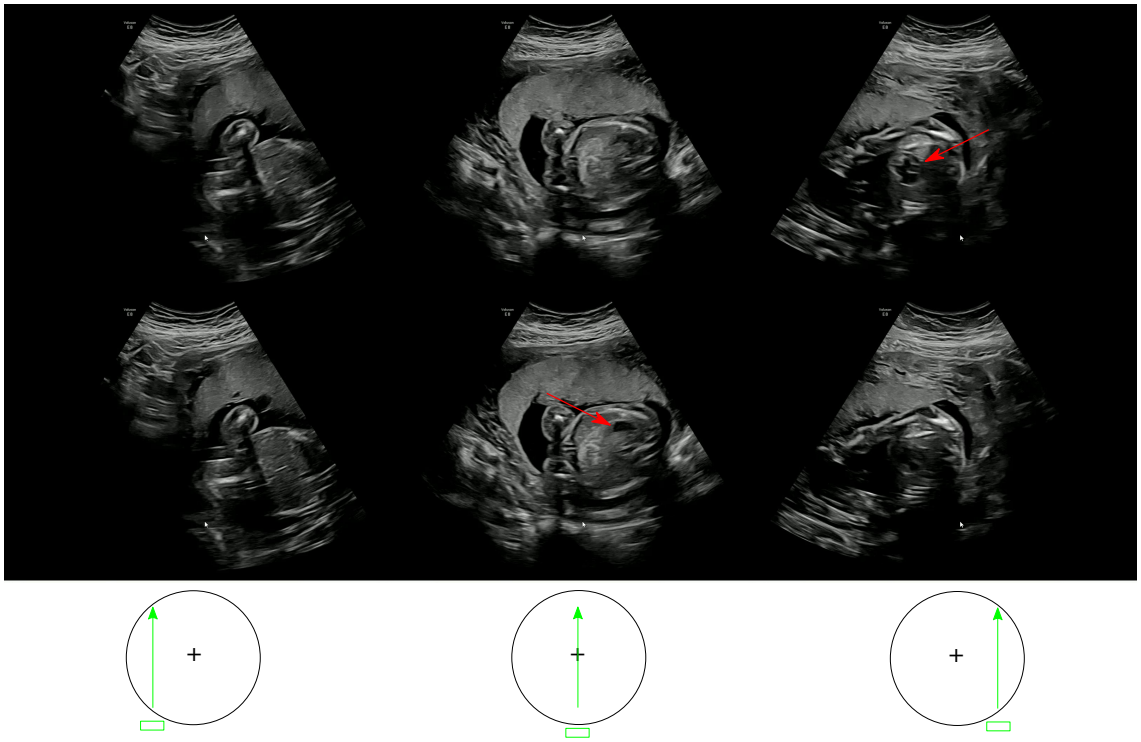

Figure 6.13: Fetal structure mismatch in temporally aligned frames of subject C. Heart (top) and stomach bubble (bottom) are shown with red arrows.

We used 2-D global affine transformations to spatially align moving video sweeps. The affine transformation coefficients were found by manually identifying keypoints in two frame pairs of temporally aligned video frames of subject C, see Figures 6.8 and 6.9. Figure 6.10 shows sample spatio-temporally aligned frames. Fused, multi-view image information is shown by taking the maximum pixel value from overlapping pixels. Qualitatively, this simple rule achieved a visually reasonable combined placenta shape and texture when compared to the original whole placenta seen in the middle sweep. The main errors in spatial alignment occur at the placenta tips, see red arrows in Figure 6.10, something also observed in the volume rendering of Figures 6.11. Interestingly, the US sectors of the left and right sweeps have different heights (*c.f.* left and right sectors in Figure 6.10). It is possible that this shows different US transducer pressures, something modelled by the computed 2-D affine transformation coefficients.

6.6 Conclusion

We have described early work on placenta 3-D reconstruction using spatio-temporal alignment of US video sweeps. Temporal alignment was found by using a temporal approximation and a 1-D affine transformation. Spatial alignment was achieved using 2-D global affine transformations from a set of manually found keypoints, which realistically mapped the position of one US transducer to another.

The results in this chapter show the feasibility of anatomical 3-D reconstruction from untracked US video, a challenging video analysis task. Methods to reconstruct 3-D geometries of human anatomy are invaluable for developing advanced computational representations of anatomical structures, so that we can better understand anatomy structure and function.

Chapter 7

Conclusions

This thesis has developed a series of US image analysis and machine learning-based algorithms that analyse obstetric US video sweeps. A key theme has been to develop new ways to understand video sweep content, which substantially differs from freehand US scans taken by a sonographer. This thesis is a step towards enabling US-inexperienced operators to perform basic pregnancy assessments, by reducing the training gap with algorithms. More broadly, we have contributed to scientific understanding of how to build effective algorithms that understand and combine US video sweeps, using real-world obstetric video data recorded in hospitals.

