Deep Learning Sonographer Visual Attention

Yifan Cai

Keble College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Hilary 2019

Abstract

Current automated fetal ultrasound (US) analysis methods are heavily influenced by the recent success of deep learning in computer vision tasks. Models built on convolutional neural networks (CNN) for fetal biometry planes detection have surpassed classic models built on hand-crafted features, but training such networks requires large dataset, especially sonographer annotations, which is normally not available in US image analysis. Meanwhile, sonographer visual attention has proven to be a strong prior for human interpretation of US video frames. This thesis attempts to utilize sonograher visual attention in the form of gaze-tracking data in deep learning frameworks to assist US image analysis tasks.

We created a single sweep dataset on fetal abdominal videos with retrospective gaze-tracking, then implemented deep learning frameworks that utilize gaze-tracking data to assist fetal biometry plane detection. We first developed a CNN called *SonoEyeNet* for standardized abdominal circumference plane (ACP) detection informed by sonographer visual attention. We demonstrate that with the assistance of human visual attention information, ACP detection performance is increased compared to models not using gaze information.

We extended this framework by proposing a novel multi-task CNN called *Multi-task SonoEyeNet (MSEN)* that learns to generate clinically relevant spatial visual attention maps using sonographer gaze tracking data, and used the predicted visual attention maps to assist ACP detection. This framework expands the potential clinical usefulness of the previous framework by eliminating the requirement of input gaze-tracking data during inference without compromising its ACP detection performance.

With the availability of a novel dataset containing real-time screen recordings of US anomaly scans coupled with simultaneous gaze-tracking, we further extended the CNN framework by introducing a bi-directional convolutional long-short term memory (LSTM) as a recurrent module to model spatio-temporal visual attention as well as to detect all standard biometry planes of fetal abdomen (ACP), head (HCP) and femur (FLP). It was demonstrated that by modeling spatio-temporal visual attention, standard biometry planes detection performance can be further improved.

This work constitutes the first demonstration that learning sonographer visual attention in an ultrasound video in a deep learning framework is an efficient method to assist other US image analysis tasks.

Deep Learning Sonographer Visual Attention

Yifan Cai

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2019

Acknowledgements

I will begin by expressing my utmost gratitude towards my supervisor, Professor Alison Noble, for her guidance over the years. Throughout my many periods of self-doubt she has remained supportive and understanding, and it has been nothing short of a privilege to work with her.

I must also thank my friends Yixing Wu, Yang Cao, Xue Min, Xiaochun Meng, Haoyu Wu, for distracting me from my work and providing me with a sense of sanity. I am grateful to my lab mates Huan Qi, Yuan Gao, Ruobing Huang, for their support and encouragement.

I am deeply grateful to my parents, Guiqiu Bian and Hongping Cai, for encouraging me to pursue my passions and for their steadfast belief in me. I would not be who I am without their unwavering support and unconditional love.

I want to finally thank my wife, Liaoliao Zhang. This thesis cannot be done, without her support and understanding, whether separated by thousands of miles or together under the same roof. Per aspera ad astra.

Abstract

Current automated fetal ultrasound (US) analysis methods are heavily influenced by the recent success of deep learning in computer vision tasks. Models built on convolutional neural networks (CNN) for fetal biometry planes detection have surpassed classic models built on hand-crafted features, but training such networks requires large dataset, especially sonographer annotations, which is normally not available in US image analysis. Meanwhile, sonographer visual attention has proven to be a strong prior for human interpretation of US video frames. This thesis attempts to utilize sonograher visual attention in the form of gaze-tracking data in deep learning frameworks to assist US image analysis tasks.

We created a single sweep dataset on fetal abdominal videos with retrospective gaze-tracking, then implemented deep learning frameworks that utilize gaze-tracking data to assist fetal biometry plane detection. We first developed a CNN called *SonoEyeNet* for standardized abdominal circumference plane (ACP) detection informed by sonographer visual attention. We demonstrate that with the assistance of human visual attention information, ACP detection performance is increased compared to models not using gaze information.

We extended this framework by proposing a novel multi-task CNN called *Multi-task SonoEyeNet (MSEN)* that learns to generate clinically relevant spatial visual attention maps using sonographer gaze tracking data, and used the predicted visual attention maps to assist ACP detection. This framework expands the potential clinical usefulness of the previous framework by eliminating the requirement of input gaze-tracking data during inference without compromising its ACP detection performance.

With the availability of a novel dataset containing real-time screen recordings of US anomaly scans coupled with simultaneous gaze-tracking, we further extended the CNN framework by introducing a bi-directional convolutional long-short term memory (LSTM) as a recurrent module to model spatio-temporal visual attention as well as to detect all standard biometry planes of fetal abdomen (ACP), head (HCP) and femur (FLP). It was demonstrated that by modeling spatio-temporal visual attention, standard biometry planes detection performance can be further improved.

This work constitutes the first demonstration that learning sonographer visual attention in an ultrasound video in a deep learning framework is an efficient method to assist other US image analysis tasks.

Contents

List of Figures

List of Tables

Glossary

1-D, 2-D, 3-D	One- , two- , or three-dimensional
US	ultrasound
IUGR	Intra-Uterine Growth Restriction
FASP	Fetal Anomaly Screening Programme
ISUOG	International Society for Ultrasound in Obstetrics and Gynaecology
AC, ACP	Abdominal Circumference, Abdominal Circumference Plane
HC, HCP	Head Circumference, Head Circumference Plane
FL, FLP	Femur Length, Femur Length Plane
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GAN	Generative Adversarial Networks
LSTM, convLSTM	Long-Short Term Memory, Convolutional Long-Short Term Memory
GRU, convGRU	Gated Recurrent Unit, Convolutional Gated Recurrent Unit
SEN, MSEN	SonoEyeNet, Multi-task SonoEyeNet
T-SEN	Temporal SonoEyeNet
TAM, VCM	Temporal Attention Module and Video Classification Module, two recurrent modules in T-SEN
SGD	Stochastic Gradient Descent
FIT	Feature Integration Theory
KLD	Kullback-Leibler divergence
MSE	Mean Squared Error
CE, BCE	Cross-Entropy, Binary Cross-Entropy
DTW, sDTW	Dynamic Time Warping, soft-Dynamic Time Warping
CPU, GPU	Central Processing Unit, Graphical Processing Unit

Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit...

There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain...

— Cicero's *de Finibus Bonorum et Malorum*

1 Introduction

Contents

1.1 Clinical Motivation

An estimated 7.5 million babies die during the perinatal period (deaths occured after 22 weeks' gestation and within the first 7 days after birth) worldwide [**Yu2003**]. Research has shown that early detection of Intra-Uterine Growth Restriction (IUGR) and subsequent early prevention of low birth weight are crucial in the reduction of perinatal mortalities [**bernstein2000morbidity**]. 2-D B-mode Ultrasound (US) is the most common imaging modality for fetal growth monitoring as well as screening for potential anomalies during fetal development [**haram2006intrauterine**].

In the UK Fetal Anomaly Screening Programme (FASP) [**kirwan2010nhs**], pregnant women receive two or three ultrasound scans, from which biometric measurements are obtained from standardised imaging planes defined by the FASP protocol. However, challenges remain for the detection of those standardised imaging

planes. First, ultrasound image appearance is highly variably due to areas of poor contrast, acoustic shadow, attenuations and other artifacts. Second, due to possible different positions of the fetus, orientations and positions of the key anatomical features can also be variable. Third, ultrasound travels through a larger amount of adipose and other soft tissues with increasing gestational age, which reduces the quality of ultrasound images. Finally, maneuvering the ultrasound probe to acquire scans of different anatomical structure is a complex task and requires lengthy training, thus inter-operator and intra-operator error contributes greatly to the variation in ultrasound image quality [**Sarris2012**].

Because of these challenges, automated fetal US analysis algorithms using machine learning methods have been investigated in research settings. Until the relatively recent introduction of deep learning, models used hand-crafted features for classification, segmentation, registration and detection tasks in 2-D US images or 2-D+t US videos. Such an approach is traditionally referred to as a "bottom-up" approach as the hand-crafted features are local descriptors of the image. Early attempts to combine human gaze information as a source of "top-down" information that are specific to a certain task or context with "bottom-up" features were made. However, applicability of the model was constrained due to limited speed (not able to take advantage of computational parallelisation capabilities of GPUs) and performance (with mean classification accuracy of fetal head, abdomen, and others of 79.19%) [**AhmedThesis**]. Meanwhile, as impressive progress has been achieved by deep learning methods in computer vision fields, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have started to be used for US image and video analysis. Therefore, obvious research questions emerge: can sonographer visual attention as a source of "top-down" information be incorporated in an end-to-end trainable deep learning model for US image and video analysis? And will a model with humn aknowledge embedded in it improve performance?

This thesis attempts to answer these questions. Specifically, gaze-tracking experiments were designed and conducted to collect sonographers' gaze data; deep models for incorporation of sonographer gaze information in CNN, learning

sonographer spatial visual attention, and an extension to spatio-temporal visual attention modelling were investigated so as to produce efficient models for standard plane detection on 2-D US images as well as 2-D+t US videos.

1.2 Contributions

This thesis contributes a series of efficient deep learning methods for standard biometry planes detection in fetal 2-D US images and 2-D+t US videos as well as several methods to model sonographer visual attentions.

The first contribution is a novel automated approach for detection of standard abdominal circumference (AC) planes in fetal ultrasound built in a CNN framework, called SonoEyeNet, that utilizes gaze-tracking data of a sonographer in automatic interpretation. Gaze-tracking data was collected from experienced sonographers (3 to 10 years of experience) as they identified an AC plane in fetal ultrasound video clips. A visual heatmap was generated from the gaze points for each video frame. A CNN model was built using ultrasound frames and their corresponding visual heatmaps. Different methods of processing visual heatmaps and their fusion with image feature maps were investigated. It is shown that with the assistance of human visual fixation information, the precision, recall and F1-score of AC plane detection was increased to 96.5%, 99.0% and 97.8% respectively, compared to 73.6%, 74.1% and 73.8% without using gaze information.

The second contribution is a novel multi-task convolutional neural network called Multi-task SonoEyeNet (M-SEN) that learns to generate clinically relevant visual attention maps using sonographer gaze tracking data on input ultrasound (US) video frames so as to assist standard abdominal circumference (AC) plane detection. The architecture consists of a generator and a discriminator, which are trained in an adversarial scheme. The generator learns sonographer attention on a given US video frame to predict the frame label (standard AC plane / background). The discriminator further fine-tunes the predicted attention map by encouraging it to mimick the ground-truth sonographer attention map. The novel model expands the potential clinical usefulness of SonoEyeNet by eliminating the requirement of

input gaze tracking data during inference without compromising its plane detection performance (Precision: 96.8, Recall: 96.2, F-1 score: 96.5).

The third contribution of this thesis is a novel network that learns sonographers' spatio-temporal visual attention on input US video sequences to assist the detection of standard biometry planes of fetal abdomen, head and femur. The proposed architecture consists of two modules: a Temporal Attention Module (TAM) that predicts Dynamic Attention Maps, and a Video Classification Module (VCM) for video frames classification, both of which utilizes bi-directional convolutional LSTM to encode spatio-temporal information from input sequence. The soft Dynamic Time Warping (sDTW) loss was found to be an excellent loss function to regularise temporal visual attention synchronisation between ground-truth and predicted visual attention maps, and focal loss an effect loss function for frame classification. The best performing model out-performs MSEN in both static saliency scores and as well as scanpath similarity scores, and improves F1-scores for ACP, HCP and FLP to 83.7%, 89.9% and 81.1%, compared 68.3%, 68.1% and 60.0% from MSEN.

Apart from the contribution in algorithms, a dataset was collected with 8 sonographers' visual attention on 33 US videos (1616 frames) of the fetal abdomen where they used a keyboard to select standard AC planes. This dataset has been used in a publication by other researchers [**droste2019ultrasound**].

1.3 Thesis Structure

Chapter 1 outlines the clinical motivation for the doctoral research to follow, the structure of this thesis, and the original contributions and publications related to the thesis.

Chapter 2 first describes the importance of US imaging in fetal growth monitoring, the challenges associated with it and traditional automated US image analysis methods. Then it details developments achieved by deep learning methods in recent years with regard to both computer vision and medical image analysis, with a focus on US image analysis. Finally, a review of visual attention modelling methods

is provided with a focus on deep learning inspired algorithms; its application in computer vision and medical image analysis is also discussed.

Chapter 3 describes two US video datasets used in this thesis.

Chapter 4-6 describe the original algorithm contributions of this thesis. Specifically:

Chapter 4 presents a noval automated approach for detection of standard abdominal circumference (AC) planes in fetal ultrasound built in a CNN framework, called SonoEyeNet (SEN), that utilizes gaze-tracking data of a sonographer in automatic interpretation. Gaze tracking experiments as well as data processing procedures are described. Basic theories for convolutional neural networks as well as performance metrics used are described. Visual attention maps based on gaze data were generated and used for standardised AC plane detection, and results presented show better performance than state of the art at that time [**baumgartner2017sononet**].

Chapter 5 presents a multi-task convolutional neural network called Multi-task SonoEyeNet (M-SEN) that learns to generate clinically relevant visual attention maps using sonographer gaze tracking data on input ultrasound (US) video frames so as to assist standard abdominal circumference (AC) plane detection. It begins with an early exploration of patch-wise saliency prediction. Then it describes a CNN model that is trained using a multi-task learning objective by simultaneously predicting sonographer visual attention as well as input image classification. Finally, an adversarial regulariser in the form of a discriminator network is introduced to further fine-tune predicted visual attention maps. This novel model expands the potential clinical usefulness of the model from Chapter 4 by eliminating the requirement of input gaze tracking data during inference without compromising its plane detection performance.

Chapter 6 presents a novel network that learns sonographers' spatio-temporal visual attention on input US video sequences to assist the detection of standard biom-

etry planes of fetal abdomen, head and femur. The chapter begins by introducing key concepts regarding Recurrent Neural Networks (RNNs), especially convolutional RNNs, that are important for modelling spatio-temporal information. The proposed architecture consists of two modules: a Temporal Attention Module (TAM) that predicts Dynamic Attention Maps, and a Video Classification Module (VCM) for video frames classification, both of which utilizes bi-directional convolutional LSTM to encode spatio-temporal information from input sequence. It demonstrates the usefulness of soft Dynamic Time Warping (sDTW) loss as a temporal regularizer, and focal loss as an effective loss function for frame classification.

Chapter 7 summarises the main contributions of this thesis and discusses potential areas for future work.

1.4 Peer Reviewed Publication

Chapter 4

Yifan Cai, Harshita Sharma, Pierre Chatelain, and J. Alison Noble, 2018, April. "SonoEyeNet: standard fetal ultrasound plane detection informed by eye tracking." In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018).*

Chapter 5

Yifan Cai, Harshita Sharma, Pierre Chatelain, and J. Alison Noble, 2018, September. "Multi-task SonoEyeNet: Detection of Fetal standard Planes Assisted by Generated Sonographer Attention Maps." In: *The 21st International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2018).*

Chapter 6

(in preparation) **Yifan Cai**, Richard Droste, Pierre Chatelain, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble, 2019, "Temporal-aware Visual Attenion Modelling for FASP scans classification." In: *Mecial Image Analysis.*

Chapter 7

Richard Droste, **Yifan Cai**, Pierre Chatelain, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble, 2019, "Ultrasound Image Representation Learning by Modelling Sonographer Visual Attention" In: *The 26th international conference on Information Processing in Medical Imaging (IPMI 2019).*

(submitted) **Yifan Cai**[1], Arijit Patra[1], and J. Alison Noble, 2019, "Creating lightweight memory efficient ultrasound image models using sonographer gaze-assisted knowledge distillation" In *The 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019).*

Richard Droste, **Yifan Cai**, Pierre Chatelain, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble, 2019, "Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction" In: *23rd Conference on Medical Image Understanding and Analysis (MIUA 2019).*

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe

2 Literature Review

Contents

2.1 Introduction

2-D B-mode ultrasound is the most common imaging modality for fetal growth monitoring as well as screening for potential anomalies during fetal development and fetal cardio-vascular pathologies [**bonnet1999detection**]. In the United Kingdom, the National Health Service (NHS) regulates the routine pregnancy ultrasound

scans under the Fetal Anomaly Screening Programme (FASP) [**kirwan2010nhs**]. Pregnant women receive three routine scans during pregnancy: a *dating scan* in the first trimester (between 11^{+2} to 14^{+1} weeks of gestation) to estimate gestational age, an *anomaly scan* in the second trimester (between 18^{+0} and 20^{+6} weeks) that detects anomalies during fetal development, and a *growth scan* in the third trimester.

Small for gestational age (SGA), defined as a baby born with a body weight less than $10^{th}percentile$, is a very prominent cause of infant mortality mainly caused by intra-uterine growth restriction (IUGR) [**yu2003global**], and it is normally detected during the anomaly scan. Biometric measurements on six anatomical sections, including the head (two views), face, spine (two views), abdomen, and femur, are taken during the scan. The head circumference (HC), abdominal circumference (AC), and femur length (FL) are measured on standard planes of each anatomy and compared to corresponding standard growth curves. These standard planes are defined by clinical bodies such as the International Society for Ultrasound in Obstetrics and Gynaecology (ISUOG), British Medical Ultrasound Society (BMUS), Royal College of Obstetricians and Gynaecologists (RCOG) [**unterscheider2014definition**, **salomon2011practice**, **loughna2009fetal**], and FASP [**kirwan2010nhs**]. The detection of these standard planes requires a high level of expertise from sonographers [**maraci2014searching**], leading to common problems such as intra-observer [**chan2009volumetric**] and geographical variability [**garne2001evaluation**].

Automatic standard plane detection algorithms can potentially be helpful with regards to increasing sonographers' time efficiency as well as providing training tools for trainees. Early automated methods explored the potential of "bottom-up", hand-crafted local features of ultrasound video frames and employed classic machine learning tools such as random forests [**yaqub2015guided**] and boosting [**rahmatullah2011quality**]. However, such methods were limited by (1) the incompatibility of hand-crafted features with clinical meaningful regions (2) lack of high level, "top-down" constraints. Ahmed *et al.* [**Ahmed2016**] compensated for the second limitation through the utilisation of gaze-tracking data of sonographers as a

powerful, "top-down" interest-point operator to assist the abdominal standard-plane detection task, but the method was still limited by the use of hand-crafted features.

Recent success of deep neural networks (DNNs) in computer vision [**krizhevsky2012imagenet**] has inspired new methods and techniques in medical image analysis, and it has been shown that publication of medical image analysis papers using deep learning has grown exponentially between 2015 and 2017 [**litjens2017survey**]. The use of gaze-tracking data for medical image analysis, though not absent, is scarce. It is hypothesized that by leveraging the high-level information inherent in gaze-tracking data as well as the powerful feature-learning capability of DNNs could further improve model performance in standard plane detection.

The review is divided into three parts. First, a brief description of a routine ultrasound examination is provided. The second part is a review of the most important developments in deep learning research in image and video analysis, with a focus on papers published in the medical imaging domain. Third, research in visual attention and human gaze, and their applications in medical image analysis is discussed.

2.2 Fetal Ultrasonography

2.2.1 Standard Fetal Biometry

Three key fetal biometry planes are obtained so as to determine gestational age and fetal size during dating and anomaly scans: the fetal head, the fetal abdomen, and the fetal femur, as can be seen in Fig. 2.1. For the fetal head, 3 measurements are made: occipito-frontal diameter (OFD), bi-parietal diameter (BPD), and the head circumference (HC). For the fetal abdomen, three measurements are made as well: Anterio-posterior abdominal diameter (APAD), Transverse abdominal diameter (TAD), and the abdominal circumference (AC). For the fetal femur, only the length (FL) is measured, as the structure is quite simple. Key measurements based on standard biometry planes are compared to corresponding standard curves plotting biometric measurements between 5^{th} and 95^{th} percentile against gestational age. For example, an AC chart can be seen in Fig. 2.2.