7.1 Contribution

The contribution of this thesis is threefold:

1. **US image analysis algorithm to aid an inexperienced operator assess placenta location**

   We first explored single video sweep analysis to assess placenta location, an important task in a basic pregnancy risk assessment. We characterised 2-D placenta shapes with a statistical data visualisation algorithm, illustrating the spectrum of shapes in the problem space. We used this knowledge to build an

automatic segmentation algorithm (mDice = 0.83 ± 0.15, $2,127$ test frames) that provides image guidance to an operator, translating the clinical criteria into visually assistive information.

2. **Statistical characterisation and graph-based analysis of multiple US video sweeps**

   Capitalising on knowledge gained in single video sweep analysis, we next explored multi-sweep analysis. We performed a statistical characterisation of different video sweeps in the CALOPUS protocol to find common anatomy detection patterns. From this, we developed a graph-based representation of multiple video sweeps which combines frame-level anatomy detection patterns into a relational model. In particular, we showed how fetal pose can be used as a statistical prior in graph structures to aid machine learning.

3. **Exploratory results demonstrating the feasibility of placenta 3-D reconstruction**

   Last, we explored placenta 3-D reconstruction in the context of multi-sweep analysis. We posed this problem as spatio-temporal alignment of US video. We temporally aligned video using a 1-D affine transformation and a temporal approximation that aligned placenta frame detection segments, using the first and last detected placenta frames. We spatially aligned frames in temporally aligned video using 2-D global affine transformations. We demonstrated the feasibility of placenta 3-D reconstruction using multiple, untracked video sweeps.

7.2 Limitations and Future Work

We next discuss some limitations of our contribution and outline some possible directions of future work.

Automatic Image Guidance for Assessment of Placenta Location

1. **Real-time automatic image guidance**

   The algorithm was designed for offline analysis, where an operator interprets the overlay after obtaining US video. It would be interesting to deploy the algorithm for real-time use. This would require an US transducer and hardware capable of both streaming US video frames in real-time and performing image computations using the algorithm. From a technical perspective, it would be necessary to consider the run-time and memory overhead of the algorithm, to provide real-time, automatic image guidance at the same frame rate of the transducer. Of direct relevance to this thesis, the algorithm could be deployed on a tablet computer which would be suitable for translation to a rural or LMIC environment. Future work would need to explore the necessary steps to port from a desktop workstation to an embedded environment on a tablet computer. From a clinical perspective, real-time automatic image guidance would allow an operator to explore taking either the same U-shaped video sweep (step E) or a freehand US scan.

2. **Point-of-care US video data**

   The video data is obtained using GE Voluson E8 US machines (GE Healthcare). These high-quality machines may not be present in environments such as LMICs. Future work could explore using the algorithm with video data obtained using a point-of-care US device. An example of a suitable, commercially available device is the portable GE Vscan Air (GE Healthcare). As the algorithm was trained on Voluson E8 video data, it would be necessary to evaluate if performance generalises to new video data. Performance could be improved using domain adaptation to learn categorical features relevant to the task which are independent of the US transducer used [131, 132].

Graph-Based Representations of Multiple Ultrasound Sweeps

1. **Graph-based analysis of other relevant clinical tasks**

   Although we have explored automated detection of breech presentation, the graph-based analysis could be applied to other tasks. For example, future work could explore automated labelling of video segments that contain a particular anatomy (*e.g.* fetal heart), using knowledge of the same anatomy in a different video sweep. This could be modelled as node prediction task, where a GCN is trained on a partially-labelled graph to predict missing node feature descriptors. This would be interesting, as it would mimic how a trained sonographer assesses a video sweep and builds understanding of where anatomies are spatially located. The fetal heart is particularly interesting to study because of its small size and the clinical interest in different imaging planes of the heart (*e.g.* four-chamber view, left ventricular outflow tract, and others). Furthermore, a user could initialise a labelled video segment as input, using a trained GCN to label the remaining video sweeps.

2. **Probability density function modelling**

   Currently, the statistical heatmaps model the frame-level probability of the presence of an anatomy. Future work could model heatmaps as probability density functions, meaning for a specific video sweep, all probabilities sum to one. This approach would allow computation of an event (*e.g.* frame-level anatomy detection) falling within a specific time range, rather than a single time-step or frame. Early published work in the group has explored this for automated video quality assessment using a kernel density-based estimation quality metric [133]. By using probability density functions, it would also be possible to model combined distributions with maximum likelihood estimation, something interesting in the context of multiple video sweeps which overlap.

Placenta 3-D Reconstruction

1. **Characterisation of subject pool**

   We used a subset of 3 subjects, selected because of the high-quality placenta appearance in video frames. For the method to be more generalisable, future work should assess the frequency of the placenta appearing across three linear sweeps of step B, such that 3-D reconstruction is possible. This likely goes beyond assessing only statistical heatmaps and would require knowledge of placenta position, subject BMI, fetal GA, video sweep quality, and other clinical factors. More generally, this relates to automated video quality assessment.

2. **Quantitative evaluation of temporal alignment**

   We used a 1-D affine transformation to achieve placenta video frame segments of the same length. Future work could quantitatively evaluate temporal alignment by comparing manually annotated frame-level anatomies across each aligned video sweep. This would assess which labelled frames match after temporal alignment. Future work could also explore more sophisticated temporal alignment by using constraints other than just the placenta detection segment. For example multiple fetal anatomies could be used with multiple 1-D affine transformations to constrain segments of moving video sweeps.

3. **Placenta keypoint detection**

   We used manually identified keypoints on frames to calculate 2-D affine transformations. This was possible because the entire placenta was seen in the middle sweep of step B. In general, it would be more effective to find and use image-based keypoints. This could be deep learning-based [134, 135], US specific [136] or placenta specific. The latter is only currently solved for fetoscopy images [137]. It may also be possible to explore other spatial transformations, applied locally or globally, such as planar homography.

7.3 Summary

This thesis has advanced computational understanding of simple US video sweeps, using medical image analysis and machine learning. The interdisciplinary effort of the CALOPUS project has obtained a rich dataset of real-world obstetric US video sweep data. This has been used throughout this thesis to characterise the problem space and train machine learning-based algorithms. In Chapter 4, we visualised the spectrum of 2-D placenta shapes and developed an automatic image guidance algorithm. In Chapter 5, we performed a graph-based analysis of multiple video sweeps and leveraged fetal pose as a statistical prior. In Chapter 6, we demonstrated the feasibility of placenta 3-D reconstruction using multiple video sweeps. In summary, the research in this thesis is a step towards universal access to obstetric care, to reduce future maternal and fetal mortalities.

Bibliography

[1] G. L. Darmstadt, A. C. C. Lee, S. Cousens, L. Sibley, Z. A. Bhutta, F. Donnay, D. Osrin, A. Bang, V. Kumar, S. N. Wall, et al. "60 million non-facility births: who can deliver in community settings to reduce intrapartum-related deaths?" In: *International Journal of Gynecology and Obstetrics* 107 (2009), S89–S112.

[2] World Health Organisation and others. "Trends in maternal mortality 2000 to 2017". In: *Estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division* (2019).

[3] H. E. Knight, A. Self, and S. H. Kennedy. "Why are women dying when they reach hospital on time? A systematic review of the 'third delay'". In: *PloS one* 8.5 (2013), e63846.

[4] Organisation for Economic Co-operation, Development, and Wellcome Trust. *Low- and middle-income countries defined by the Organisation for Economic Co-operation and Development*. `https://wellcome.org/grant-funding/guidance/low-and-middle-income-countries`. (Accessed on 25/09/2023). 2020.

[5] J. Arroyo, T. J. Marini, A. C. Saavedra, M. Toscano, T. M. Baran, K. Drennan, A. Dozier, Y. T. Zhao, M. Egoavil, L. Tamayo, et al. "No sonographer, no radiologist: New system for automatic prenatal detection of fetal biometry, fetal presentation, and placental location". In: *PloS one* 17.2 (2022), e0262107.

[6] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. "Conditional random fields as recurrent neural networks". In: *International Conference on Computer Vision*. 2015, pp. 1529–1537.

[7] OpenCV. *Open Source Computer Vision Library*. `https://opencv.org/`. (Accessed on 06/04/2023). 2015.

[8] A. Self, Q. Chen, B. K. Desiraju, S. Dhariwal, A. D. Gleed, D. Mishra, R. Thiruvengadam, V. Chandramohan, R. Craik, E. Wilden, et al. "Developing Clinical Artificial Intelligence for Obstetric Ultrasound to Improve Access in Underserved Regions: Protocol for a Computer-Assisted Low-Cost Point-of-Care UltraSound (CALOPUS) Study". In: *JMIR Research Protocols* 11.9 (2022), e37374.

[9] M. Fey and J. E. Lenssen. "Fast Graph Representation Learning with PyTorch Geometric". In: *Representation Learning on Graphs and Manifolds Workshop*. 2019.

[10] P. Ramachandran and G. Varoquaux. *Mayavi: 3D visualization of scientific data*. `https://github.com/enthought/mayavi`. (Accessed on 06/04/2023). 2011.

[11] A. Abuhamad, Y. Zhao, S. Abuhamad, E. Sinkovskaya, R. Rao, C. Kanaan, and L. Platt. "Standardized six-step approach to the performance of the focused basic obstetric ultrasound examination". In: *American Journal of Perinatology* 2.01 (2016), pp. 090–098.