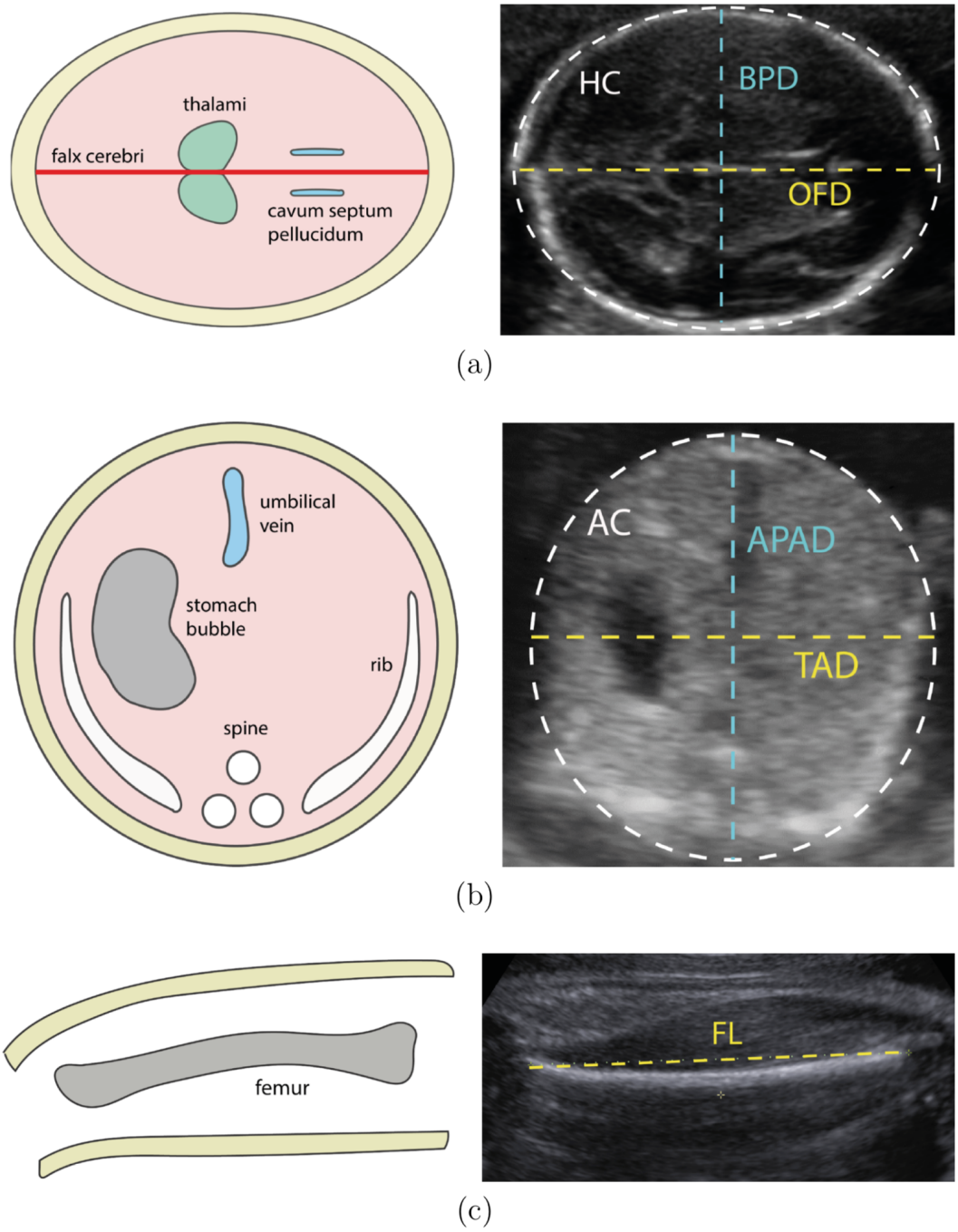


Figure 2.1: Cartoon representation of three key fetal biometry planes. (A) Standard fetal head plane where occipito-frontal diameter (OFD), bi-parietal diameter (BPD), and head circumference (HC) are measured. (B) Standard fetal abdominal plane where Anterio-posterior abdominal diameter (APAD), Transverse abdominal diameter (TAD), and abdominal circumference (AC) are measured. (C) Standard fetal femur plane, where femur length (FL) is measured. The cartoon is adapted from [**AhmedThesis**].

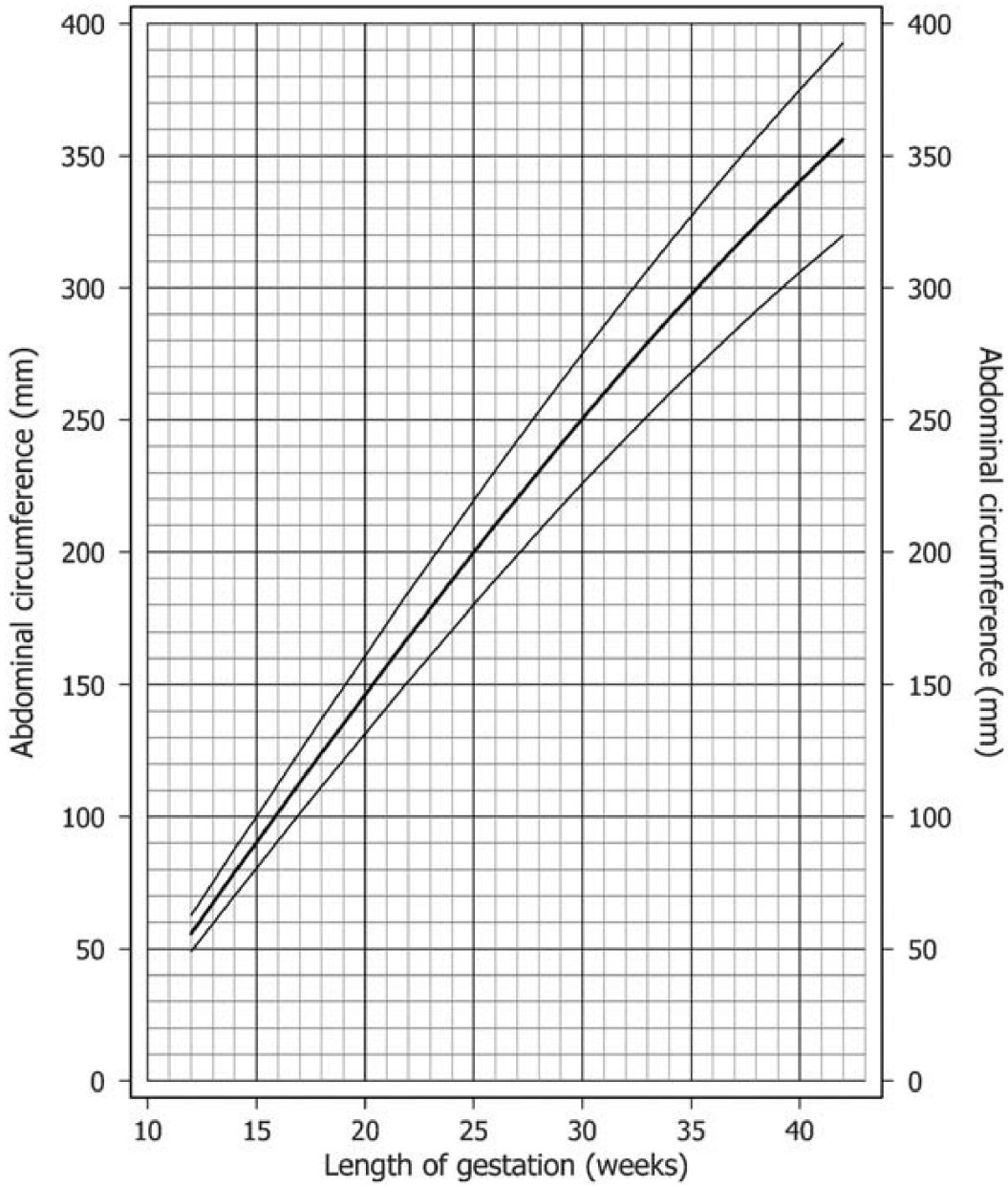


Figure 2.2: Abdominal Circumference (AC) against gestational age, plotting AC between 5^{th} and 95^{th} percentile. The figure is adapted from [**loughna2009fetal**].

2.2.2 Challenges for Standard Biometry Plane Detection

Standard plane identification, however, is challenging for clinicians and sonographers for several reasons. First, due to poor contrast, signal attenuation, acoustic shadows and other artifacts [**van2010prospective**, **germanakis2012assessment**, **szabo2004diagnostic**], ultrasound image quality is variable. Second, orientation and position of key anatomical structures, for instance the stomach bubble and umbilical vein on the AC plane , are variable. Third, with increasing gestational age, image quality deteriorates due to increased amount of adipose and other soft tissues. Due to these complications, it requires lengthy training (1-2 years) and a considerable amount of practice before sonographers can take high quality ultrasound images of the standard planes for measurements.

However, like many other developed countries, UK is experiencing a nationwide shortage of sonographer resources, both in quantity and quality, which is leading to severe difficulties for many NHS Trusts and Health Boards in meeting increasing demand, government targets and especially delivery of the national obstetric screening program [**parker2015educating**]. Many sonographers opt for early retirement, decrease in working hours or even leaving services completely, while the number of sonographers in training can barely keep up with the wastage and there is limited scope for increased training activities due to human resource constraints and financial limitations [**thomson2009developing**]. On the other hand, inter- and intra-observer variability in fetal ultrasound measurements is significant enough not to be ignored for the purpose of monitoring fetal growth [**sarris2012intra**]; the variability sometimes leads to erroneous diagnosis, causing unnecessary intervention, maternal anxiety, or in other cases inadvertently overlooking growth-restricted fetuses.

2.2.3 Classic Automated Ultrasound Analysis Methods

Given these challenges, as well as the demand for high-quality prenatal ultrasound in a larger population, automated ultrasound analysis methods, including the segmentation and localisation of key anatomical structures as well as

classification of ultrasound image contents, are needed. Classic computer vision methods have been successfully applied to medical image analysis problems. Most of these methods depend on local, hand-crafted image features including intensity [**yaqub2016plane**], local gradients [**nithya2009detection**], local phase [**rahmatullah2011quality**], Haar-like features [**rahmatullah2011quality**, **bridge2015object**, **namburete2015learning**, **AhmedThesis**] as well as SIFT [**lowe1999object**, **maraci2014searching**, **maraci2017framework**]. Machine learning methods based on these features have been applied to segmentation [**noble2006ultrasound**, **chalana1996automatic**, **rackham2013ultrasound**, **nithya2009det**
foi2014difference, **rueda2014evaluation**], localisation [**rahmatullah2011quality**, **rahmatullah2012image**, **rahman2016optimizing**, **ryou2016automated**] and classification [**carneiro2008detection**, **yaqub2010weighted**, **yaqub2016plane**, **maraci2014searching**, **maraci2017framework**, **bridge2015object**] problems in US image analysis.

One potential drawback of these methods is the use of only local features, lacking high-level constraints, such as the geometric relationship between anatomical structures [**liao2013representation**, **kim2013unsupervised**]. One way of compensation is the incorporation of gaze-tracking data collected from sonographers as a source of top-down information, and use the collected gaze data to build an interest-point operator that guides spatial selection of features to assist localisation of anatomical structures [**Ahmed2016**]. Another drawback is that hand-crafted image features may not necessarily identify anatomically salient regions; it also forfeits the opportunity to learn clinical meaningful features from the available data. In recent years, medical image analysis has been heavily influenced by the successful application of deep learning methods in computer vision [**litjens2017survey**], for example AlexNet [**krizhevsky2012imagenet**]. In the next section, important developments in deep learning for computer vision that is relevant to medical image analysis is reviewed.

2.3 Deep Learning for Image and Video Analysis

Impressive success achieved by deep learning methods in computer vision has attracted researchers in the field of computational medical imaging to investigate the potential of deep learning for medical image analysis. For example, success of neural networks in image classification tasks [**krizhevsky2012imagenet**, **simonyan2014very**, **he2016deep**, **szegedy2015going**, **huang2017densely**] inspired many applications of neural networks in medical image classification [**rajpurkar2017chexnet**, **hosseini2016alzheimer**, **korolev2017residual**, **roth2015anatomy**, **yan2015bodypart**, **baumgartner2017sononet**]. One of the major contributions of this thesis is related to US image classification, and all the classification networks used in this thesis are based on the architecture VGGNet [**simonyan2014very**]. In order to provide a comprehensive overview of the development of classification networks, several winning networks for Large Scale Visual Recognition Competition (ILSVRC) [**deng2009imagenet**, **russakovsky2015imagenet**], together with close-competitors, are discussed in the following section.

In addition to US image classification, the thesis also contributes to the detection of standard biometry planes in US video sequences. Notable video classification networks are discussed in order to understand how to process temporal information in video sequences.

Finally, because the thesis contributes to the study of visual attention modelling using deep networks, Attention in computer vision context and human visual attention modelling, often referred to as saliency prediction, are defined separately; inspiration on how to utilize predicted visual attention to assist classification tasks was drawn from literature in computer vision Attention.

2.3.1 Image Classification

The gold-standard image classification dataset was the one used by ILSVRC, commonly known as ImageNet. AlexNet [**krizhevsky2012imagenet**] achieved a top 5 error rate of 15.3%, a reduction of 43% in error rate from previous year. This work, for the first time, showed how powerful convolutional neural network is for

large-scale image classification. It made several important discoveries. First is the use of Rectified Linear Units (ReLUs), rather than hyperbolic tangent function ($tanh$), as non-linearity activation functions, which demonstrated better performance the benefit of less training time. Second is data augmentation techniques, such as translation, horizontal reflection, and mean subtraction, which all helped improve the model's generalisation capability. Third is the use of dropout layers to tackle overfitting problems. Finally, it employed a structure of successive convolution and pooling layers, followed by fully-connected layers in the end. This structure is still the basis for many of the state-of-the-art networks today, and its network training techniques are used in the work of this thesis.

ZF-net [**zeiler2014visualizing**] used an architecture very similar to AlexNet with the exception that it used 7×7 convolutional kernel in the first layer, instead of the 11×11 used by AlexNet. In addition, the network was more efficient as it only trained on 1.3 million images, instead of 15 million in the case of AlexNet.

VGGNet [**simonyan2014very**] (Fig. 2.3) further extended the idea of AlexNet. The main contribution is that simply making the network deeper can improve classification accuracy. It used 3×3 convolutional kernels throughout the network and demonstrated successive 3×3 kernels have equivalent receptive fields to those of 5×5, 7×7 or 11×11 kenerls used in AlexNet. Top-5 error rate was reduced to 8.0%. It also introduced scale jittering as a new type of data augmentation technique. GoogLeNet [**szegedy2015going**] of the same year explored, in addition to the depth, the width of the network. It introduces the inception module that implements 1×1, 3×3, and 5×5 convolutions in parrallel, and let the network itself decide how much each convolution should be used; also, the module captures both local features via small convolutions and higher abstracted features through larger convolutions.

ResNet [**he2016deep**] was the first architecture to pass human level performance on ImageNet, with a top-5 error rate of 3.75%. It was shown that stacking many layers of convolution together doesn't help improve model performance due to gradient vanishing or gradient explosion. The solution, residual module, utilises an additive skip connection as a shortcut, so deep layers have direct access to features

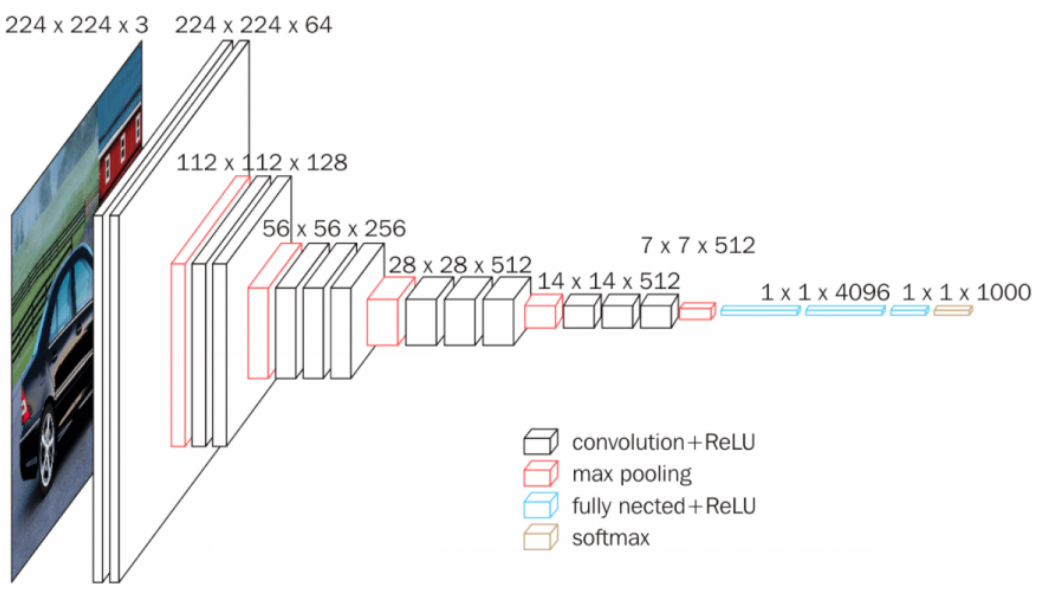


Figure 2.3: Architecture Of VGGNet. This figure was reproduced from [**simonyan2014very**].

from previous layers and gradient can easily propagate through network. With the help of residual module, networks can be extremely deep, commonly over 100-200 layers. The concept of short-cut connection was taken even further by DenseNets [**huang2017densely**], which connects each layer to every other layer in a feed-forward fashion through concatenation. The architecture further helps alleviate the vanishing-gradient problem and substantially reduces the number of parameter used.

CUImage [**ouyang2016factors**] further improved classification results using an ensemble method that combines several pre-trained models, including Inception-v3 [**szegedy2016rethinking**], Inception-v4 [**szegedy2017inception**], Pre-Activation ResNet-200 [**he2016deep**] and Wide ResNet [**zagoruyko2016wide**]. It was the first model to reach classification error on test set under 3.0% (2.99%).

A comparison of network performances can be seen in Fig. 2.4.

In medical imaging, classification networks have been applied to computer-aided diagnostics (CADx) in x-ray [**rajpurkar2017chexnet**], MRI [**hosseini2016alzheimer**, **korolev2017residual**] and CT [**roth2015anatomy**, **yan2015bodypart**], and many other problems. In the area of Ultrasound imaging, Chen *et.al* [**chen2015standard**] developed a model based on AlexNet to detect FASP standard planes on a dataset

acquired using a sweeping protocol, achieving precision of 71.4%. Gao *et. al* [**gao2016describing**] used pre-trained weights of AlexNet on ImageNet to classify fetal US images including fetal skull, abdomen, heart, and demonstrated that the features learnt on natural images could be transferred to US image dataset. The mean classification accuracy reached 91.5% comparing to 87.9%, the performance achieved by a network of the same architecture but initialized using random weights. SonoNet [**baumgartner2017sononet**] was built on VGGNet for FASP standard plane detection on routine free-hand US scan. Three variants of the SonoNet were tested. The largest was SonoNet-64, which uses the same architecture as VGG-16 [**simonyan2014very**], while the other, namely SonoNet-32 and SonoNet-16, adopt architectures with halved and quartered number of kernels in all layers. The best performing model was SonoNet-64, achieving mean F1-score of 82.8%.

The models explored in this Thesis, namely the SonoEyeNet [**cai2018sonoeyenet**], Multi-task SonoEyeNet [**cai2018multi**], and Temporal SonoEyeNet were all based on the architecture of VGG-16 with quartered number of kernels in each layer, instead of on those of more advanced architectures such as ResNet or DenseNet. The reasons for this choice are two-fold. First, VGG-16 is simple in the sense that it uses 3×3 convolutional kernels in all layers, making it easier to compare the effectiveness of gaze information in improving classification performances by fusing gaze information with CNN feature representations on differnet layers, as seen in [**cai2018sonoeyenet**]. Second, it makes it easier to compare with contemporary models for US image classification [**baumgartner2017sononet**].

2.3.2 Video Classification

The first attempt to use modern Neural Networks for video classification was [**karpathy2014large**] in which several ways of processing temporal information were explored using a single-stream CNN. However, as the authors discovered, the model couldn't reach the performance of the state-of-the-art models built on hand-crafted features as the learnt spatial-temporal features didn't capture motion patterns in the data.

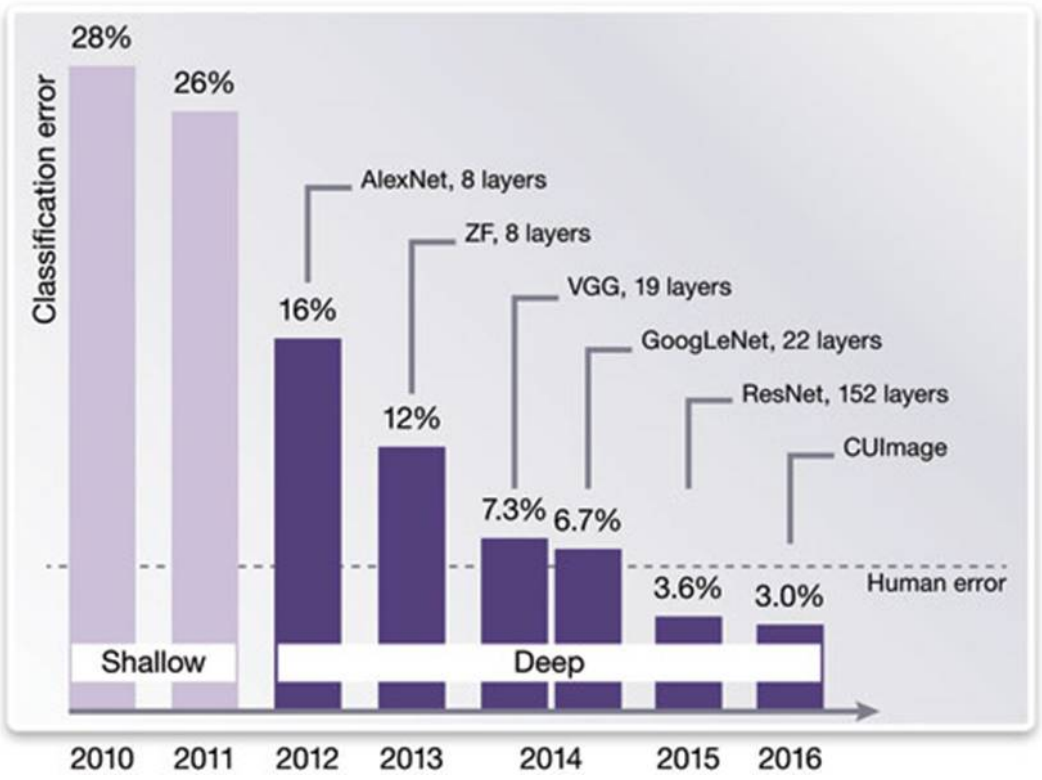


Figure 2.4: A summary of notable networks on ImageNet. This figure was reproduced from [**cooper_2017**].

Simonyan and Zisserman [**simonyan2014two**] adopted a two-stream architecture: one stream for spatial information extraction using a CNN [**karpathy2014large**], while the other stream explicitly modelled motion features in the form of stacked optical flow. Input to the first stream was a single video frame, while that to the second stream was a stack of optical flow maps across 10 consecutive frames. The model improved video classification performance, yet it required pre-processing videos to calculate optical flows, making end-to-end training impossible.