[12] J. Ferrer, T. Chaumont, L. Trujillo, I. Fernandez, J. Guerrero, P. Stewart, G. Garra, M. F. Campos, K. Garra, N. Stephens, et al. "New tele-diagnostic model using volume sweep imaging for rural areas". In: *International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2017, pp. 2622–2625.

[13] K. K. DeStigter, G. E. Morey, B. S. Garra, M. R. Rielly, M. E. Anderson, M. G. Kawooya, A. Matovu, and F. R. Miele. "Low-cost teleradiology for rural ultrasound". In: *Global Humanitarian Technology Conference*. IEEE. 2011, pp. 290–295.

[14] T. L. A. van den Heuvel, H. Petros, S. Santini, C. L. de Korte, and B. van Ginneken. "Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries". In: *Ultrasound in Medicine and Biology* 45.3 (2019), pp. 773–785.

[15] M. A. Maraci, C. P. Bridge, R. Napolitano, A. Papageorghiou, and J. A. Noble. "A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat". In: *Medical Image Analysis* 37 (2017), pp. 22–36.

[16] M. A. Maraci, W. Xie, and J. A. Noble. "Can Dilated Convolutions Capture Ultrasound Video Dynamics?" In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pp. 116–124.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105.

[18] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[19] Y. Gao, M. A. Maraci, and J. A. Noble. "Describing ultrasound video content using deep convolutional neural networks". In: *International Symposium on Biomedical Imaging*. IEEE. 2016, pp. 787–790.

[20] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *International Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 248–255.

[21] A. Self, Q. Chen, J. A. Noble, and A. T. Papageorghiou. "OC10. 03: Computer-assisted low-cost point of care ultrasound: an intelligent image analysis algorithm for diagnosis of malpresentation". In: *Ultrasound in Obstetrics and Gynecology* 56 (2020), pp. 28–28.

[22] T. L. A. van den Heuvel, H. Petros, S. Santini, C. L. de Korte, and B. van Ginneken. "Combining Automated Image Analysis with Obstetric Sweeps for Prenatal Ultrasound Imaging in Developing Countries". In: *Imaging for Patient-Customized Simulations and Systems for Point-of-Care Ultrasound*. Springer, 2017, pp. 105–112.

[23] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv:1409.1556* (2014).

[24] O. Ronneberger, P. Fischer, and T. 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.

[25] T. L. A. van den Heuvel, C. L. de Korte, and B. van Ginneken. "Automated interpretation of prenatal ultrasound using a predefined acquisition protocol in resource-limited countries". In: *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*. 2019.

[26] T. K. Ho. "Random decision forests". In: *International Conference on Document Analysis and Recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[27] R. G. Gomes, B. Vwalika, C. Lee, A. Willis, M. Sieniek, J. T. Price, C. Chen, M. P. Kasaro, J. A. Taylor, E. M. Stringer, et al. "A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment". In: *Communications Medicine* 2.1 (2022), p. 128.

[28] A. C. Saavedra, J. Arroyo, L. Tamayo, M. Egoavil, B. Ramos, and B. Castaneda. "Automatic ultrasound assessment of placenta previa during the third trimester for rural areas". In: *International Ultrasonics Symposium*. IEEE. 2020, pp. 1–4.

[29] M. Schilpzand, C. Neff, J. van Dillen, B. van Ginneken, T. Heskes, C. de Korte, and T. L. A. van den Heuvel. "Automatic Placenta Localization From Ultrasound Imaging in a Resource-Limited Setting Using a Predefined Ultrasound Acquisition Protocol and Deep Learning". In: *Ultrasound in Medicine and Biology* (2022).

[30] J. Arroyo, A. C. Saavedra, L. Tamayo, M. Egoavil, B. Ramos, and B. Castaneda. "Automatic fetal presentation diagnosis from ultrasound images for rural zones: head location as an indicator for fetal presentation". In: *Medical Imaging 2021: Computer-Aided Diagnosis*. Vol. 11597. International Society for Optics and Photonics. 2021, p. 1159718.

[31] M. Prabhudas, E. Bonney, K. Caron, S. Dey, A. Erlebacher, A. Fazleabas, S. Fisher, T. Golos, M. Matzuk, J. M. McCune, et al. "Immune mechanisms at the maternal-fetal interface: perspectives and challenges". In: *Nature Immunology* 16.4 (2015), pp. 328–334.

[32] G. N. Stevenson, S. L. Collins, J. Ding, L. Impey, and J. A. Noble. "3-D ultrasound segmentation of the placenta using the random walker algorithm: reliability and agreement". In: *Ultrasound in Medicine and Biology* 41.12 (2015), pp. 3182–3193.

[33] P. Looney, G. N. Stevenson, K. H. Nicolaides, W. Plasencia, M. Molloholli, S. Natsis, and S. L. Collins. "Automatic 3D ultrasound segmentation of the first trimester placenta using deep learning". In: *International Symposium on Biomedical Imaging*. IEEE. 2017, pp. 279–282.

[34] P. Looney, G. N. Stevenson, K. H. Nicolaides, W. Plasencia, M. Molloholli, S. Natsis, and S. L. Collins. "Fully automated, real-time 3D ultrasound segmentation to estimate first trimester placental volume using deep learning". In: *JCI insight* 3.11 (2018).

[35] R. Hu, R. Singla, R. Yan, C. Mayer, and R. N. Rohling. "Automated Placenta Segmentation with a Convolutional Neural Network Weighted by Acoustic Shadow Detection". In: *International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2019, pp. 6718–6723.

[36] P. Hellier, P. Coupé, X. Morandi, and D. L. Collins. "An automatic geometrical and statistical method to detect acoustic shadows in intraoperative ultrasound brain images". In: *Medical Image Analysis* 14.2 (2010), pp. 195–204.

[37] V. A. Zimmer, A. Gomez, E. Skelton, R. Wright, G. Wheeler, S. Deng, N. Ghavami, K. Lloyd, J. Matthew, B. Kainz, et al. "Placenta segmentation in ultrasound imaging: Addressing sources of uncertainty and limited field-of-view". In: *Medical Image Analysis* 83 (2023), p. 102639.

[38] H. Qi, S. Collins, and J. A. Noble. "Weakly supervised learning of placental ultrasound images with residual networks". In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2017, pp. 98–108.

[39] H. Qi, S. Collins, and J. A. Noble. "Automatic lacunae localization in placental ultrasound images via layer aggregation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 921–929.

[40] H. Qi, S. Collins, and J. A. Noble. "UPI-Net: Semantic Contour Detection in Placental Ultrasound". In: *International Conference on Computer Vision Workshops*. IEEE/CVF, 2019.

[41] S. L. Collins, A. Ashcroft, T. Braun, P. Calda, J. Langhoff-Roos, O. Morel, V. Stefanovic, B. Tutschek, and F. Chantraine. "Proposal for standardized ultrasound descriptors of abnormally invasive placenta (AIP)." In: (2016).

[42] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu. "GCNet: Non-local networks meet squeeze-excitation networks and beyond". In: *International Conference on Computer Vision Workshops*. IEEE/CVF, 2019.

[43] X. Li, X. Hu, and J. Yang. "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks". In: *arXiv:1905.09646* (2019).

[44] R. Prevost, M. Salehi, S. Jagoda, N. Kumar, J. Sprung, A. Ladikos, R. Bauer, O. Zettinig, and W. Wein. "3D freehand ultrasound without external tracking using deep learning". In: *Medical Image Analysis* 48 (2018), pp. 187–202.

[45] J. F. Chen, J. B. Fowlkes, P. L. Carson, and J. M. Rubin. "Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test". In: *International Journal of Imaging Systems and Technology* 8.1 (1997), pp. 38–44.

[46] Q. Li, Z. Shen, Q. Li, D. C. Barratt, T. Dowrick, M. J. Clarkson, T. Vercauteren, and Y. Hu. "Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames". In: *arXiv:2211.04867* (2022).

[47] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Institute for Cognitive Science, 1985.

[48] M. Tan and Q. Le. "EfficientNet: Rethinking model scaling for convolutional neural networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[49] P. H. Yeung, M. Aliasi, A. T. Papageorghiou, M. Haak, W. Xie, and A. I. L. Namburete. "Learning to map 2D ultrasound images into 3D space with minimal human annotation". In: *Medical Image Analysis* 70 (2021), p. 101998.

[50] A. I. L. Namburete, W. Xie, M. Yaqub, A. Zisserman, and J. AA Noble. "Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning". In: *Medical Image Analysis* 46 (2018), pp. 1–14.

[51] B. Hou, A. Alansary, S. McDonagh, A. Davidson, M. Rutherford, J. V. Hajnal, D. Rueckert, B. Glocker, and B. Kainz. "Predicting slice-to-volume transformation in presence of arbitrary subject motion". In: *Medical Image Computing and Computer-Assisted Interventions*. Springer. 2017, pp. 296–304.

[52] P. H. Yeung, L. Hesse, M. Aliasi, M. Haak, W. Xie, A. I. L. Namburete, et al. "ImplicitVol: sensorless 3D ultrasound reconstruction with deep implicit representation". In: *arXiv:2109.12108* (2021).

[53] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[54] M. Luo, X. Yang, H. Wang, H. Dou, X. Hu, Y. Huang, N. Ravikumar, S. Xu, Y. Zhang, Y. Xiong, et al. "RecON: Online learning for sensorless freehand 3D ultrasound reconstruction". In: *Medical Image Analysis* 87 (2023), p. 102810.