Long-term Recurrent Convolutional Networks (LRCN) [**donahue2015long**] is a one-stream network with a RNN-based architecture (Long Short-Term Memory, or LSTM modules) to learn a spatio-temporal feature representation of video clips. The architecture also enabled end-to-end training. C3D [**tran2015learning**] further improved the single-stream model by using 3-D convolution, instead of 2-D convolution, as a feature extractor before feeding spatial-temporal features into an LSTM. Further improvements on this idea, including using an attention

mechanism [**yao2015describing**], were developed and achieved better results on the Sports1M dataset [**karpathy2014large**]. Improvements in two-stream models [**feichtenhofer2016convolutional**, **wang2016temporal**, **girdhar2017actionvlad**] were also achieved, yet the requirement of pre-computing optical flow maps for training was problematic. HiddenTwoStream [**zhu2017hidden**] solved this problem by generating optical flow maps on the fly using a second network.

Even though the above video classification models all performed video-level classification, rather than frame-level classification, the way they extracted spatio-temporal information was of interest to the research work of this thesis, especially for learning spatio-temporal feature representations of US video sequence to model temporal variations of visual attention. One way to extract spatio-temporal information is through optical flow; the other was using recurrent neural network. An example of the first method for US video analysis was [**gao2017detection**], where a 2-stream based model (video sequence and optical flow) was proposed for a US video sequence classification tasks (skull, abdomen, heart, background) and achieved 90.3% precision on heart detection. An example of the latter is the Temporal HeartNet [**huang2017temporal**], which learns to predict the visibility, viewing plane, location and orientation of the fetal heart in a multi-task manner at the frame level using LSTM module to learn spatial-temporal features. A third way of learning spatio-temporal information was demonstrated in [**maraci2018can**], where 1D dilated convolutions [**van2016wavenet**] were used.

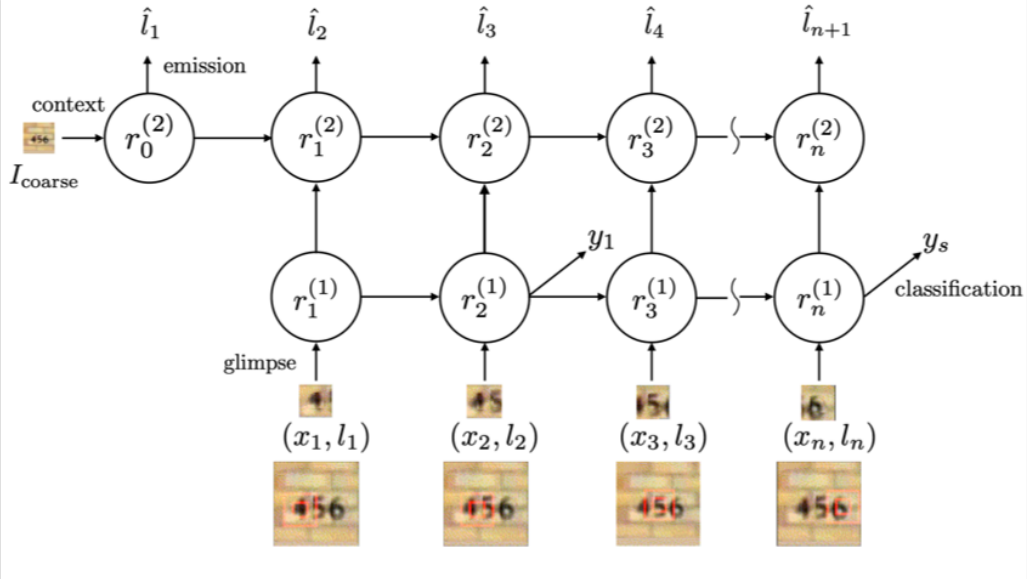
Temporal SonoEyeNet (in Chapter 6) chose to extract spatio-temporal features using recurrent neural networks. First, extracting optical-flow maps for video sequences requires an additional computational step and it can be computationally expensive. Second, as demonstrated later, current state-of-the-art dynamic visual attention model relies on recurrent neural networks to extract spatio-temporal information. Third, using RNNs to extract spatio-temporal features makes network architecture simpler without having to train a separate branch of CNN for optical flow map prediction, the method that [**gao2017detection**] used to generate optical flow maps.

2.3.3 Computer Vision Attention

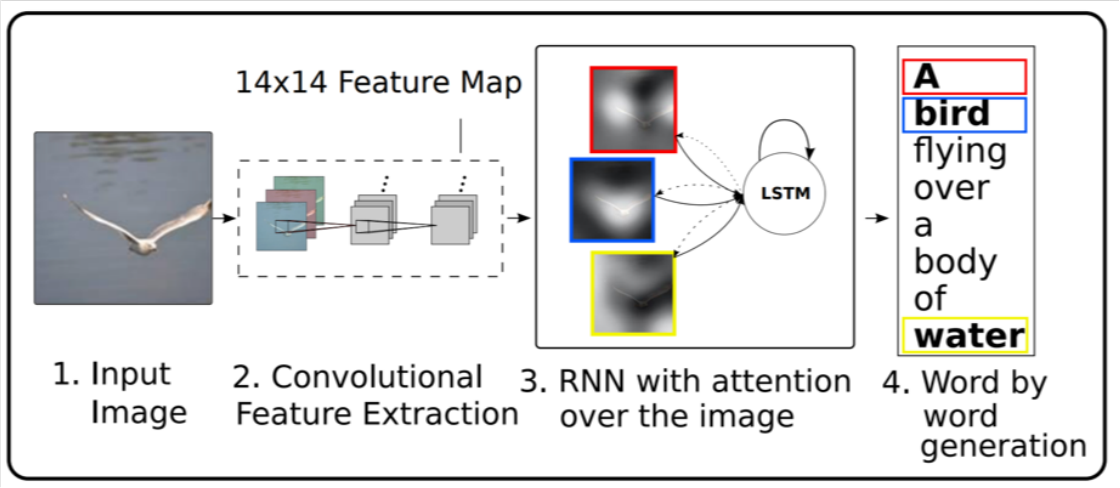
An important contribution of this thesis is to model sonographer visual attention. However, in computer vision, attention is widely used but not defined based on human gaze-tracking data. In order to avoid confusion, Attention in computer vision is discussed in this section, while visual attention based on human gaze-tracking data is discussed in the next. In the context of the rest of the thesis, unless stated otherwise, attention always refer to human gaze-based visual attention.

The Attention discussed here is purely from a computer vision point of view, especially in the context of deep learning, that involves *implicitly* learning to selectively process a region or a sequence of regions in the input image purely based on image features *for the purpose of assisting* an image analysis task. Two types of computer vision Attention mechanisms have been explored: **hard attention**, which processes only a part of the input image (usually stochastically sampled), and **soft attention**, which processes all parts of the input image through assigning weights to different regions in the feature space.

The first attempt of using **hard attention** for image classification was by Larochelle and Hinton [**larochelle2010learning**] who developed a mechanism called retinal transformation that maintains high resolution at the point of the selected "fixation" (the center point of a region that the algorithm selected to analyze) while downsampling regions with increasing distance to the fixation point. Through the use of a Restricted Boltzmann Machine (RBM), information at sequentially sampled fixation points are combined. Mnih *et al.* [**mnih2014recurrent**] developed a Recurrent Attention Model (RAM), which replaces the RBM with a recurrent neural network to accumulate information from different fixations. Its Location Network, which is used to find the next fixation point, samples fixation points through a stochastic process, making the network non-differentiable. In order to train the network, the model uses a Policy Gradient Method, a type of reinforcement learning method. Ba *et al.* [**ba2014multiple**] further extended RAM to a Deep Recurrent Attention Model (DRAM) by increasing the depth of the feature extractor and implementing two recurrent cells for classification and localisation



(A)



(B)

Figure 2.5: (A). An example of hard attention used for image classification in [**ba2014multiple**]. This figure was reproduced from [**ba2014multiple**]. (B). An example of soft attention used for image captioning in [**xu2015show**]. This figure was reproduced from [**xu2015show**].

respectively (Fig. 2.5(A)). Hard attention methods have been applied to computer vision tasks such as object recognition [**gonzalez2015active**], action detection [**mathe2014multiple**], and unsupervised feature extraction [**zou2012deep**]. Jaderberg *et al.* [**jaderberg2015spatial**] tackled hard attention using a Spatial Transformer Network (STN). A STN is differentiable because instead of using stochastic process to determine the location of attention it learns the affine transformation parameters, including translation, rotation, (isotropic) scaling and skewing, through regression.

Soft attention, on the other hand, assign weights to different regions of image features in the form of heatmaps. Xu *et al.* [**xu2015show**] developed a model that utilizes soft attention for visual scene description (Fig. 2.5(B)). It introduced a "doubly stochastic regularization" term that encourages the model to pay equal attention to every part of the image, making it possible to train this model in a deterministic manner using standard backpropagation techniques but at the same time stochastically. The feature maps extracted by the CNN were fed into a LSTM for caption generation, which assigns weights to features through attention maps, which are updated every time a new caption was generated. Variants of this model were developed for natural images [**yang2016stacked**, **wang2017residual**, **seo2016progressive**, **jetley2018learn**].

In medical imaging, soft attention has been used for its simplicity as a ready-to-use module to improve deep learning model performances. Kumar *et al.* [**kumar2016plane**] fused features from trained classification networks and saliency features from object detection networks to boost fetal US biometry plane identification performance; Schlemper *et al.* [**schlemper2018attention**] introduced a self-gated soft-attention mechanism that allows the network to contextualise local spatial information useful for detection of US standard planes and weakly-supervised object localisation. Gao *et al.* [**Gao2019learn**] built a self-gated soft-attention mechanism not only for spatial information, but also takes into consideration the temporal variation of attention in US videos.

2.4 Visual Attention Modeling

One major contribution of this thesis is the study of sonographer visual attention modelling using neural networks. In the following section, several classical saliency models using hand-crafted features were reviewed before introducing the development of saliency prediction using deep learning. As the thesis discusses visual attention prediction on independent ultrasound images as well as visual attention transitions on consecutive frames of US video clips, both static saliency models as well as dynamic saliency models are reviewed in this section.

2.4.1 Nomanclature

With the availability of eye trackers, human gaze information are recorded and two types of eye movements can be distinguished: **fixations** (points on which human eyes linger for a prolonged period of time), and **saccades** (rapid eye movements in between fixation points) [**Yarbus1967**]. To describe human visual search and detection, Kundel and Nodine developed a model, which comprised of three components: global impression, discovery search, and reflective search [**Nodine1987**]. The first glance at the image, which takes up to several hundred miliseconds, produces a global impression of the image that leaves the viewer with a fairly accurate conception of the content of the image. This initial impression, as brief as it is, is very hard to measure during recording, but sets the stage for detailed analysis of the image by central vision. Following the global impression, eyes move over the image to inspect each target in detail, accomplished by clusters of closely-spaced fixations in the discovery search stage. In the reflective search stage, the observer cross references against other targets and potential targets. Unlike the global impression stage, the discovery and reflective stages are measurable using the eye tracker, forming the basis of a widely recognized two-stage model for visual search.

Visual attention, different from the computer vision Attention discussed above, describes the human behavior when viewing images or a sequence of video frames. A **Visual Attention Map** is a representation of visual attention in a weight

matrix form; locations in the visual attention map with higher weight indicate higher probability of attracting human fixation. A **visual attention model**, in the context of this thesis, refers to a deep learning model that learns to model visual attention *explicitly* and predict visual attention maps based on input image frames.

In computer vision literature, algorithms that try to model human visual behavior are referred to as "saliency models", and metrices that measure the difference between model prediction and ground-truth human visual attention maps are referred to as "saliency metrices". In this thesis, "saliency models" are sometimes used interchangably with "visual attention models", and "saliency map" with "visual attention map". This is especially true in the following section. Similarity measures between ground-truth visual attention maps and model-predicted visual attention maps are referred to as **saliency metrices**.

The **eye-tracking data**, collected by the eye tracker, contains three parts of information: (1) pupil diameters, (2) pupil positions in the 3-D environment, and (3) **gaze data** describing where we look, a series of gaze point positions projected on 2-D sampled at frame rate. Only the gaze data collected by an eye tracker is used in the scope of this thesis; pupil diameters and pupil positions are not used. However, due to historical nomanclature [**AhmedThesis**] that used "eye-tracking data" to refer to gaze data, early publication that contributed to this thesis [**cai2018sonoeyenet**] inherited that nomenclature. In this thesis, eye-tracking data refers to the gaze data.

2.4.2 Saliency Models based on feature integration theory

Early saliency models followed the feature integration theory (FIT) [**treisman1980feature**], defining saliency by the fusion of several hand-crafted features extracted on multiple scales. Other approaches define saliency in terms of information theory, such as self-information [**zhang2008sun**], information maximization [**bruce2006saliency**], discriminant saliency [**gao2007bottom**], or spectrum-based methods [**hou2012image**]. Finally, many data-driven image feature learning algorithms have been proposed to predict saliency and achieved state-of-the-art results. The following section will

start by introducing several classical saliency prediction models based on FIT, and then describe current state-of-the-art image feature learning methods.

Saliency Models built purely from bottom-up features

Limitations of classic saliency prediction methods lie in their adoption of a pure bottom-up approach – only low-level, local features such as edges, colour, disparity and direction of movement are extracted from the images to predict visually salient regions. Koch and Ullman proposed the first biologically-plausible computational architecture using feature integration theory to define saliency maps in a bottom-up manner [**koch1987shifts**]. Itti *et al.* improved on Koch and Ullman's approach by introducing feature maps at different scales (pyramids) for each feature [**itti1998model**]. Such bottom-up saliency models work well when higher order semantics are reflected in low-level features (as is often the case for isolated objects, and even for reasonably cluttered scenes), but tends to fail if other factors dominate: for instance free-viewing of images without clearly isolated objects (forest scenes or foliage) [**itti2001computational**]. Meanwhile, only object features explicitly represented in at least one of the feature maps could lead to "pop-out" and these models failed at detecting target salient for unimplemented hand-crafted feature types.

Combining bottom-up and top-down methods

Saliency is not independent of the nature of a particular task and it can be influenced by contextual, figure-ground effects [**james2013principles**]. In real world images, the semantic content of the scene, the co-occurrence of objects, and task constraints have been shown to influence where attention and eye movement go [**chun1998contextual**, **henderson2003human**, **neider2006scene**].

Combination of bottom-up and top-down features for saliency prediction was first attempted by Cerf et al. by simply combining Itti *et al.*'s model with an additional face detector [**cerf2008predicting**]. They used a bottom-up saliency model based on intensity, colour, and orientation and introduced an additional top-down "face channel" based on the established Viola & Jones feature-based template matching

face detector algorithm [**viola2001rapid**]. Judd et al. further introduced mid- and high-level features to the model and labelled all features using gaze data from 15 users viewing 1003 natural images (referred to as MIT1003 database) as fixated (positively-labelled) and non-fixated (negatively-labelled). On the low-level, local energy of the steerable pyramid filters in additional to the classic Itti *et al.* features such as intensity, orientation, colour, contrast were used [**itti1998model**]. On the mid-level, they employed a trained horizon line detector; high-level features included face detector [**viola2001rapid**], person detector [**felzenszwalb2008discriminatively**] and car detectors etc. These features with their corresponding labels were fed into a support vector machine to train a saliency prediction model.

Even though combination of bottom-up and top-down features achieved state-of-the-art performance at that time, limitations were also obvious. High level features such as a face detector and person detectors were hand-crafted, designed specifically for each category of features, which made it very hard to scale. Improvement on the state of the art has been achieved mostly by incrementally adding more hand-tuned features to existing models. Meanwhile, image feature learning using neural networks gained momentum in the computer vision community, as a result of the approach's superior performance on several vision tasks ranging from scene classification to object and face recognition [**krizhevsky2012imagenet**, **cox2011beyond**]. Saliency prediction community was inspired by such development and employed image feature learning as a novel method to train more powerful models.

2.4.3 Saliency Models built on Neural Networks

Static Saliency Models

Static saliency models are those that predict saliency maps on independent input images, rather than a sequence of consecutive video frames. Vig *et al.* was among the pioneers in adopting image feature learning for saliency prediction by creating the Ensemble of Deep Networks (eDN) [**vig2014large**]. They used an automated hyper-parameter optimization algorithm [**bergstra2013making**] to search through large numbers of randomly initialised multilayer (1-3 layers) feature extractors (CNNs)

that fed the extracted features into an L2-regularized, linear, L2-loss SVM so as to train to predict for each location in a new test image its probability of fixation using the right ensemble of these feature extractors. Kümmerer *et al.* developed Deep Gaze [**kummerer2014deep**]. Similar to eDN, it used fixed deep neural networks as feature extractors. However, Deep Gaze specifically solved the over-fitting problem by building the model in a transfer learning manner, utilizing early layers of AlexNet [**krizhevsky2012imagenet**] as the fixed feature extractors and linearly combining responses from each layer. Instead of searching through complex spaces for richly-parameterized CNNs, Deep Gaze learns weights for the linear combination of responses from each layer of AlexNet.

In contrast to top-performing hand-tuned models, these approaches made no assumption about what lower-level features or higher-level concepts attract eyes; instead, they allowed the hierarchical models to learn complex patterns and their linear combination from gaze-labelled natural images, making such methods more generic and cleaner than approaches that rely on domain-specific hand-crafted features. A limitation of eDN is heavy over-fitting. Deep Gaze avoided over-fitting using transfer learning, yet the transferred early layers lack further fine-tuning for saliency prediction purposes.

Learning from Deep Gaze that uses networks trained for object recognition tasks to perform saliency prediction, Huang et al. [**huang2015salicon**] designed a saliency prediction model using CNNs pre-trained on ImageNet, including AlexNet [**krizhevsky2012imagenet**], VGG-16 [**simonyan2014very**] and GoogLeNet [**szegedy2015goin**] and implemented them in a fully convolutional way [**long2015fully**]. Saliency evaluation metrics, such as Kullback-Leibler divergence (KLD) [**tatler2005visual**], Normalized Scanpath Saliency (NSS) [**peters2005components**], Correlation Coefficient (CC) [**jost2005assessing**], were used as loss functions to fine-tune the CNNs. Also, inspired by work on visual recognition and scene labelling [**he2015spatial**, **farabet2013learning**], which found that using images at multiple scales improves accuracy of prediction, the authors scaled all training images to two resolutions and fed them into two streams of the CNN, which share the same filters, so that

neurons are trained to detect the same image patterns at different resolutions [**bylinskii2015saliency**]. [**liu2015predicting**] also took advantage of the multi-resolution concept, but used three different scales, rather than two.

One recent development in saliency prediction was the exploration of relationships between regions in an image using recurrent modules. The DSCLRCN model (Deep Spatial Contextual Long-term Recurrent Convolutional Neural network) [**liu2018deep**] uses an LSTM module to explore the global relationships between learnt local saliency of small image regions. Saliency Attentive Models (SAM-ResNet &SAM-VGG) [**cornia2018predicting**] developed a spatial attentive model that combines LSTM modules a fully connected CNN to improve saliency prediction performance. A further development is the idea of rather than dividing images into smaller regions and exploring spatial relationships between features of these regions so as to predict visual attention, the algorithm directly models pixel-wise saliency. The Deep Visual Attention (DVA) model [**wang2018deep**] proposes a CNN encoder-decoder structure to capture hierarchical saliency information, combining feature representations from shallow layers containing global saliency information with those from deep layers with local saliency response. Saliency GAN (SalGAN) [**pan2017salgan**] utilizes a GAN structure: a generator that predicts human visual attention and a discriminator that discerns whether the attention map was generated or not; binary cross-entropy (BCE) loss is used so that each pixel is modelled as a binary random variable either being attended to or not. State of the art at the time is claimed by several heavily-engineered models, including EML-Net [**jia2018eml**], FUCOS [**bruce2016deeper**], and Attention Push [**gorji2017attentional**].

Dynamic Saliency Models

Dynamic saliency models are those that predict saliency maps on a sequence of consecutive video frames. Early attempts to model saliency on video utilize only short-term temporal information in the form of optical flow, accounting for temporal differences between consecutive frames. The earliest work was a two-stream network

[**bak2018spatio**], in which one stream of the CNN extracts information from individual video frames while the other extracts optical flow maps. Chaabouni *et al.* [**chaabouni2016transfer**] made use of early layers of a trained CNN network for image classification as a fixed feature extractor for saliency prediction; it also utilized residual motion as an additional input, but this additional temporal information was also very short-term (extracted between consecutive frames). [**leifman2017learning**] also used optical flow in tandem with a predicted visual attention map on the previous frame to assist attention prediction on the current map. [**bazzani2016recurrent**] utilized longer-term temporal information by using a 3-D convolution to extract spatial-temporal information from 16 consecutive frames; locations of visual attention are parameterised using a mixture of gaussians.

Recurrent modules (LSTM) are used to model temporal variation in visual attention [**bazzani2016recurrent**, **liu2015predicting**]. Object-to-motion CNN (OM-CNN) [**jiang2017predicting**] improved saliency prediction performance through the use of convolutional LSTM (convLSTM) [**gorji2017attentional**], replacing the inner-production operation within vanilla LSTM by convolution operation, thus allowing the LSTM to learn both spatial and temporal information. ConvLSTM was also used to improve video saliency prediction [**gorji2018going**].