[55] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation". In: *IEEE Transactions on Medical Imaging* 37.2 (2017), pp. 384–395.

[56] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, et al. "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography". In: *IEEE Transactions on Medical Imaging* 38.9 (2019), pp. 2198–2210.

[57] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert. "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound". In: *IEEE Transactions on Medical Imaging* 36.11 (2017), pp. 2204–2215.

[58] M. Komatsu, A. Sakai, R. Komatsu, R. Matsuoka, S. Yasutomi, K. Shozu, A. Dozen, H. Machino, H. Hidaka, T. Arakaki, et al. "Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning". In: *Applied Sciences* 11.1 (2021), p. 371.

[59] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". In: *International Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9.

[61] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *International Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[62] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. "Aggregated residual transformations for deep neural networks". In: *arXiv:1611.05431* (2016).

[63] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks". In: *International Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4700–4708.

[64] M. Lin, Q. Chen, and S. Yan. "Network in network". In: *arXiv:1312.4400* (2013).

[65] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. "MobileNets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv:1704.04861* (2017).

[66] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5 MB model size". In: *arXiv:1602.07360* (2016).

[67] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou. "Fixing the train-test resolution discrepancy". In: *Advances in Neural Information Processing Systems* 32 (2019).

[68] R. Ranftl, A. Bochkovskiy, and V. Koltun. "Vision transformers for dense prediction". In: *International Conference on Computer Vision*. IEEE/CVF, 2021, pp. 12179–12188.

[69] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. "Scaling vision transformers". In: *International Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2022, pp. 12104–12113.

[70] Z. Dai, H. Liu, Q. V. Le, and M. Tan. "CoAtNet: Marrying convolution and attention for all data sizes". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3965–3977.

[71] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation". In: *International Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

[72] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. "Attention U-Net: Learning where to look for the pancreas". In: *Medical Imaging with Deep Learning*. 2018.

[73] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre. "Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network". In: *Biomedical Signal Processing and Control* 61 (2020), p. 102027.

[74] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. M. Eslami, D. J. Rezende, and O. Ronneberger. "A probabilistic u-net for segmentation of ambiguous images". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6965–6975.

[75] K. Sohn, H. Lee, and X. Yan. "Learning structured output representation using deep conditional generative models". In: *Advances in Neural Information Processing Systems* 28 (2015).

[76] Z. Zeng, Y. Xulei, Y. Qiyun, Y. Meng, and Z. Le. "SeSe-Net: Self-Supervised deep learning for segmentation". In: *Pattern Recognition Letters* 128 (2019), pp. 23–29.

[77] R. Huang, J. A. Noble, and A. I. L. Namburete. "Omni-supervised learning: scaling up to large unlabelled medical datasets". In: *International Conference on Medical Image Computing and Computer-Assisted Interventions*. Springer. 2018, pp. 572–580.

[78] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask R-CNN". In: *International Conference on Computer Vision*. IEEE/CVF, 2017, pp. 2961–2969.

[79] K. K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. "Deep Extreme Cut: From extreme points to object segmentation". In: *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 616–625.

[80] R. Ke, A. Bugeau, N. Papadakis, M. Kirkland, P. Schuetz, and C. B. Schönlieb. "Multi-task deep learning for image segmentation using recursive approximation tasks". In: *arXiv:2005.13053* (2020).

[81] M. A. Maraci, R. Napolitano, A. Papageorghiou, and J. A. Noble. "Fetal Head Detection on Images from a Low-Cost Portable USB Ultrasound Device". In: *IEEE Transactions on Biomedical Engineering* (2012).

[82] J. Canny. "A computational approach to edge detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1986), pp. 679–698.

[83] I. Sobel and G. Feldman. *An Isotropic 3x3 Image Gradient Operator. Presentation at Stanford AI Project*. 1968.

[84] S. Xie and Z. Tu. "Holistically-nested edge detection". In: *International Conference on Computer Vision*. IEEE/CVF, 2015, pp. 1395–1403.

[85] Z. Yu, C. Feng, M. Y. Liu, and S. Ramalingam. "CaseNet: Deep category-aware semantic edge detection". In: *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5964–5973.

[86] R. Droste, Y. Cai, H. Sharma, P. Chatelain, L. Drukker, A. T. Papageorghiou, and J. A. Noble. "Ultrasound image representation learning by modeling sonographer visual attention". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 592–604.

[87] F. Yu and V. Koltun. "Multi-scale context aggregation by dilated convolutions". In: *arXiv:1511.07122* (2015).

[88] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. "Feature pyramid networks for object detection". In: *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 2117–2125.

[89] T. N. Kipf and M. Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations*. 2017.

[90] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, et al. "A deep learning approach to antibiotic discovery". In: *Cell* 180.4 (2020), pp. 688–702.