The state of the art video saliency prediction model is [**wang2018revisiting**] (Fig. 2.6). It proposes a CNN-LSTM architecture to exploit both the spatial and temporal information for predicting video saliency. With a supervised attention mechanism, it explicitly captures static saliency information; the static saliency information is fed into LSTM to focus on learning dynamic information. In this way, models already trained on static gaze-tracking dataset can be directly used in combination with the attentive module (residual operation and convLSTM) to predict dynamic saliency.

2.4.4 Deep vs. classic Saliency Models

It is shown that classic (FIT) models' performances lack behind those of deeper models [**borji2018saliency**], as can be seen in Fig. 2.7. Three reasons establish

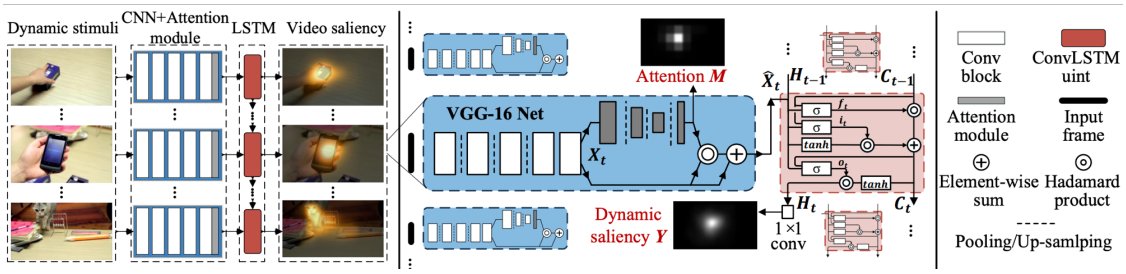


Figure 2.6: Network architecture of the proposed video saliency model from [**wang2018revisiting**] with an attentive CNN-LSTM architecture. This figure was reproduced from [**wang2018revisiting**].

deep learning-based models as the current state-of-the-art for saliency prediction. The first is the combination of bottom-up approach, which captures local features, and top-down approach, which provides context and semantic meanings. While some classic models may incorporate top-down features such as an object detector or face detector, deep models learn the complex high-level visual cues implicitly in deep layers and are able to combine them with bottom-up features in shallow layers. The second factor is the transition from the accumulation of hand-crafted features to the usage of back-propagation and stochastic gradient-descent to learn features and their combination for saliency prediction. The third factor is transfer learning, which utilizes feature extractors trained from large image database for other tasks and then are specifically fine-tuned towards saliency prediction purposes. Classic models lack such capabilities.

2.4.5 Visual Attention Applications

Human visual attention has been used in a number of ways for image analysis. A first class of algorithms record human gaze information in order to perform inter-observer comparisons, as usually seen in clinical or pychological research studies [**Beam2006**, **nodine1987using**, **kundel2007holistic**, **kundel2008using**, **Antonelli2007**, **James2007**, **Ahmed2014**]. A second class uses recorded human gaze information as input, in addition to medical images, to assist image analysis tasks [**ramanathan2009automated**, **yun2013studying**, **karthikeyan2013and**, **papadopoulos2014training**, **soliman2016towards**, **shanmuga2015eye**, **mcguinness2010comparative**, **boykov2001interactive**,

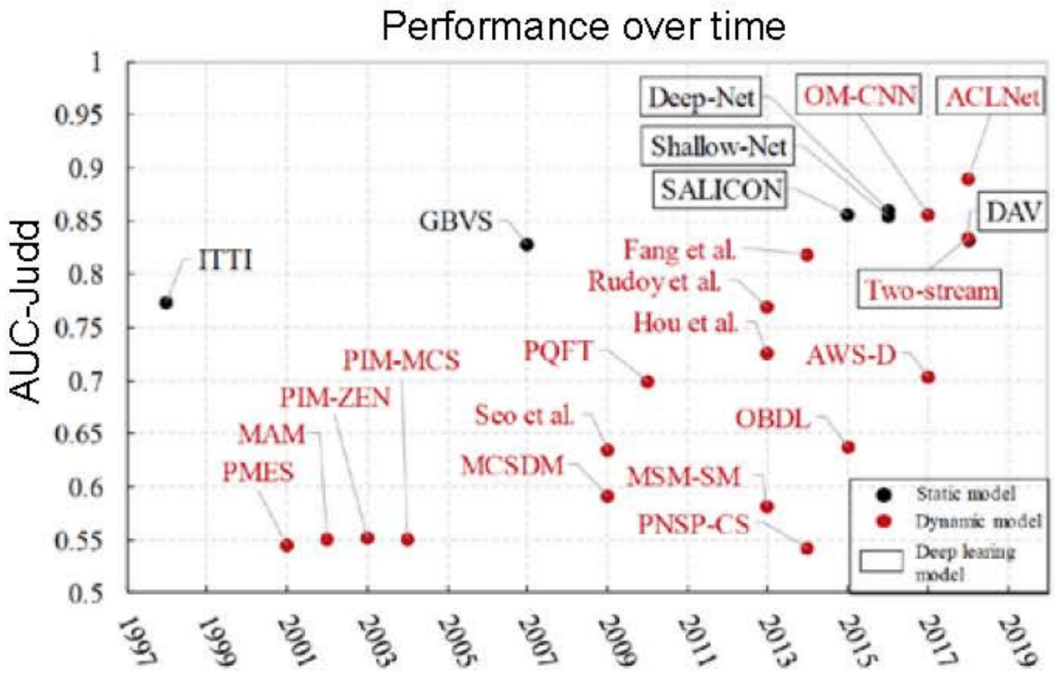


Figure 2.7: Comparison of saliency prediction models in terms of their AUC scores over time. Model names with square boxes around them are deep learning models, while the others are classic models. The figure is adapted from [**borji2018saliency**].

xu2013incorporating, **sadeghi2009hands**, **bjorkman2010active**]. Early algorithms require gaze input in real time in order to assist its task; later algorithms took advantage of the development of saliency prediction models that learnt expert visual attention, so that no additional gaze information was needed during inference.

The first class of algorithms has been shown to be very useful for the investigation of visual search strategies of different clinicians analyzing medical images on different parts of the body. Nodine and Kundel [**nodine1987using**] investigated visual search strategies of radiologists for lung cancer detection, followed by similar studies on mammography [**kundel2007holistic**, **kundel2008using**]. Antonelli and Yang [**Antonelli2007**] recorded eye movements of radiologists when they are searching for abnormalities in lung CT volumes; James *et al.* [**James2007**] tracked the eye movement of surgeons when they are performing surgery. In the realm of US image analysis, Ahmed and Noble [**Ahmed2014**] carried out the first research to analyze eye movement of sonographers viewing fetal abdominal ultrasound images and volumes.

The second class of algorithms have been successfully applied for object detection [**ramanathan2009automated**, **yun2013studying**, **karthikeyan2013and**, **papadopoulos2014training**, **soliman2016towards**, **shanmuga2015eye**], segmentation [**mcguinness2010comparative**, **boykov2001interactive**, **xu2013incorporating**, **sadeghi2009hands**, **bjorkman2010active**], visual question answering [**das2017human**], and image captioning [**sugano2016seeing**] for natural images. However, application of visual attention in medical image analysis is scarce. Krupinski *et al.* [**krupinski1996visual**] found that gaze duration is a useful predictor of missed lesions in mammography, making gaze duration a potential tool for perceptual feedback. Gandomkar *et al.* [**gandomkar2017model**, **gandomkar2017icap**] combined gaze pattern as well as image features to classify radiologists' decisions. In US image analysis, Ahmed and Noble [**Ahmed2016**] built a vocabulary of visual words trained on SURF descriptors extracted around eye fixations to classify head, abdominal and femoral image frames, achieving accuracies of 76%, 68% and 64% for the head, abdominal and femoral images respectively. No deep learning models were developed prior to the research in this thesis.

2.5 Conclusions

In the first section of the literature review, fetal ultrasonography and classic automated US image analysis methods were discussed. In the second section, development of deep learning models in natural images as well as medical image analysis, especially US image analysis, were reviewed. The third section discusses attention in the context of computer vision so as to distinguish from the visual attention discussed in the fourth section. Key visual attention modelling models, either based on feature integration theory or deep learning-based, together with their applications in natural image or medical image analysis were discussed. It is found that even though visual attention modelling was proven to assist other image analysis tasks, research on sonographer visual attention has just begun, and more research on sonographer visual attention modelling as well as their application for US image analysis is needed. That is exactly what this thesis tries to address.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe

3 Datasets

Contents

3.1 Introduction

Two different US video datasets were used in this thesis. The first is a single sweep dataset collected under *Intergrowth* 21^{st} Project [**villar2014international**], and it is used for the contributions in Chapter 4 and Chapter 5. The second is a free-hand anomaly scan dataset collected under the project *Perception Ultrasound by Learning Sonographic Experience (PULSE)*, and it is used for the contribution in Chapter 6. Both datasets consist of US video frames and corresponding gaze-tracking data. However, they are different in terms of data size, scan type, length, image quality, gaze-tracking protocol, and sampling rate of gaze data. The difference in US video acquisition protocol can be seen in Fig. 3.1.

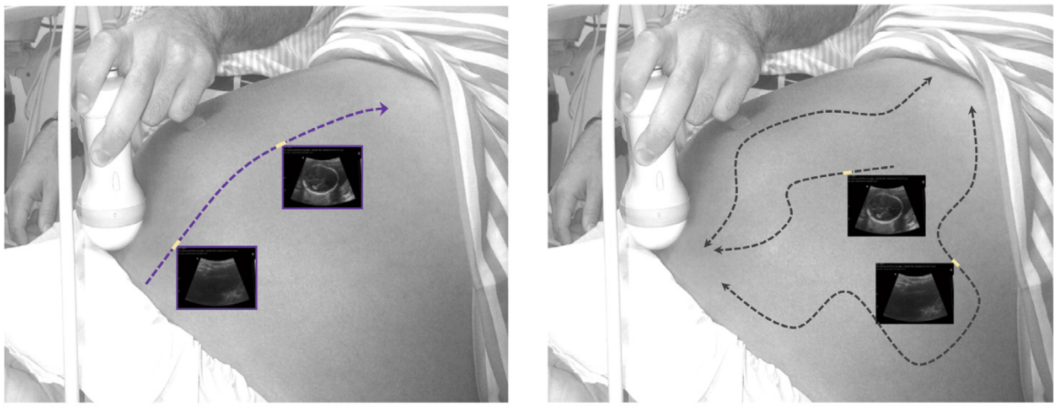


Figure 3.1: Left: probe movement during US video acquisition in single sweep dataset. Right: probe movement during US video acquisition in PULSE dataset. This figure was reproduced from [**maraci2017framework**].

3.2 Single Sweep Dataset

This dataset is used in Chapter 4 and 5 of this thesis. It consists of 33 fetal US videos (1616 frames) of fetal abdomen, each lasting 1-3 seconds. The distribution of frame number in each video clip can be seen in Fig. 3.2. This dataset was selected from a larger dataset of 412 fetal US videos [**maraci2014searching**] each lasting approximately 6–8 seconds of healthy volunteers with fetuses of 28 weeks gestation or higher. Only the abdominal section of the videos were selected. All frames contain at least one of the following anatomical structures: stomach bubble, umbilical vein, a circular abdominal wall, and spine, and did not show a kidney. These abdominal video clips were manually selected by an experienced sonographer from a larger dataset of 80 fetal abdominal US videos that were acquired according to a freehand US sweep protocol. The videos were acquired on a mid-range US machine (Philips HD9 with a V7-3 transducer) by moving the probe from the maternal cervix to the fundus following the longitudinal axis of the uterus, as presented in Fig. 3.1(left). Each video $V(x, y, t)$ has spatial dimensions x and y corresponding to the transverse view of the fetal abdomen, and temporal dimension t corresponding to the progression of time (movement of the probe is along the longitudinal axis of the fetal abdomen)(Fig. 3.3).

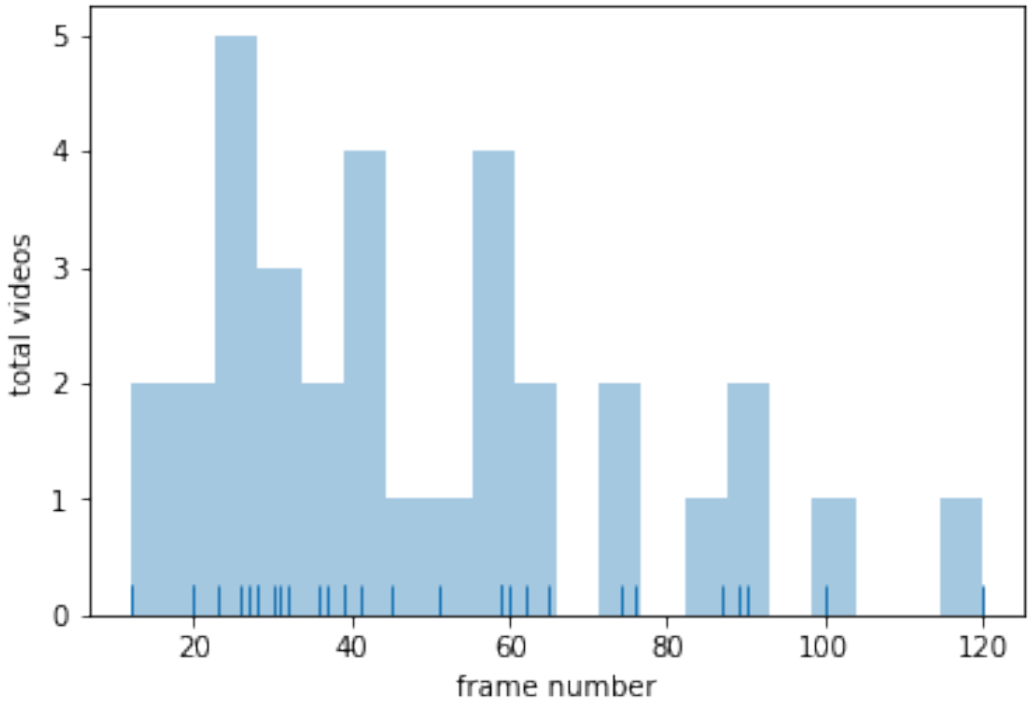


Figure 3.2: Distribution of frame numbers in each video clip in the single sweep dataset.

The videos were use as stimuli in gaze tracking experiments and divided into independent datasets for training and testing of the automated standardised plane detection algorithms in Chapter 4 and 5. It is worth noting that the algorithms presented in these two chapters are trained with only videos containing fetal abdominal frames, thus are limited in applicability to only abdominal standard planes detection task. An example of a video clip can be seen in Fig. 3.4.

3.2.1 Gaze tracking experiment

The sonographer was presented with consecutive frames from 33 fetal abdominal ultrasound videos on the screen, and had unlimited viewing time for each frame. The sonographer was asked to identify the standard AC plane from these frames through free navigation, using up- and down- keys on a keyboard to move to the next or the previous frame. Once the sonographer had identified the standard AC plane, the space key is pressed and the index of standard AC plane is recorded. Sonogrpaher viewing behaviour for each video was recorded by an eye tracker

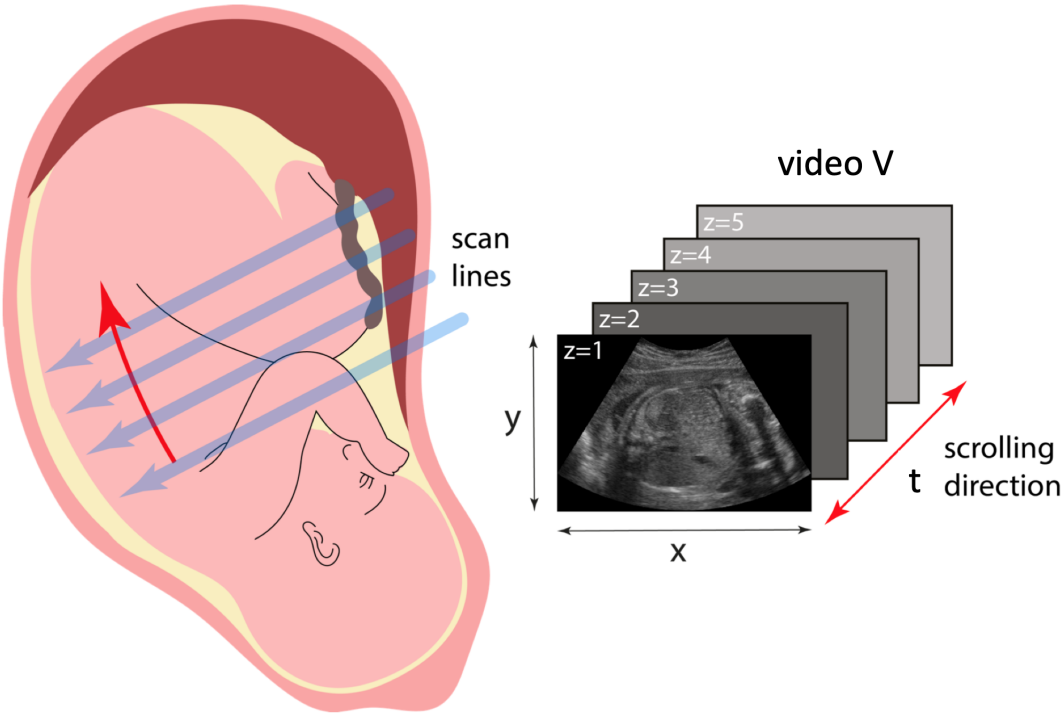


Figure 3.3: Left: Schematic of the freehand US sweep acquisition of fetal abdominal videos. Transducer scan lines (blue) are parallel to the acquired image planes; probe sweeps along the longitudinal axis of the fetal abdomen (z axis, red). Right: In gaze tracking experiment, sonographers scroll frames along the z axis. The figure is adapted from [**AhmedThesis**].

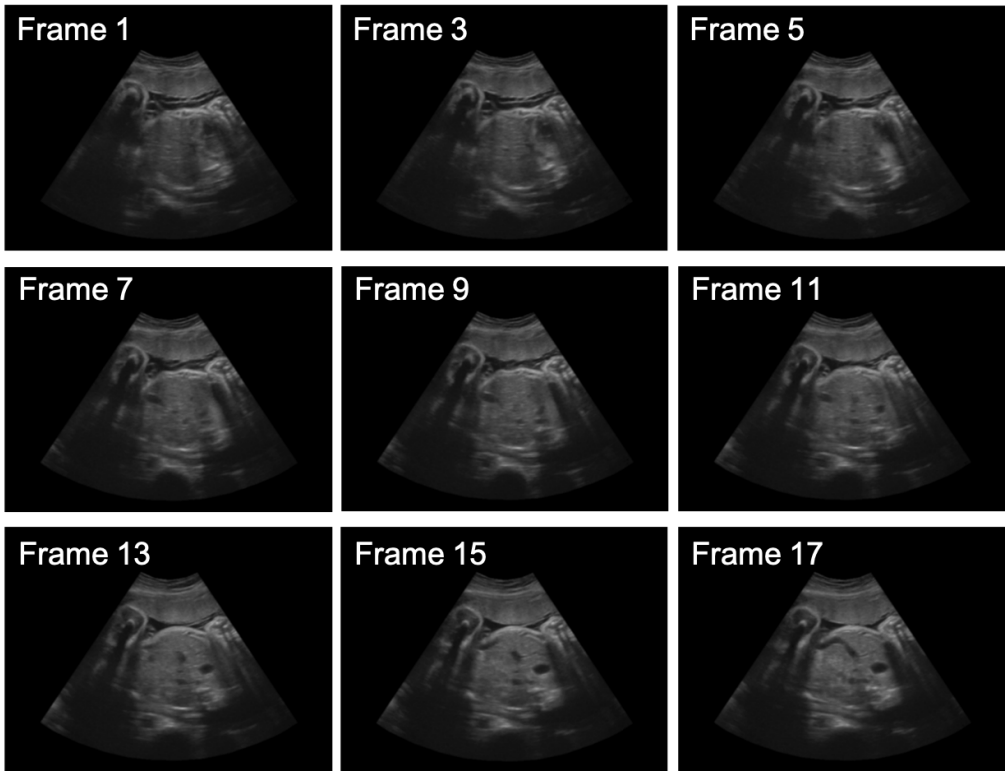


Figure 3.4: An example of consecutive frames in a video clip in the Free-hand Sweep Dataset

(The EyeTribe) placed in front of the screen, approximately 0.5 m from the viewer. Gaze-tracking data was recorded (x-y coordinates and time stamps) at 30 Hz. To avoid viewer fatigue, the sonographer was given 10 minutes break after viewing 10 videos. Futher details of the gaze tracking experiment can be seen in Chapter 4.

3.3 PULSE Anomaly Scan Dataset

This dataset is used in Chapter 6 of this thesis. It consists of 280 video clips containing one of the three biometry planes: 89 video clips of fetal abdomens, 71 video clips of fetal brains, 76 video clips of fetal femurs, and 44 video clips of "other" class that does not contain any biometry planes, each lasting 3-7 seconds. A distribution of frame numbers in each video clip can be seen in Fig. 3.5. Except for the "other" class, all video clips contain standard planes of one of the key anatomies: abdomen, brain, and femur. All video clips were manually selected by myself (the author) from a larger dataset of more than 500 2^{nd} trimester fetal anomaly full scan videos, each lasting between 30 to 45 minutes. All free-hand anomaly scans were conducted on a GE Voluson E8 scanner(General Electric, USA) while the video signal of the machine's monitor was recorded lossless at 30 Hz. Each video was processed to remove software user interface and manually annotated by the sonographer during the anomaly scan according to the FASP protocol [**kirwan2010nhs**].

The dataset was divided into independent datasets for training and testing of the automated standardised plane selection algorithm in Chapter 6. An example of a video clip can be seen in Fig. 3.6.

3.3.1 Gaze tracking experiment

While video signal of the ultrasound machine's monitor is recorded, gaze tracking data were recorded using a Tobii Eye Tracker 4C (Tobii, Sweden) that records the point-of-gaze (relative x, y-coordinates with corresponding timestamp) and 3-D eye position of each eye at a rate of 90 Hz, effectively recording 3 gaze points per frame. The eye tracker was rigidly attached under the display area with a magnetic mounting bracket as per the instruction of the product. The eye tracker was calibrated for

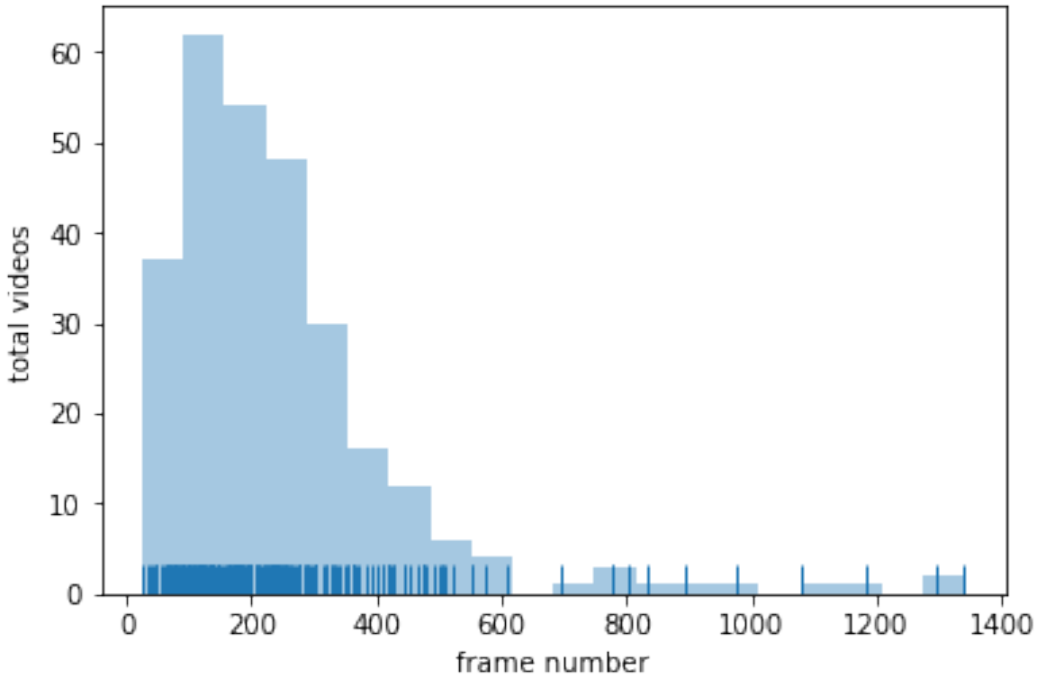


Figure 3.5: Distribution of frame numbers in each video clip in the PULSE dataset.

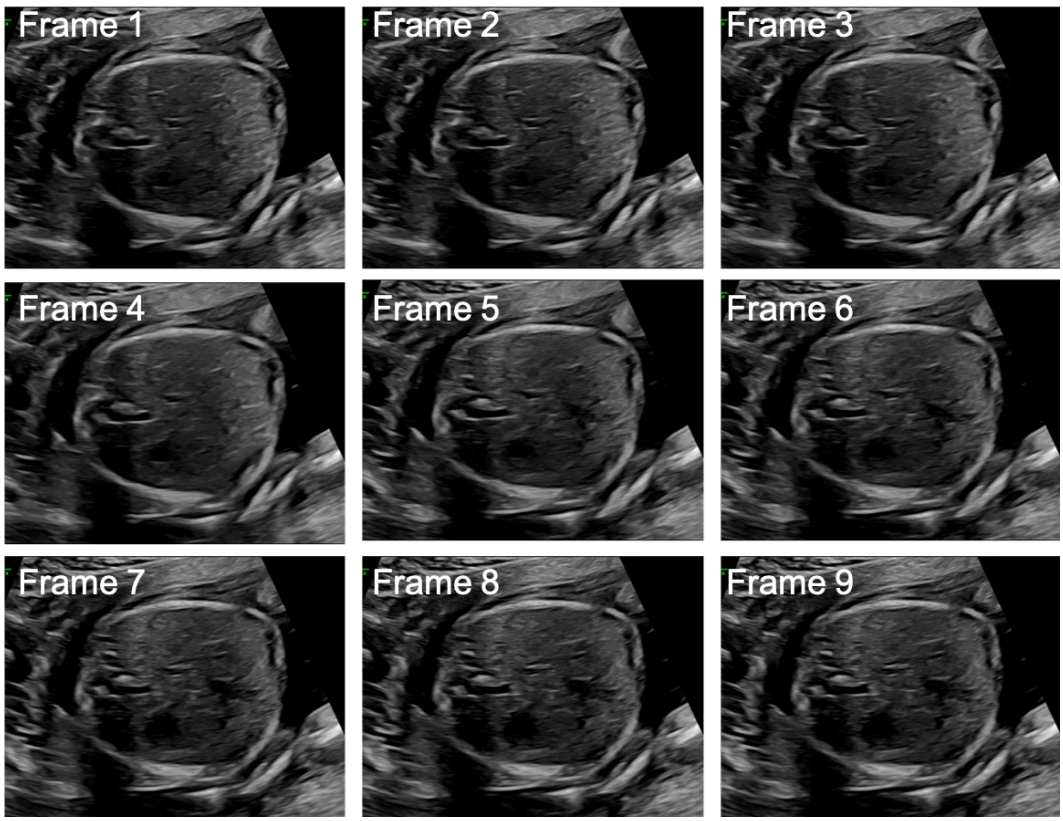


Figure 3.6: An example of consecutive frames in a video clip in the PULSE Anomaly Scan dataset.

each sonographer following a 9-point calibration protocol. Sonographers were free to adjust the height of the chair and the inclination of the monitor, and to operate ultrasound probe in order to perform the routine of ultrasound examinations without being affected by the existance of an eye tracker so that realistic gaze data were recorded. Further details of the gaze tracking experiment is provided in Chapter 6.

3.4 Summary

Two datasets were used for the gaze-tracking experiments as well as the training of automated US image and video analysis algorithms within this thesis. Both datasets contained B-mode 2-D ultrasound videos. However, these two datasets were different. The first dataset was acquired following a single sweep protocol with retrospective gaze-tracking data. The data size was small, with 33 video clips and 1616 US video frames in total, covering only fetal abdomen. Annotations are made retrospectively during gaze-tracking experiment. The second dataset was 2^{nd} trimester freehand anomaly scan video dataset with simultaneous gaze-tracking data. The data size was larger, with 280 video clips and more than 20 thousand frames in total, covering all biometries according to the FASP protocol. Annotations are made according to the FASP criteria.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe

4

Standard Fetal Ultrasound Plane Detection Informed by gaze-tracking

Contents

4.1 Introduction

The Abdominal Circumference (AC) is an important biometric measurement used to monitor fetal development. Sonographers identify standard AC planes by observation of the presence of several anatomical landmarks defined by clinical

bodies, e.g. International Society for Ultrasound in Obstetrics and Gynaecology (ISUOG) [**salomon2011practice**] and Fetal Anomaly Screening Program (FASP) [**kirwan2010nhs**]. An example of the standard AC plane can be seen in Fig. 4.1. Automatic detection of standard AC planes in 2-D and 3-D dataset aided by sonographer gaze-tracking was previously attempted using sonographer visual heatmaps as interest operators to first learn explicitly the location of key anotomical structures, i.e. stomach bubble and umbilical vein, using a sliding window detector; the localization of key anatomical structures was then refined using a pictorial model [**Ahmed2016**]. The model was able to recognize standard AC planes using the detected anatomical structures on each frame. However, it requires extensive tuning of the hand-crafted features, *e.g.* Haar-like features or intensity gradient magnitude and orientation; it requires prior knowledge to define geometric constraints, *e.g.* relative positions and lengths of anatomical structures; it is not readily generalizable, and requires separate models for separate standard planes.

This chapter takes a different approach to standard AC planes detection. First, it does not explicitly localize the stomach bubble and the umbilical vein in 2-D fetal abdominal US images; instead, it uses a Convolutional Neural Network(CNN) to directly learn a standard AC plane classifier, which extracts features around regions of sonographer visual attention (i.e. regions that sonographers fixated on) and uses them to distinguish standard AC planes from background planes, allowing end-to-end training of the model. Second, it does not assume prior knowledge of anatomy as constraints. Third, it does not use hand-crafted features extracted from images; instead, the CNN is able to learn a set of features from low to high level in different layers, thus making the model more flexible. It is also possible to scale to other standard plane detection tasks using the same model.

I conducted retrospective gaze-tracking experiments on a single sweep dataset whereby 8 experienced sonographers were asked to annotate the standard AC planes in a series of abdominal fetal ultrasound videos. Sonographers navigated through the frames from the ultrasound video using a keyboard's up- and down-keys. While they were exploring the frames presented on the screen their gaze positions

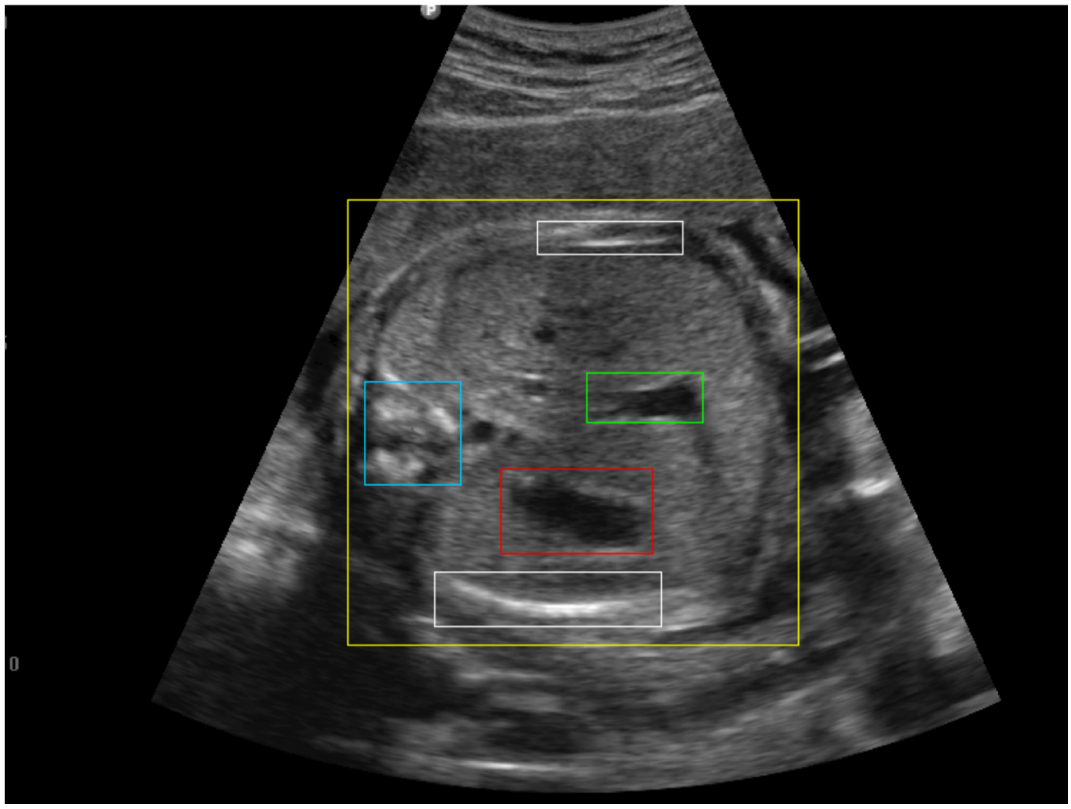


Figure 4.1: Standard abdominal circumference plane with key anatomical landmarks and regions of interests labeled by bounding boxes of different color. Yellow: abdominal wall. Red: stomach bubble. Green: umbilical vein. Blue: spine. White: ribs. This figure was reproduced from [**AhmedThesis**].

were recorded using an eye-tracker. In order to use gaze data as an input in the CNN model, fixation points were extracted using filters for time and visual angle, and visual heatmaps were subsequently generated for each frame by convolving a Gaussian kernel with a standard deviation that corresponds to typical human visual angle of 1.5 degree [**strasburger2011peripheral**].

This chapter is inspired by Ahmed *et al.* [**Ahmed2016**] but, to our knowledge, considers for the first time how gaze-tracking data can inform plane detection within a CNN framework, which I call the ***SonoEyeNet (SEN)***. Specifically, this chapter considers different CNN models, which combine information from gaze-tracking data and corresponding US video frames for standard AC plane detection. Sonographer visual heatmaps were generated from gaze-tracking data, and different methods of processing visual heatmaps and their fusion with image feature maps were investigated. We show that with the assistance of human visual

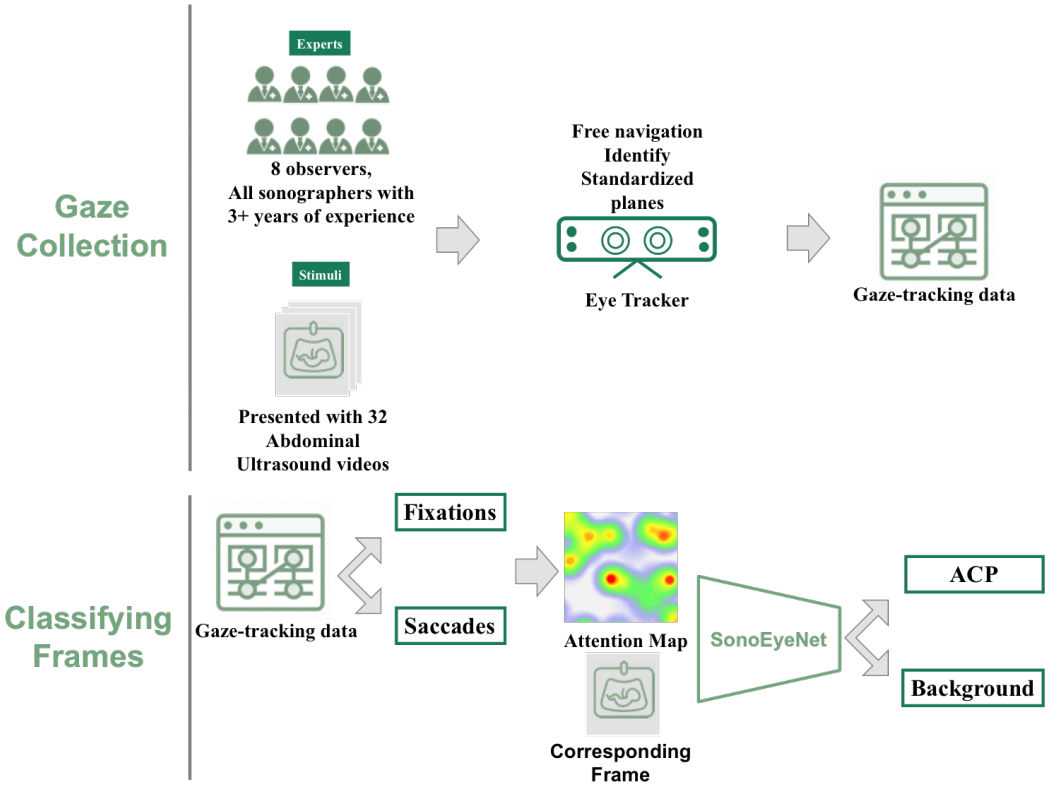


Figure 4.2: This chapter is divided into two main parts. The first describes the procedure of the gaze-tracking experiment, and the second elaborates on how gaze-tracking data are used to attempt frame classification task.

fixation information, the precision, recall and F1-score of AC plane detection was increased to 96.5%, 99.0% and 97.8% respectively, compared to 73.6%, 74.1% and 73.8% without using eye fixation information [**Baumgartner2017**]. The overall structure of the methods presented in this chapter can be seen in Fig. 4.2.

4.2 Originality and Individual Role

I made changes to an early version of a Python application *EyeSpy* developed by Maryam Ahmed, a former D.Phil student at the University of Oxford Institute of Biomedical Engineering, to perform eye-tracking experiments. I processed frames from the abdominal fetal ultrasound video in the single sweep dataset, and collected and processed the eye-tracking data recorded by an eye-tracker. Video frames were annotated by 8 experience sonographers. Using Keras and TensorFlow, both open source Python libraries, I trained and tested *SonoEyeNet* to classify abdominal AC

planes. This work was accepted to and presented at IEEE International Symposium on Biomedical Imaging (ISBI) 2018 [**cai2018sonoeyenet**].

4.3 Gaze-tracking Experiment

I conducted gaze-tracking experiments by inviting 8 experienced sonographers to view a series of frames from 2-D fetal abdominal ultrasound videos in order to gain insights on the exploratory and decision-making strategy of sonographers. Their gaze-tracking data were used as interest-point operators to assist frame classification under a CNN framework.

4.3.1 Methods

Stimuli

The single sweep dataset was used as gaze-tracking stimuli, which consisted of 33 fetal US videos (1616 frames) from 33 pregnant women, each lasting one to three seconds. These abdominal video clips were manually selected by an experienced sonographer from a larger dataset of 412 fetal US videos each lasting approximately six to eight seconds of healthy volunteers with fetuses of 28 weeks gestation or higher according to a freehand US sweep protocol [**maraci2017framework**]. Only the abdominal section of the videos were selected. The videos were acquired on a mid-range US machine (Philips HD9 with a V7-3 transducer) by moving the probe from the bottom to top of the woman's abdomen. Video frames were presented to observers on a 15-inch 1920×1080 pixels LCD monitor, through a custom built user interface *EyeSpy* which received x and y gaze co-ordinates averaged across the left and right eyes, and timestamps at a sampling frequency of 30 Hz.

Hardware

Gaze data were recorded with an EyeTribe v1.0 (the Eye Tribe, Denmark) eye tracker. The setup of the experiment can be seen in Fig. 4.3.

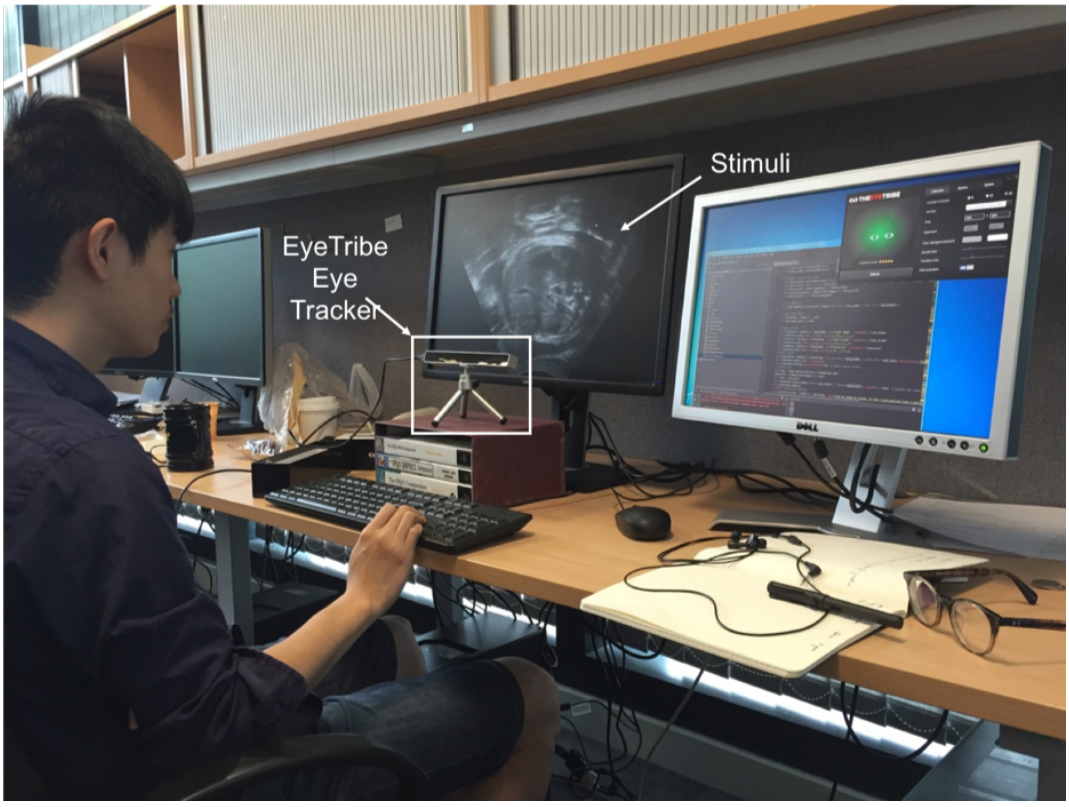


Figure 4.3: The experimental setup of gaze-tracking experiment on fetal abdominal US videos, consisting of a screen that displays stimuli, an eye-tracking device, and a keyboard for navigation through ultrasound frames.

Participants

Eight observers, with normal acuity, participated in the study. All participants were sonographers affiliated with Xuzhou Medical School with 3 to 10 years of experience.

Collection Procedure

Calibration of the eye tracker was conducted before each gaze collection experiment. During calibration, 9 circular targets were displayed at random sequence on the 4 corners, 4 edge mid-points as well as the center of the screen. The calibration procedure was repeated until the error between all recorded gaze positions and their corresponding target positions fell below the threshold of 1.5° visual angle.

Sonographers were presented with consecutive frames from 32 fetal abdominal ultrasound videos on the screen, and they had unlimited viewing time for each frame. Sonographers were asked to identify the standard AC plane from the these

frames through free navigation, using up and down keys to move to the next or the previous frame. Once they had identified the standard AC plane and pressed the space key, eye movement data was saved and the sonographer proceeded to the next video. To avoid viewer fatigue, sonographers were given 10 minutes break after viewing 10 videos. Sonographer viewing behavior for each video was recorded by an eye tracker placed in front of the screen, approximately 0.5 m from the viewer.