[91] L. Backstrom and J. Leskovec. "Supervised random walks: predicting and recommending links in social networks". In: *International Conference on Web Search and Data Mining*. ACM, 2011, pp. 635–644.

[92] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein. "Geometric deep learning on graphs and manifolds using mixture model cnns". In: *International Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2017, pp. 5115–5124.

[93] D. Chen, X. Wu, J. Dong, Y. He, H. Xue, and F. Mao. "Hierarchical Sequence Representation with Graph Network". In: *International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2020, pp. 2288–2292.

[94] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert. "Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease". In: *Medical Image Analysis* 48 (2018), pp. 117–130.

[95] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert. "Metric learning with spectral graph convolutions on brain connectivity networks". In: *NeuroImage* 169 (2018), pp. 431–442.

[96] P. Lu, W. Bai, D. Rueckert, and J. A. Noble. "Multiscale graph convolutional networks for cardiac motion analysis". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2021, pp. 264–272.

[97] Q. Men, C. Teng, L. Drukker, A. T. Papageorghiou, and J. A. Noble. "Multimodal-GuideNet: Gaze-probe bidirectional guidance in obstetric ultrasound scanning". In: *Medical Image Computing and Computer Assisted Interventions*. Springer. 2022, pp. 94–103.

[98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30 (2017).

[99] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An image is worth 16x16 Words: Transformers for image recognition at scale". In: *International Conference on Learning Representations*. 2020.

[100] C. Vondrick, D. Patterson, and D. Ramanan. "Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling". In: *International Journal of Computer Vision* 101 (2013), pp. 184–204.

[101] *GitHub - opencv/cvat: Powerful and efficient Computer Vision Annotation Tool (CVAT)*. `https://github.com/opencv/cvat`. (Accessed on 08/25/2020).

[102] A. Self, S. Dhariwal, S. Tomar, Q. Chen, D. Mishra, V. Chandramohan, B. K. Desiraju, R. Thiruvengadam, S. Bhatnagar, J. A. Noble, et al. "VP34. 14: Machine learning algorithms in ultrasound: Quality assurance metrics for annotations used in training". In: *Ultrasound in Obstetrics and Gynecology* 56 (2020), pp. 199–200.

[103] L. Oppenheimer, A. Armson, D. Farine, L. Keenan-Lindsay, V. Morin, T. Pressey, M. F. Delisle, R. Gagnon, W. R. Mundle, and J. Van Aerde. "Diagnosis and management of placenta previa". In: *Journal of Obstetrics and Gynaecology Canada* 29.3 (2007), pp. 261–266.

[104] E. Jauniaux, S. Collins, and G. J. Burton. "Placenta accreta spectrum: pathophysiology and evidence-based anatomy for prenatal ultrasound imaging". In: *American Journal of Obstetrics and Gynecology* 218.1 (2018), pp. 75–87.

[105] C. Zotti, Z. Luo, A. Lalande, and P. M. Jodoin. "Convolutional neural network with shape prior applied to cardiac MRI segmentation". In: *IEEE Journal of Biomedical and Health Informatics* 23.3 (2018), pp. 1119–1128.

[106] S. M. Abulnaga, E. A. Turk, M. Bessmeltsev, P. E. Grant, J. Solomon, and P. Golland. "Volumetric parameterization of the placenta to a flattened template". In: *IEEE Transactions on Medical Imaging* 41.4 (2021), pp. 925–936.

[107] L. van der Maaten and G. Hinton. "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9.11 (2008).

[108] G. E. Hinton and S. Roweis. "Stochastic neighbor embedding". In: *Advances in Neural Information Processing Systems* 15 (2002).

[109] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[110] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel. "Left-ventricle quantification using residual U-Net". In: *Statistical Atlases and Computational Models of the Heart*. Springer International Publishing. Springer, 2019, pp. 371–380.

[111] P. Krähenbühl and V. Koltun. "Efficient inference in fully connected CRFs with Gaussian edge potentials". In: *Advances in Neural Information Processing Systems*. 2011, pp. 109–117.

[112] A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. 2019, pp. 8024–8035.

[113] M. K. Hu. "Visual pattern recognition by moment invariants". In: *IRE transactions on information theory* 8.2 (1962), pp. 179–187.

[114] A. Reinke, M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina, et al. "Common limitations of image processing metrics: A picture story". In: *arXiv:2104.05642* (2021).

[115] A. Bhide, F. Prefumo, J. Moore, B. Hollis, and B. Thilaganathan. "Placental edge to internal os distance in the late third trimester and mode of delivery in placenta praevia". In: *BJOG: An International Journal of Obstetrics and Gynaecology* 110.9 (2003), pp. 860–864.

[116] W. B. Dawson, M. D. Dumas, W. M. Romano, R. Gagnon, R. J. Gratton, and R. D. Mowbray. "Translabial ultrasonography and placenta previa: does measurement of the os-placenta distance predict outcome?" In: *Journal of Ultrasound in Medicine* 15.6 (1996), pp. 441–446.

[117] M. Gori, G. Monfardini, and F. Scarselli. "A new model for learning in graph domains". In: *International Joint Conference on Neural Networks*. Vol. 2. IEEE. IEEE, 2005, pp. 729–734.

[118] D. K. Hammond, P. Vandergheynst, and R. Gribonval. "Wavelets on graphs via spectral graph theory". In: *Applied and Computational Harmonic Analysis* 30.2 (2011), pp. 129–150.

[119] M. Defferrard, X. Bresson, and P. Vandergheynst. "Convolutional neural networks on graphs with fast localized spectral filtering". In: *Advances in Neural Information Processing Systems* 29 (2016).

[120] T. Eiter and H. Mannila. "Computing discrete Fréchet distance". In: (1994).

[121] H. Alt and M. Godau. "Computing the Fréchet distance between two polygonal curves". In: *International Journal of Computational Geometry and Applications* 5.01n02 (1995), pp. 75–91.

[122] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (1978), pp. 43–49.

[123] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: Computational and Biological Learning Society, 2015, pp. 1–14.

[124] L. Drukker, H. Sharma, R. Droste, M. Alsharid, P. Chatelain, J. A. Noble, and A. T. Papageorghiou. "Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video". In: *Scientific Reports* 11.1 (2021), pp. 1–12.

[125] W. J. Watson, S. Welter, and D. Day. "Antepartum identification of breech presentation". In: *The Journal of Reproductive Medicine* 49.4 (2004), pp. 294–296.

[126] A. H. Gee, G. M. Treece, R. W. Prager, C. J. C. Cash, and L. Berman. "Rapid registration for wide field of view freehand three-dimensional ultrasound". In: *IEEE Transactions on Medical Imaging* 22.11 (2003), pp. 1344–1357.

[127] H. Guo, H. Chao, S. Xu, B. J. Wood, J. Wang, and P. Yan. "Ultrasound volume reconstruction from freehand scans without tracking". In: *IEEE Transactions on Biomedical Engineering* (2022).

[128] R. Aughwane, E. Ingram, E. D. Johnstone, L. J. Salomon, A. L. David, and A. Melbourne. "Placental MRI and its application to fetal intervention". In: *Prenatal Diagnosis* 40.1 (2020), pp. 38–48.

[129] W. Cheung, G. N. Stevenson, A. E. G. de Melo Tavares, J. Alphonse, A. W. Welsh, et al. "Feasibility of image registration and fusion for evaluation of structure and perfusion of the entire second trimester placenta by three-dimensional power doppler ultrasound". In: *Placenta* 94 (2020), pp. 13–19.

[130] Y. Caspi and M. Irani. "Spatio-temporal alignment of sequences". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.11 (2002), pp. 1409–1424.

[131] Q. Chen, Y. Liu, Y. Hu, A. Self, A. T. Papageorghiou, and J. A. Noble. "Cross-device cross-anatomy adaptation network for ultrasound video analysis". In: *Advances in Simplifying Medical Ultrasound*. Springer International Publishing. Springer, 2020, pp. 42–51.

[132] Q. Meng, J. Matthew, V. A. Zimmer, A. Gomez, D. F. A. Lloyd, D. Rueckert, and B. Kainz. "Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging". In: *IEEE Transactions on Medical Imaging* 40.2 (2020), pp. 722–734.

[133] J. Kwon, J. Jiao, A. Self, J. A. Noble, and A. Papageorghiou. "A kernel density estimation based quality metric for quality assessment of obstetric ultrasound video". In: *Trustworthy Machine Learning for Healthcare Workshop*. ICLR.

[134] D. DeTone, T. Malisiewicz, and A. Rabinovich. "SuperPoint: Self-supervised interest point detection and description". In: *International Conference on Computer Vision and Pattern Recognition Workshops*. IEEE/CVF, 2018, pp. 224–236.

[135] P. E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. "SuperGlue: Learning feature matching with graph neural networks". In: *International Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2020, pp. 4938–4947.

[136] C. Zhao, R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble. "US-Point: Self-supervised interest point detection and description for ultrasound-probe motion estimation during fine-adjustment standard fetal plane finding". In: *Medical Image Computing and Computer Assisted Interventions*. Springer. 2022, pp. 104–114.

[137] S. Bano, F. Vasconcelos, L. M. Shepherd, E. Vander Poorten, T. Vercauteren, S. Ourselin, A. L. David, J. Deprest, and D. Stoyanov. "Deep placental vessel segmentation for fetoscopic mosaicking". In: *Medical Image Computing and Computer Assisted Interventions*. Springer. 2020, pp. 763–773.