The recorded gaze-tracking data include a timestamp t (the eye tracker operates at 30 Hz so the difference between consecutive timestamps is 33 ms), and the mean x and y gaze co-ordinates across left and right eyes.

Fixation Filtering

After the raw data recorded by the eye tracker were extracted, it was necessary to filter these data to separate the fixation points (where the gaze lingers) from saccades (fast eye movement between fixations). Gaze-tracking data (x, y-coordinates and timestamp t) were processed according to the protocol in [**Mathe2012**]. Gaps in the tracking data, due to tracking errors, were filled in by interpolation. In order to remove high frequency noise (due to eye tremor), a moving average filter with a window size of 3 data points was applied. Any eye movement with angular velocity below 30° s^{-1} was classified as a fixation, and all the other points were classified as saccades [**Ahmed2016**]. This angular velocity threshold was translated into a pixel per second threshold taking into account that the distance between the eyes of the observer and the screen was 0.5 m. Fixations less than 80 ms in duration were discarded and labeled as a saccade; fixations points 0.5° visual angle in space were merged by taking the mean of the merged points. These filtering were conducted in accordance with the parameters employed by Tobii [**olsonEyeTracking**].

4.3.2 Results

An example of sonographer gaze distribution on frames of a fetal abdominal US video during free navigation can be seen in Fig. 4.4. Each red dot represents a gaze-tracking data point (x and y coordinates) specific to a timestamp on a US video

frame. In the case that a sonographer revisited a frame several times during free naviation, gaze points are accumulated and overlaid on a single frame. The viewing behavior shows a global-focal pattern, confirming the discovery of [**AhmedThesis**]. For early frames, the sonographer's gaze points are more global, exploring prominent patterns out of the abdominal wall first and then moving into the wall, as can be seen in *Frame 1* in Fig. 4.4. As a sonographer explores subsequent video frames (*Frame 2* to *Frame 9*), no significant anatomical landmark emerges, resulting in brief glimpse with only a few gaze points concentrated in the centers on those frames. Starting from *Frame 10*, anatomical structures *i.e.* umbilical vein and stomach bubble, start to appear and the sonographer fixated on those regions. These two structures are even more prominent in *Frame 11* and *Frame 12*, and the sonographer actively compares these prominent regions to make a decision, as reflected by the dense gaze points distribution. The sonographer finally decided that *Frame 12* is the standard AC plane. Frames after *Frame 12* are only viewed briefly to confirm the decision, and *Frame 20* was not viewed.

In order to better understand how sonographers move their gaze within each frame as well as how they navigate through different frames, i.e. decisions to view the next or previous frame, sonographers' visual tracks are plotted in Fig. 4.5. Visual tracks represent the sequence of regions that sonographer gaze visits during exploration. The x and y axis define the image plane, while the z axis defines the frame number, with the first frame at the bottom. The gaze points distribution plotted in Fig. 4.4 can also be represented as *Sonographer 1*'s visual track in Fig. 4.5, as can be seen on the top left corner.

Two observations can be made from Fig. 4.5. First, inter-observer variability in visual tracks is high for the same fetal abdominal US video: different sonographers adopt different viewing strategies to find standard AC planes. On this particular ultrasound video, *Sonographer 1, 2, 4, 5* adopt a viewing strategy that starts with a global search (rapid scanning of the whole image) but gradually narrow down to certain regions for a focal search, while others combines global and focal searches without a specific order. Second, final decisions for the standard AC plane vary

among sonographers. For this particular video, the majority of the final decisions concentrate on frame 14, 15 and 16 (Fig. 4.6).

Due to experimental limitations and eye-tracker instability [1], gaze data were collected on all videos for only one sonographer. That sonographer had extremely still posture during experiments. We chose that data to train a classifier for standard AC plane. The classifier design is described in the next section.

4.4 Convolutional Neural Networks

Supervised learning driven by convolutional neural networks (CNNs) has contributed greatly to recent success in computer vision and machine learning. CNNs are highly suited for image classification tasks as demonstrated by recent developments in the ImageNet competitions [**krizhevsky2012imagenet**, **simonyan2014very**, **he2016deep**]. Our problem of standard AC plane detection is formulated here as a supervised classification problem, where each video frame has a corresponding label indicating whether it is a standard plane or a background frame. We selected CNNs as a classifier for its versatility and scalability, thanks to the available toolboxes for Graphical Processing Unit (GPU) acceleration.

4.4.1 Problem Formulation

The aim of the algorithm is to fit a non-linear function, *i.e.* a CNN denoted as $f(\cdot)$ that maps two input domains $\mathcal{X}$ (ultrasound video frames) and $\mathcal{M}$ (their corresponding visual attention maps) to an output domain $\mathcal{Y}$ (binary class labels). We assume $(x^{(i)}, m^{(i)}, y^{(i)}) \in \mathcal{D} \sim \mathcal{X} \times \mathcal{M} \times \mathcal{Y}$ represents a data point sampled from a set $\mathcal{D}$ that has N samples, $i \in \{1, 2, ...N\}$, $x^{(i)}$ represents an ultrasound video frame, $m^{(i)}$ a corresponding visual attention map, and $y^{(i)}$ the desired output. We further denote that the CNN is parameterised by θ, so we can write it as $f(\cdot; \theta)$. Then the problem is simplified to optimizing the parameter θ so that it achieves minimum prediction error on a unseen dataset $D_{test} \sim p_{data}$, the data-generating distribution. We design a training dataset D_{train} with empirical distribution $\hat{p}_{data}$

[1] The EyeTribe eye tracker has over-heating and occasional drift issues

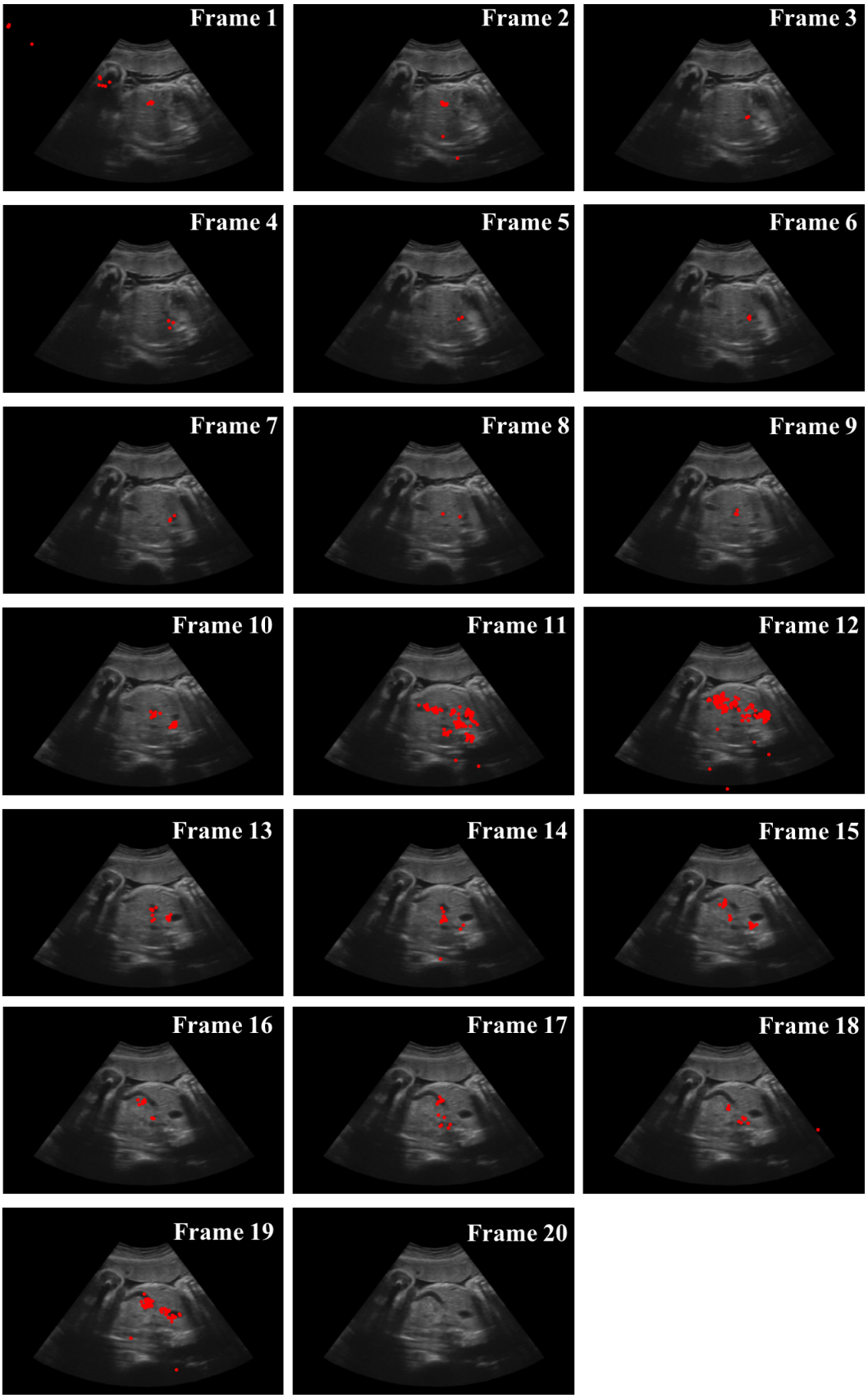


Figure 4.4: Gaze points (red) of a single sonographer plotted onto all frames in a single fetal abdominal US video.

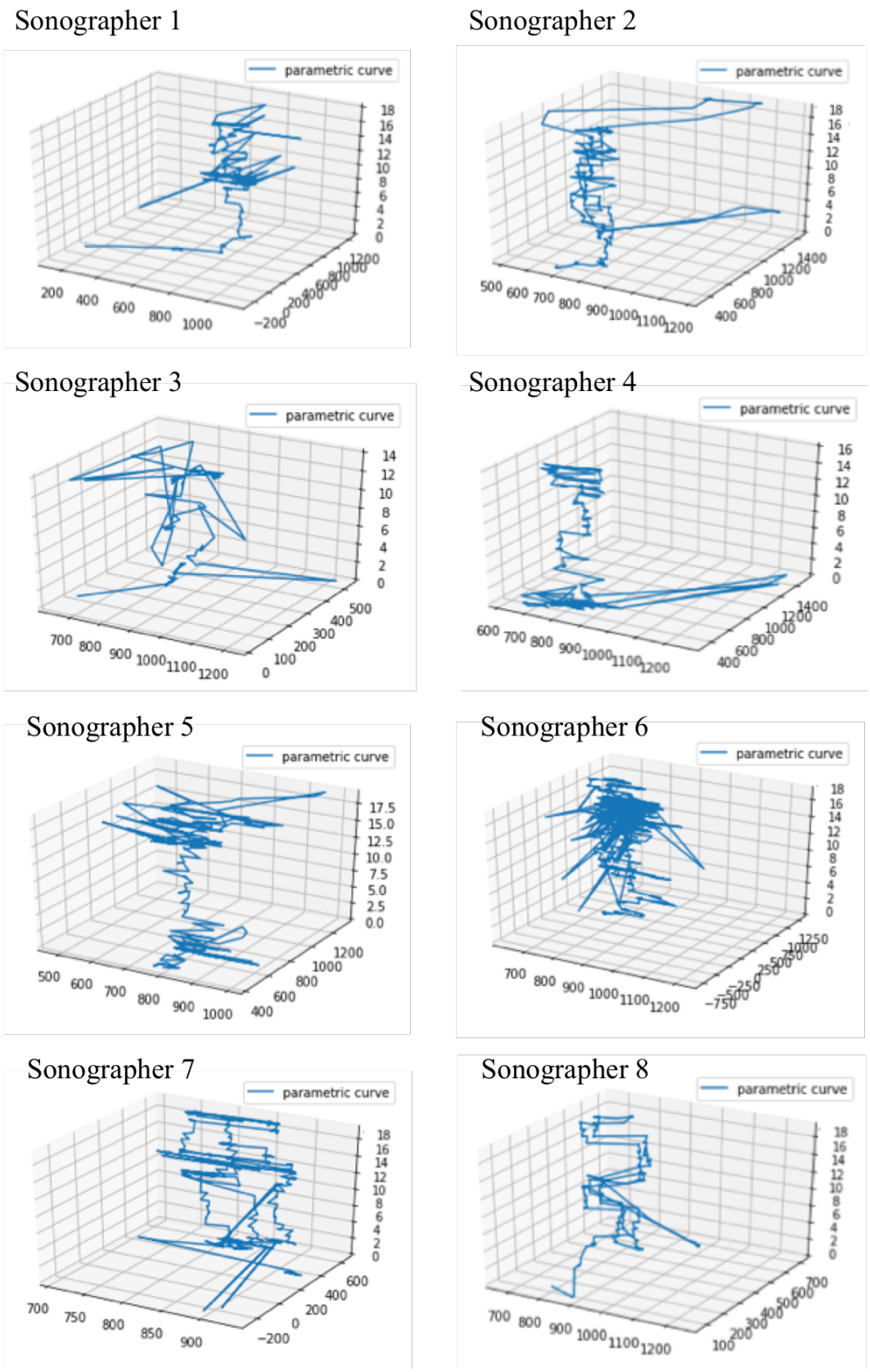


Figure 4.5: Visual tracks of 8 different sonographers on the same fetal abdominal video clip.

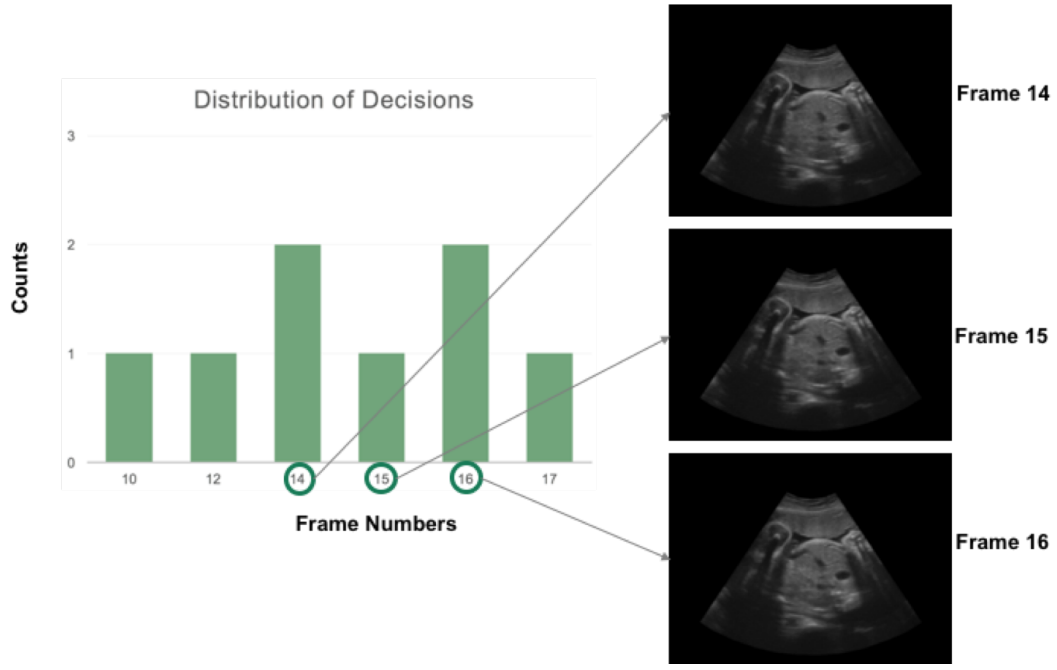


Figure 4.6: Distribution of 8 sonographers' final decision on the standard AC plane for a fetal abdominal US video.

so that we aim to minimize the prediction error on D_{test} by means of minimizing the prediction error on D_{train} using parameter update through stochastic gradient descent (SGD). The prediction error is represented by a cost function in the form of *cross-entropy loss*, which will be defined in the next section.

4.4.2 Loss Function

Since the frame labels are binary, we know $y \in \{0, 1\}$ and $f(x^{(i)}, m^{(i)}; \theta) \in [0, 1]$. We can model the predictions as a conditional distribution $\Pr(y^{(i)}|x^{(i)}, m^{(i)}; \theta)$:

$$\Pr(y^{(i)}|x^{(i)}, m^{(i)}; \theta) = f(x^{(i)}, m^{(i)}; \theta)^{y^{(i)}} \times (1 - f(x^{(i)}, m^{(i)}; \theta))^{1-y^{(i)}} \tag{4.1}$$

Thus, we can define the likelihood of θ as:

$$L(\theta) = \prod_{i=1}^{N} \Pr(y^{(i)}|x^{(i)}, m^{(i)}; \theta) \tag{4.2}$$

plug Eq. 4.1 in Eq. 4.2, we get:

$$L(\theta) = \prod_{i=1}^{N} f(x^{(i)}, m^{(i)}; \theta)^{y^{(i)}} \times (1 - f(x^{(i)}, m^{(i)}; \theta))^{1-y^{(i)}}$$

Taking the log of likelihood function does not change monotonicity of the function, so optimizing a likelihood function is equivalent to optimizing a log-likelihood function $l(\theta)$:

$$l(\theta) = \sum_{i=1}^{N} y^{(i)} \log(f(x^{(i)}, m^{(i)}; \theta) + (1 - y^{(i)}) \log(1 - f(x^{(i)}, m^{(i)}; \theta)) \quad (4.3)$$

We want to maximise $l(\theta)$, which is effectively minimizing the $-l(\theta)$. Thus, we arrive at the *cross-entropy* function as the cost function. For N samples:

$$\mathcal{J}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} y^{(i)} \log(f(x^{(i)}, m^{(i)}; \theta) + (1 - y^{(i)}) \log(1 - f(x^{(i)}, m^{(i)}; \theta)) \quad (4.4)$$

and the cross-entropy loss for a single sample is:

$$\mathcal{L}(f(x^{(i)}, m^{(i)}; \theta), y^{(i)}) = -(y^{(i)} \log(f(x^{(i)}, m^{(i)}; \theta) + (1 - y^{(i)}) \log(1 - f(x^{(i)}, m^{(i)}; \theta))) \quad (4.5)$$

Thus, the cost function that's being optimized can be expressed as:

$$\mathcal{J}(\theta) = \mathbb{E}_{(x,m,y)\sim\hat{p}_{data}} \mathcal{L}(f(x, m; \theta), y) \quad (4.6)$$

so that when the model is not overfitted, this cost function is not only minimized on the empirical distribution but also on test samples from the data-generating distribution p_{data}:

$$\mathcal{J}^*(\theta) = \mathbb{E}_{(x,m,y)\sim p_{data}} \mathcal{L}(f(x, m; \theta), y) \quad (4.7)$$

We can use gradient-based optimization methods to optimize the cost function through backpropagation.

4.4.3 Stochastic Gradient Descent

We have defined the *cross-entropy loss* from Eq. 4.4 and Eq. 4.6. The gradient of the cost function is:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{(x,m,y)\sim\hat{p}_{data}} \nabla_\theta \mathcal{L}(f(x, m; \theta), y) \quad (4.8)$$

However, computing this expectation exactly is very expensive as it requires evaluating the model on every example in the entire dataset. Optimization

algorithms that use the entire training set are called *deterministic gradient methods* as they process all the training examples simultaneously in a large batch. The entire dataset cannot be fit into a commercial GPU which has a memory ranging from 6-32 GB. Another extreme is the *online method*, which updates the parameters using gradient from a single example at a time. *Online method* is much faster than *deterministic gradient method* but it performs frequent updates with a high variance that cause the objective function to fluctuate heavily, making it harder for the loss function to converge (takes longer time). On the other hand, online method enables it to jump to new and potentially better local minima of the loss function. Empirically, convergence is almost guaranteed if learning rate is slowly decreased. Optimization algorithms used in deep learning falls somewhere in between: it randomly samples a small number of examples from the dataset (*i.e.* a *mini-batch*) then update network parameters using the average over only those examples, which is commonly referred to as *stochastic gradient descent* (SGD):

Algorithm 1 Stochastic Gradient Descent (SGD)

Require:
Learning rate ϵ
Initial parameter θ
1: **while** true **do**
2: Sample a mini-batch of n samples from the training set D_{train} $\{x^{(1)}, ..., x^{(n)}\}$ with corresponding label $y^{(i)}$
3: Compute and accumulate gradients: $\hat{g} \leftarrow +\frac{1}{m}\nabla_\theta \sum_i \mathcal{L}(f(x^{(i)}, m^{(i)}; \theta), y^{(i)})$
4: Apply parameter update: $\theta \leftarrow \theta - \epsilon\hat{g}$
5: **end while**

The crucial parameter in the SGD algorithm is the learning rate ϵ_k where k denotes the number of epochs. Normally, to ensure that learning converges, one applies learning rate decay. One way of learning rate decay is the linear decay where the learning rate linearly decreases till the epoch τ: $\epsilon_k = (1-\alpha)\epsilon_0 + \alpha\epsilon_\tau$ with $\alpha = \frac{k}{\tau}$ and k represents epoch number. However, linear decay requires hand-tuning the beginning and ending learning rates: $\epsilon_0, \epsilon_\tau$. Using non-linear decay, one ensures that ϵ decays by a certain percentage after each training epoch: $\epsilon_k = \epsilon_0 \times \mu^k$, with μ representing the percentage decay for every epoch. Non-linear decay renders

higher learning rate drop at the beginning of training and less decrease towards the end of training, making the training process more stable.

4.4.4 Components of CNNs

In this section, several functions used in convolutional neural networks are described.

2-D Convolution

The convolution operation in 1-D can be written in the form:

$$(x * w)(t) = \int x(a)w(t-a)da. \tag{4.9}$$

where x is the *input* and w the *kernel*. When stored and processed on a computer, data are discretized and we may consider 1-D discrete convolution:

$$(x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a). \tag{4.10}$$

In the case of medical image analysis, the input I is a 2-D array, so a 2-D kernel K is also used:

$$(I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n). \tag{4.11}$$

where i, j defines the horizontal and vertical coordinates on the image. Since convolution operation is commutative, this operation is equivalent to:

$$(K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n). \tag{4.12}$$

However, many neural network libraries, such as the one used for this thesis **TensorFlow**, implements a similar operation **cross-correlation**:

$$(I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n). \tag{4.13}$$

which essentially flips the kernel. Using convolutional kernels instead of fully-connected layers in CNNs is an important improvement: it creates sparse connectivity between the kernel and the input, which significantly reduces the number of parameters that needs to be stored, thus reduces memory requirements and increases computational efficiency.

Max-Pooling

Max-pooling operation outputs the maximum value within a rectangular neighborhood. It can be used as a strong prior to enforce translation invariance, as when input is translated by a small amount, most of the output from max-pooling operation will not change. More importantly, this pooling operation gradually increases the receptive field of each neuron in deep convolutional layers, enabling the network to learn higher level features as the network becomes deeper. A neuron is defined as a convolutional kernel, and the receptive field is defined as the region in the input space that a particular convolutional kernel processes.

Batch Normalization

The use of Batch Normalization [**ioffe2015batch**] is mainly to remove the effect of internal covariate shift, the change in the feature activations to internal layers of a CNN caused by parameter updates in convolutional kernels, so as to accelerate deep network training. Within a deep neural network with l layers where $l > 2$, the existence of high-order effects (up to order l) when updating network parameters through back-propagation contributes to the change in the distribution of network activations. This is problematic, as the change in distribution of activations will cause activation functions to saturate in deeper layers, causing gradients to vanish and significantly slow down convergence. Batch Normalization algorithm tackle this problem through normalizing activations H in each layer into H' by the mini-batch activations' mean μ and standard deviation σ so that activations in each layer follow standard normal distribution:

$$\mu = \frac{1}{m}\sum_i H_i \tag{4.14}$$

and

$$\sigma = \sqrt{\delta + \frac{1}{m}\sum_i (H-\mu)_i^2} \tag{4.15}$$

where the δ is a small positive value to avoid zero standard deviation, so that:

$$H' = \frac{H-\mu}{\sigma} \tag{4.16}$$

This way, updates of the parameters in early layers will not saturate activations in deeper layers as all activation now follow a standard normal distribution. However, normalizing activations in each layer comes at a cost of reducing the expressive power of the neural network. To tackle this problem, the common practice is to allow the activations to have arbitrary means and standard deviations parameterized by γ and β, which are learnable, rather than depending on complicated computations from all previous layers:

$$H \leftarrow \gamma H' + \beta \tag{4.17}$$

Dropout

The main purpose of Dropout is to reduce overfitting of the network. During training, randomly sample neurons are silenced, or masked, simply by multiplying its output to 0, then run forward propagation, back-propagation and parameter update as usual. It provides an inexpensive way to train and evaluate an ensemble of many neural networks derived from a base model. During inference, no neuron is silenced.

These basic components will be used to build SonoEyeNet, as proposed in the next section.

4.5 SonoEyeNet

The objective of SonoEyeNet is to recognize standard abdominal circumference (AC) planes from background planes using sonographer gaze data as an explicit attention mechanism. Different from previous work [**AhmedThesis**] that identifies standard AC planes by first localizing key anatomical structures, SonoEyeNet is trained to find the standard AC plane in an end-to-end manner without the requirement of labels for key anatomical structures. This work investigates different ways of combining gaze-tracking data and deep ultrasound image features in a deep learning network.

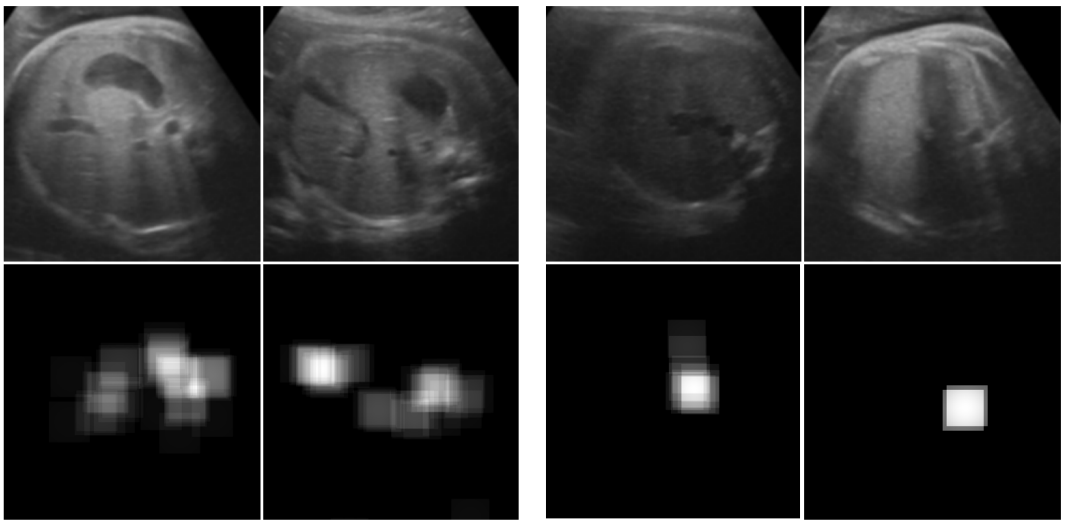


Figure 4.7: (a) Top row shows examples of standard AC planes and the bottom row shows corresponding eye tracking based visual heatmaps (b) examples of background frames and corresponding visual heatmaps.

4.5.1 Methods

Gaze data processing

Using the filtered fixations points, a binary map B of the same dimension as the corresponding frame is generated, with pixels corresponding to a fixation labeled as 1 and others labeled as 0. A sonographer visual attention map A is generated for each binary map by convolving with a truncated Gaussian Kernel $G(\sigma_{x,y})$: $A = B * G(\sigma_{x,y})$, where $G(\sigma_{x,y}) = 30$ *pixels* corresponding to visual angle of 1.5° with an observer-to-screen distance of 0.5 m. Examples of sonographer visual attention maps on both standard AC planes and background planes can be seen in Fig. 4.7.

Baseline Model

The baseline model that is the *SonoNet* [**Baumgartner2017**], which is inspired by *VGG-16* model that consists of 13 convolutional layers and 3 fully-connected layers [**simonyan2014very**]. All layers use 3×3 convolutional kernels with stride of 1, followed by a pooling layer with 2×2 receptive field. Instead of using 2 fully connected layers before output, it uses 2 adaptation layers that convolves the previous feature maps with 1×1 convolutional kernels. Also, it forgoes the max

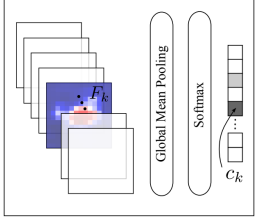
	Feature Extractors	Adaptation Layers	Classification Layer
SonoNet-64	2x[3x3x64/1], MP, 2x[3x3x128/1], MP, 3x[3x3x256/1], MP, 3x[3x3x512/1], MP, 3x[3x3x512/1]	1x[1x1x256/1, 1x[1x1x*K*]	
SonoNet-32	2x[3x3x32/1], MP, 2x[3x3x64/1], MP, 3x[3x3x128/1], MP, 3x[3x3x256/1], MP, 3x[3x3x256/1]	1x[1x1x128/1, 1x[1x1x*K*]	
SonoNet-16	2x[3x3x16/1], MP, 2x[3x3x32/1], MP, 3x[3x3x64/1], MP, 3x[3x3x128/1], MP, 3x[3x3x128/1]	1x[1x1x64/1, 1x[1x1x*K*]	
SmallNet	1x[7x7x32/2], MP, 1x[5x5x64/2], MP, 2x[3x3x128/1]	1x[1x1x64/1, 1x[1x1x*K*]	

Figure 4.8: Baseline SonoNet architectures. It used the following notation: [kernel size × number of kernels / stride]. This figure was reproduced from [**Baumgartner2017**].

pooling layers before adaptation layers; instead, it aggregates spatial features using a global-mean-pooling function on the input feature maps, resulting in a prediction vector that is subsequently fed into a softmax layer. As they were interested in implementing the network in real-time, they explored the effects of reducing the complexity of the network and called these variants *SonoNet-64*, *SonoNet-32*, and *SonoNet-16*. *SonoNet-64* uses the *VGG-16* architecture with 64 kernels in the first convolutional layer. *SonoNet-32* and *SonoNet-16* halved and quartered the number of features maps in each convolutional layer, respectively. Specific network architectures are described in Fig. 4.8. In this work, the architecture of *SonoNet-16* was used to build *SonoEyeNet*, as it was much lightweight and easier to test, given limited computational resource at disposal at the time of this work.

SonoEyeNet architecture

The *SonoEyeNet* aims to take advantage of the deep architecture of *VGG-16* to extract deep features from fetal abdominal US video frames, and at the same time explore possible methods to process the rich information in the gaze-tracking data and its ideal fusion with image features to assist the frame classification task. In this contribution, images and their corresponding heatmaps generated from gaze-tracking data are used as inputs to two streams of CNN before fusing into one stream for frame classification. Two aspects of fusing deep-image features and gaze-tracking data are investigated:

1. how to fuse the two streams;

2. where to fuse the two streams.

Three variants of the *SonoEyeNet* (referred to as *SEN* from this point on) architecture are explored in this chapter:

- *SEN-Concat* (Fig. 4.9);
- *SEN-Later Fusion* (Fig. 4.10);
- *SEN-Early Fusion* (Fig. 4.11).

The ***SEN-Concat*** architecture (Fig. 4.9) feeds an ultrasound video frame as well as the corresponding visual attention map from a sonographer into two parallel streams of CNNs. Both streams have the same architecture: each has 4 convolutional blocks, of which the first two have 2 convolutional layers and the latter has 3 convolutional layers. A max pooling layer using 2×2 pooling is used between each convolutional block. However, parameters for the two streams are initiated differently: the image stream is initiated using the trained parameters of *SonoNet-16*'s first 4 convolutional blocks [**Baumgartner2017**], while the visual attention map stream is initiated randomly from a Gaussian distribution. Image features ϕ_{c4} and visual attention map features each has a shape of $30 \times 30 \times 128$, and after fusion through concatenation the resultant feature maps $\phi^{v}_{c4concat}$ has a shape of $30 \times 30 \times 256$. $\phi^{v}_{c4concat}$ is further processed by max pooling and three convolutional layers before being fed into the adaptation layers, global max pooling layers and softmax layer before classification. The architecture of Adaptation layers can be seen in Fig. 4.8, which uses 1×1 convolutions. The output is a 2-dimensional vector indicating whether the frame belongs to a standard AC plane or background.

Apart from fusion through concatenation, other methods of combining image and gaze-tracking information were investigated. In an early experiment at the beginning of my D.Phil study, edge features were extracted from standard AC planes Fig. 4.12(A) using Edge Box [**zitnick2014edge**]. The edge feature map Fig. 4.12(C) was filtered by a sonographer's visual attention map Fig. 4.12(B) through point-wise multiplication, resulting in a feature maps that clearly highlighted the edges of key anatomical structures, i.e. stomach bubble and umbilical vein, while at the

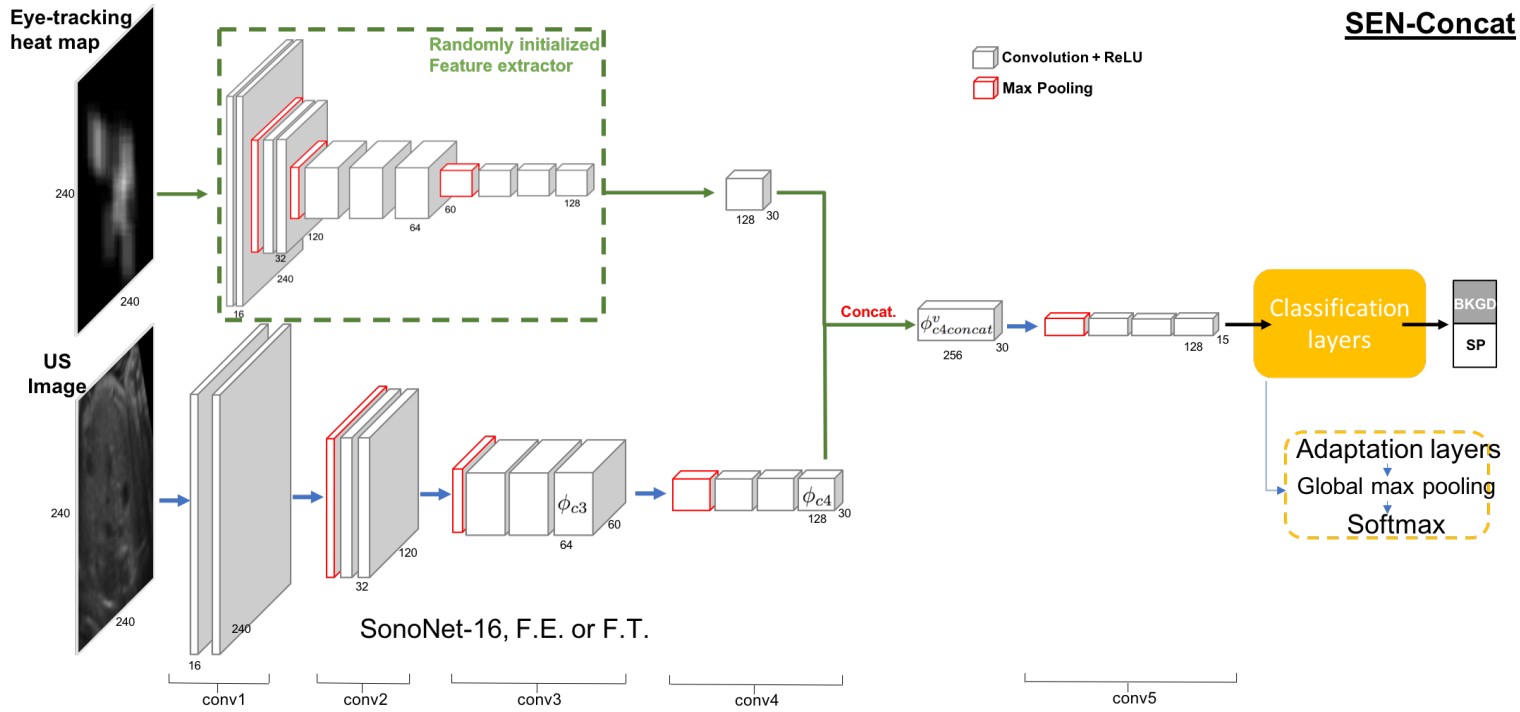


Figure 4.9: The architecture of *SonoNet-Concat*. Features from images and visual attention maps from two streams of CNNs are fused by concatenation.

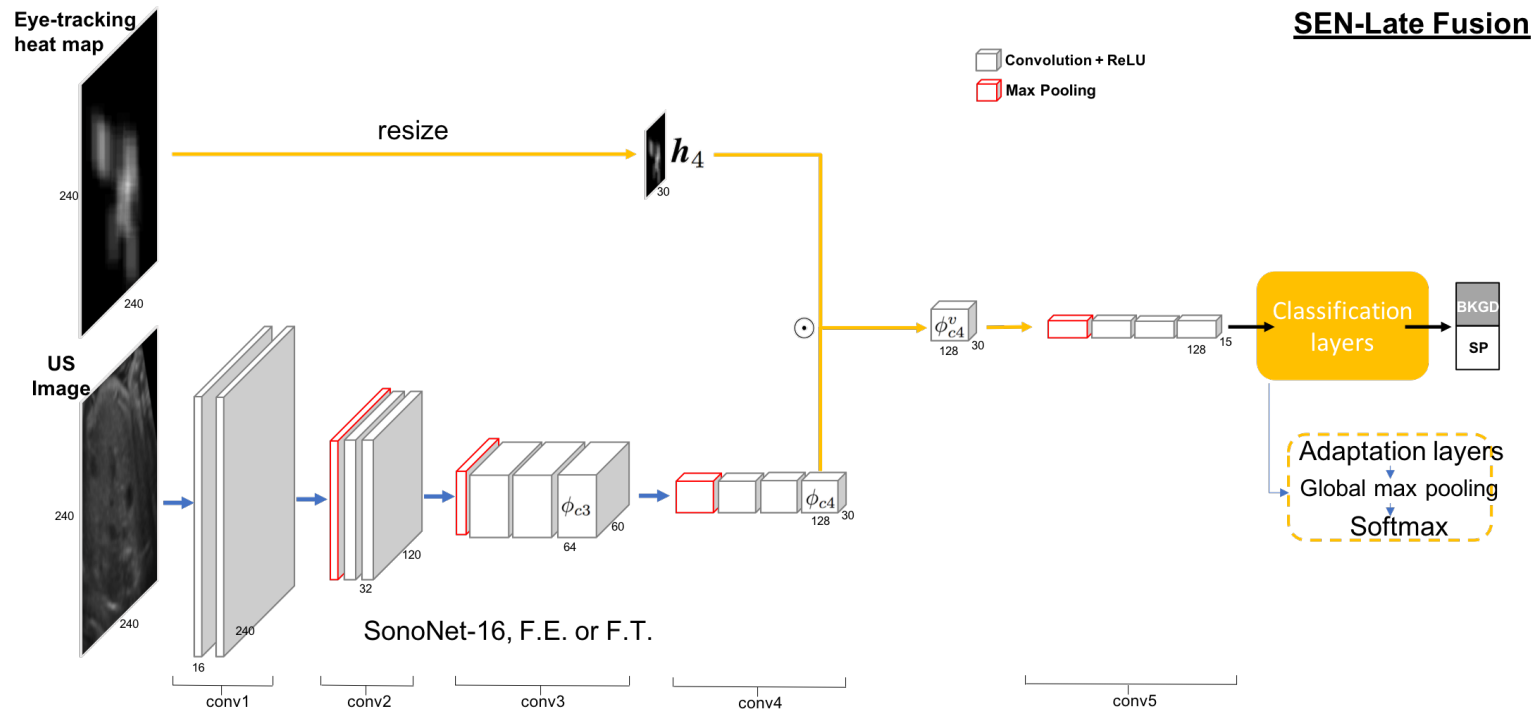


Figure 4.10: The architecture of *SonoNet-Late Fusion*. Feature maps from the 4^{th} convolutional block and the resized corresponding visual attention maps are fused by element-wise multiplication.

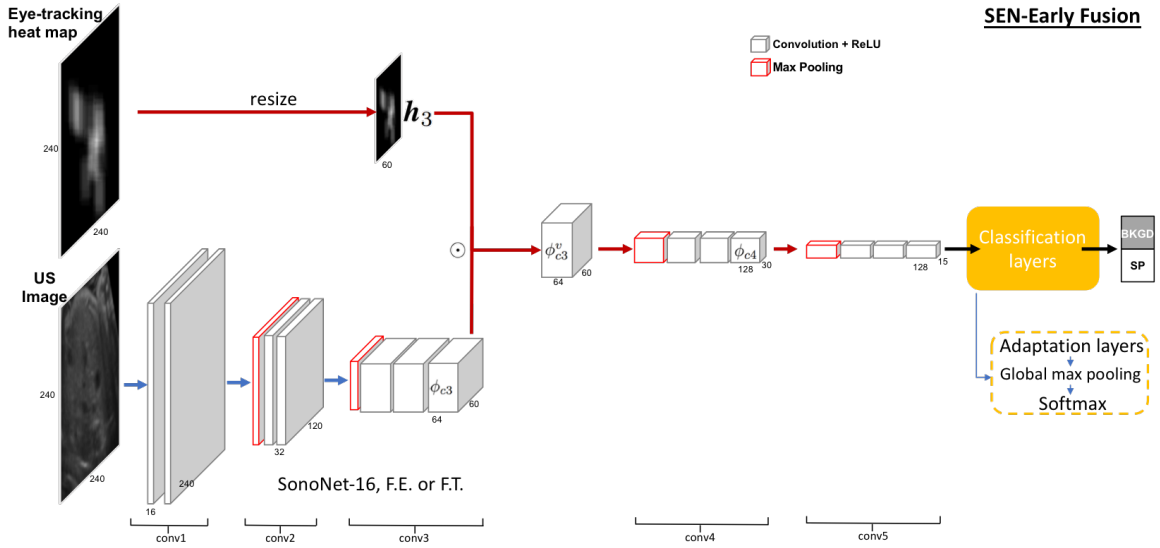


Figure 4.11: The architecture of *SonoNet-Early Fusion*. Feature maps from the 3^{rd} convolutional block and the resized corresponding visual attention maps are fused by element-wise multiplication.

same time reduced signal levels of irrelevent information, e.g. the straight-line edge of US image's field of view Fig. 4.12(D). It is natural to hypothesize that visual attention map will also filter out key information in other features.

Inspired by the edge box experiment, ***SEN-Late Fusion*** and ***SEN-Early Fusion*** explore fusion through element-wise multiplication. In both architectures, the original US image was resized to the shape of corresponding feature to allow for element-wise multiplication, and the operation is very fast and easy to implement under TensorFlow. Fusion at different levels of features was explored: in ϕ^v_{c3}, each element in the feature map corresponds to a receptive field of size 4×4 on the input US image, while in ϕ^v_{c4} each element corresponds to 8×8 receptive field. Element-wise multiplication of the resized visual attention map to feature maps (in ***SEN-Late Fusion***, visual attention map fused with ϕ^v_{c4}, and in ***SEN-Early Fusion*** visual attention map fused with ϕ^v_{c3}) were explored to see which method works better for frame classification.

Model Variants

All three architectures mentioned above all use the pre-trained *SonoNet* weights to initialize the CNN stream for images and kept these weights frozen (*i.e.* not trainable) during training. We call these models ***SEN-Concat***,***SEN-Late FE***, and ***SEN-Early FE***, where *FE* stands for *Feature Extractor*. *SonoNet* architecture is used as a fixed feature extracter in the image stream of *FE* models. However, Gao *et al.* [**Gao2016**] argues that fine-tuning a transferred model renders better performance in classification task. Thus, a variant of ***SEN-Late FE*** was created by allowing the pre-trained stream to be further fine-tuned and we call the model ***SEN-Late FT***.

As a further variant, we investigated how the overall number of trainable weights affects classification performance by removing the convolutional block *Conv5* in the ***SEN-Late FE*** and ***SEN-Late FT*** model, and named them ***SEN-Late FE truncate*** and ***SEN-Late FT truncate***.

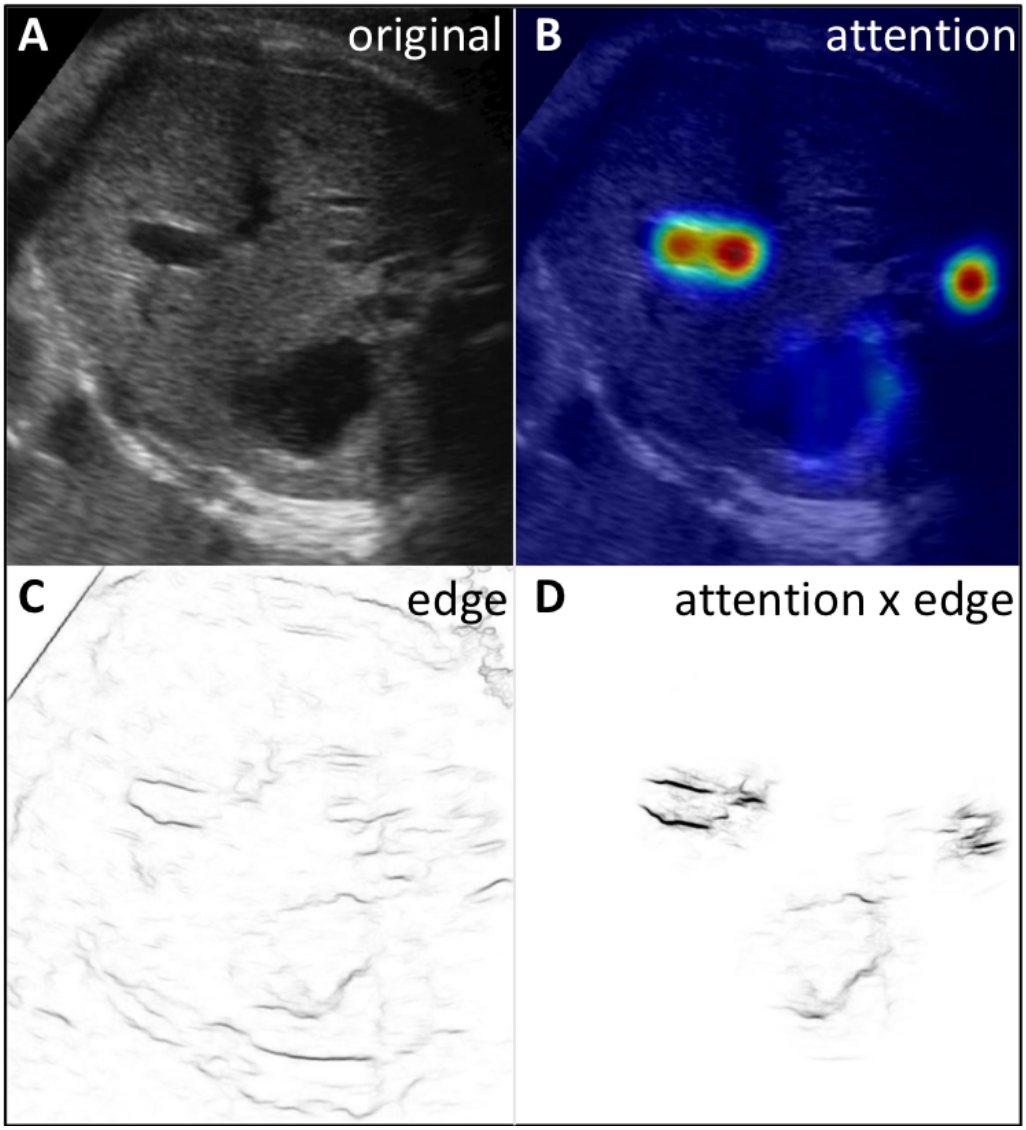


Figure 4.12: Fusion experiment of visual attention map and edge features using element-wise multiplication. (A) Standard AC plane (B) Sonographer visual attention map (C) edge feature map generated by Edge Box [**zitnick2014edge**] (D) Filtered edge feature map

Training details

Models were trained using the adaptive moment estimation (Adam) [**kingma2014adam**] algorithm with a batch size of 256 samples, and an initial learning rate of 0.05. The number of epochs was set to 100. In the stream for visual attention map in ***SEN-Concat***, convolutional kernels were not pre-trained and were initialised from a zero-mean Gaussian distribution with standard deviation of 0.01. Batch

normalization and dropout (rate = 0.5) were used for every convolutional layer before the adaptation layers. The dataset was augmented by rotating each image and its horizontal flip by 45, 90, 135, 180, 225, 270, and 315 degrees; their corresponding visual heatmaps were augmented using the same specifications.

Performance Metrics

The performance of a trained model is determined by how good the predictions reflect the true labels. In the context of ultrasound frame classification, we define standard AC planes as the positive class, and background planes as the negative class. When a standard AC plane is correctly classified, it is defined as a *true positive (TP)*; otherwise a *false negative (FN)*, or a *Type II error*. A correctly classfied background frame is defined as a *true negative (TN)*; otherwise, it is a *false positive (FP)*, or a *Type I error*.

Accuracy is the most intuitive performance measure, following the following equation;

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4.18}$$

However, using *accuracy* as a performance metric is only good for symmetric data sets where the classes are balanced. For unbalanced data, such as the dataset used in this chapter where instances of standard AC planes are fewer than those of background frames, the effect of *false positives* is much smaller than that of *false negatives*. Thus, instead of using *accuracy*, we used three other metrics whic are commonly used in image classification: *Precision, Recall*, and *F1-score*;

$$Precision = \frac{TP}{TP + FP}, \tag{4.19}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4.20}$$

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision}. \tag{4.21}$$

Precision is used to determine the costs of *false positives*, *i.e.* the cases where background frames are wrongly classified as standard AC planes, which potentially provides wrong measurements, leading to wrong diagnosis. ***Recall***, also called

Sensitivity, is used to determine the costs of *false negative*, *i.e.* the case where standard AC planes are not recognized. ***F1-score*** is a way of measuring the test's accuracy, taking into account both *precision* and *recall*.

A ***Receiver Operating Characteristic Curve***, or ***ROC Curve***, illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots *Recall* against *Fall-out*, which is the *false-positive rate* and defined as $FP/(FP+TN)$. ROC analysis provides a tool to select possibly optimal models and to discard suboptimal ones without the need to consider the cost context or class distribution. The closer the Area Under Curve (AUC) of the ROC curve to 1, the more accurate the model.

4.5.2 Results

The baseline model, ***SonoNet-16 FE***, was trained only on image data, and convolutional layers of SonoNet-16 as a fixed feature extractor; adaptation layers were randomly initiated and trained. ***SonoNet-16 FT*** also used only image data but it allowed all the convolutional layers to be fine-tuned during training. After fine-tuning, an increase of precision from 73.6% to 85.1% was observed, indicating fewer background frames were wrongly classified as standard AC planes; however, there was a corresponding decrease in recall from 74.1% to 64.7%, indicating more standard AC planes were classified as backgrounds, as shown in Table. 4.1. The *F1-scores* of both models remained under 75%.

When training with visual heatmaps in tandem with US images, an immediate classification improvement was observed for background frames, as the SonoEyeNet modelled all achieved precisions above 93%. However, ***SEN-Concat*** and ***SEN-Early FE*** did not improve the ACP classification as recall remained at 74.4% and 76.8%, respectively. The first improvement in recall was observed in ***SEN-Late FE*** where recall increased to 91.3%. Best results were achieved by ***SEN-Late FT*** where the branch for image feature extraction was allowed to be further fine-tuned. As seen in Table. 4.1, the model's ability to classify background and ACP was the best amongst all models.

Table 4.1: *Comparative Evaluation of Classification Performance. Column "Eye" indicates whether eye movement data was used. Values in bold correspond to the best results.*

Models	Eye	Precision	Recall	F1-score
SonoNet-16 FE	No	73.6	74.1	73.8
SonoNet-16 FT	No	85.1	64.7	73.5
SEN-Concat	Yes	95.3	74.4	85.4
SEN-Early FE	Yes	93.8	76.8	84.5
SEN-Late FE	Yes	96.1	91.3	93.6
SEN-Late FT	Yes	***96.5***	***99.0***	***97.8***

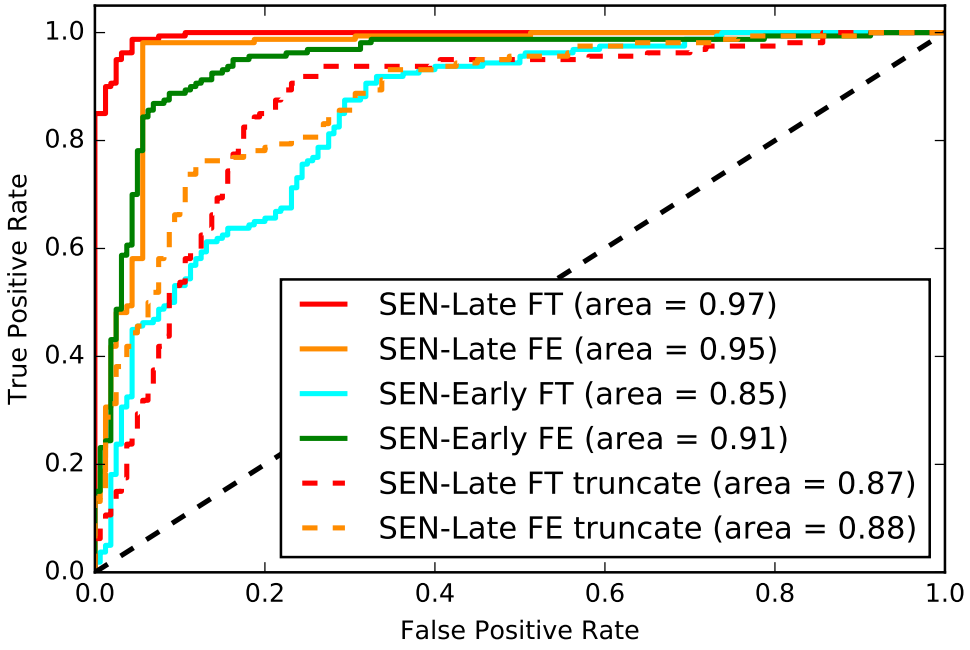


Figure 4.13: *ROC curves of selected SEN models. SEN-Late FT with AUC=0.97 is the best-performing model.*

Receiver Operating Characteristic (ROC) curves for the models are shown in Fig. 4.13. Confirming the findings in Table. 4.1, ***SEN-Late FT*** (red) performs best and is followed by ***SEN-Late FE*** (orange) and ***SEN-Early FE*** (green), with Area Under the Curve (AUC) of their ROC curves of 0.97, 0.95 and 0.91 respectively. The image feature branch in the early fusion model was further fine-tuned in ***SEN-Early FT*** (blue), but a dramatic decrease in performance was observed, as the

AUC of the ROC dropped to 0.86. In addition, ***SEN-Late FE truncate*** (orange dash) and ***SEN-Late FT truncate*** (red dash) were further trained to see whether smaller models with fewer parameters to train perform better than larger ones. However, they were found to be inferior to their close variants ***SEN-Late FT*** and ***SEN-Late FE***, as their corresponding AUC dropped to 0.89 and 0.88, respectively.

4.6 Discussions and Conclusion

It needs to be pointed out that during gaze data collection procedure, the order in which US videos were presented to sonographers were not randomized, so potential bias might be introduced due to observer habituation or fatigue. In addition, as was described earlier the eye-tracker experienced overheating issues after prolonged use, further errors (drift caused by overheating) might be introduced to the gaze tracking data. To mitigate the potential bias and errors, it is worth randomizing the order in which videos will be presented to sonographers in the future if further retrospective gaze-tracking experiments are to be conducted on multiple sonographers.

During training the input images were pre-processed through flipping and rotation with one of the eight pre-set angles, which were chosen empirically to evenly cover 360 degrees of rotation. However, in reality the position of the fetus is random, so the 8 pre-set angles may not provide enough generalization power for training. This pre-processing step can be further extended to have input images randomly rotated by an angle in the interval between 0 and 360 degrees to encourage rotational invariance of the model. This is the augmentation method used in Chapter 6.

Four trends were observed from Table. 4.1 and ROC curves in Fig. 4.13. First, models that use both eye movement data and US image data (***SonoEyeNets***) achieve higher classification accuracy than models trained purely on US image data. Second, element-wise multiplication of a resized visual heatmap and image feature maps performs better (as measured by recall and precision) than concatenation of the feature maps from both image and heatmap branches. Third, fusion of image feature maps and visual heatmaps at later stages (after the 4^{th} convolutional block) achieves better results (as measured by recall and precision) than that at an earlier

stage (after the 3^{rd} convolutional block), indicating higher level image features masked by visual heatmaps are more discriminative. Finally, fine-tuning the image feature branch during training further improves model performance.

The truncated models in the experiments over-fitted to the training set, which is mainly caused by the removal of the convolutional block immediately before the classification layer. The remaining 2 convolutional layers in the classification layer lacked descriptive power to classify fusion feature map ϕ^{v}_{c4}, indicating that the fusion of US image data and eye movement data needs to find a balance: a model that can sufficiently describe image data (higher level features) but at the same time allows more flexibility to handle fused feature maps, even if it means one additional convolutional block.

Due to small data size (33 patients in total, 1616 US images), caution is required in the interpretation of improvements in Precision, Recall and F1-score. For example, correctly predicting only 160 more images (3-4 patients) will cause an improvement of 10% in Precision. With multiple architectures compared, the chance that a model randomly improves prediction on 3-4 patients is increased. In addition, this preliminary result is not cross-validated so the difference between different models can be less dramatic than presented. However, we still consider these results valuable as they lay the foundation for future works, and provides inspirations for future network architecture design.

The results presented demonstrate a novel way to train a classification CNN for fetal US video plane finding using US video and sonographer eye movements on datasets of modest size. Clinically, this work provides a direct demonstration of sonographer visual attention distribution on US frames, and provides an opportunity for sonographers to retrospectively examine their decision-making process to find the standard plane. However, the clinical application of the model is still very constrained, as the model requires both the ultrasound frame and its corresponding sonographer attention maps as inputs. In order to expand its clinical application, the next step is to investigate models that can infer sonographer's visual attention

given an ultrasound video frame, and use the inferred attention maps to assist frame classification. Such investigations will be discussed in the next chapter.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe

5

Visual Attention Prediction and its Application

Contents

5.1 Introduction

It was shown in Chapter 4 that sonographer's visual attention on 2-D fetal abdominal ultrasound video frames provides a very strong prior to assist the task of standard abdominal circumference (AC) plane detection, with visual attention focusing on key anatomical structures such as the stomach bubble, umbilical vein, and spine. However, such a model is constrained in terms of its clinical applicability, as *SonoEyeNet* requires gaze-tracking information for inference. Hence, it is desirable to learn from visual cues contained in ultrasound video frames as well as existing sonographers' gaze data to model which regions in the frames are more likely to be visited. More formally, the objective is to learn a probability distribution of sonographer visual attention conditioned on a specific input ultrasound frame.

Generally, this type of research is called Saliency Prediction [**itti1998model**], which describes models that predict spatial locations in an image that attract human attention. In traditional saliency prediction models, the collection of gaze data requires observers to view images for several seconds each without any particular tasks in mind [**koch1987shifts**, **itti1998model**], reflecting the bottom-up process of human visual attention. However, here, we recognize that top-down features such as context and tasks are also important in modeling human visual attention [**viola2001rapid**, **felzenszwalb2008discriminatively**]. Early works modelled saliency depending on hand-crafted bottom-up features [**koch1987shifts**, **itti1998model**, **itti2001computational**]. Such bottom-up saliency models work well when higher order semantics are reflected in low-level features (as is often the case for isolated objects, and even for reasonably cluttered scenes), but tends to fail if other factors dominate: for instance free-viewing of images without clearly isolated objects [**itti2001computational**].

In this chapter, a baseline saliency prediction model using patch-wise multi-resolution convolutional neural networks inspired by [**liu2015predicting**] is trained on the dataset described in chapter 4. One sonographer's gaze data were used to train the model. The understanding of limitations of the baseline model helps design Multi-task SonoEyeNet, a CNN that learns to predict sonographer visual attention

and use that to assist standard AC plane detection. We further use an adversarial and multi-task training scheme to fine-tune the model and generate smoother visual attention maps. The produced visual attention maps is used as a strong prior for the standard AC plane detection task and expands the potential clinical usefulness of SonoEyeNet. The best model variant is able to achieve Precision of 96.8%, Recall of 96.2% and F-1 score of 96.5%, on par with the classification results of SonoEyeNet but not requiring sonographer gaze-tracking data as an input.

5.2 Originality and Individual Role

Using Keras and TensorFlow, both open source Python libraries, I trained and tested the patch-wise multi-resolution convolutional neural network as well as Multi-task SonoEyeNet to predict sonographer visual attention. Multi-task SonoEyeNet [**cai2018multi**] has been accepted and published in the proceedings of Medical Image Computing & Computer Assisted Intervention (MICCAI) 2018.

5.3 Patch-wise Saliency Prediction Model

Patch-wise model for saliency prediction was designed for two reasons. First, due to computation limitations at the time of the work, it was too expensive to train deep CNNs directly on full-size (559×745) images; constraining input sizes to 42×42 allows for efficient computation. Second, multi-scale input for saliency prediction was proven to increase model performance, with [**he2015spatial**, **farabet2013learning**] using 2 different scales and [**liu2015predicting**] using 3 different scales, allowing the model to simultaneously look at patches with small to large context and fine-grained to coarse detail levels. Because of these two reasons, the patch-wise model was designed and tested.

5.3.1 Model and Training Details

Patch Generation

Fixation data were processed the same way as in Chapter 4. In order to simultaneously learn features at different scales and their integration to predict eye fixations

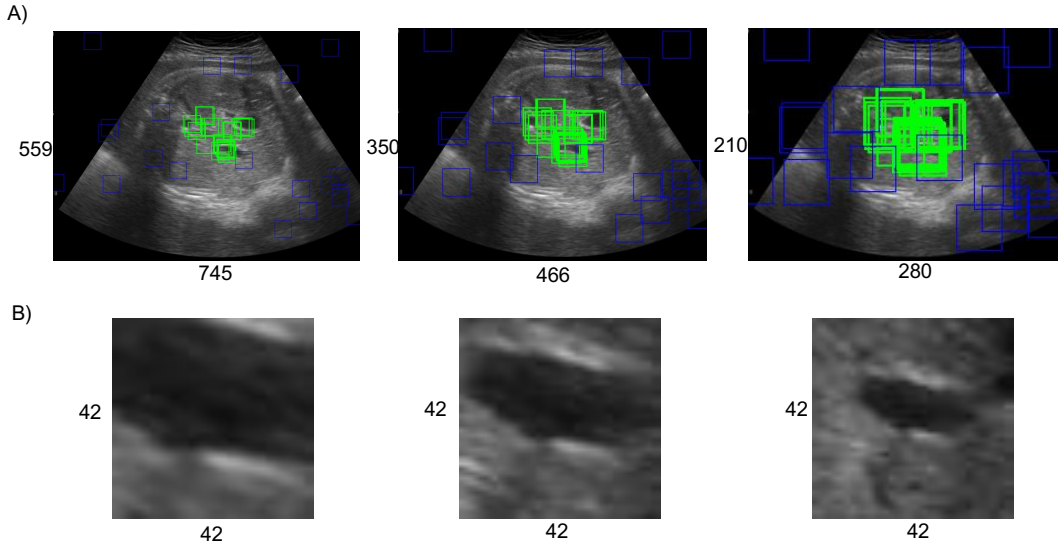


Figure 5.1: The patch generation process for each input image. Green boxes indicate patches centered around fixation points, while red boxes around saccades. (A) overview of how patches are cropped on each scales of image (B) patches generated on the same fixation point but on different scales

on ultrasound images, a multi-resolution CNN architecture inspired by Liu *et al.* [**liu2015predicting**] was built. Ultrasound images were firstly resized to three different resolutions. Image regions of a fixed size centered on the same locations from same image with different scales were extracted and fed into three streams of the CNN. Thus, three image patches contained small-to-large context and fine-to-coarse granularities. The three resolutions used here were 559×745 (original), 350×466 and 210×280. Image patches' sizes were fixed at 42×42. Fixation points and non-fixation points were randomly selected from each image in the training set. This process can be seen in Fig. 5.1. During inference, for each image 50 locations along each dimension (totalling 2500 regions) were evenly sampled. Thus, for the 64 images in the test set, 160,000 image patches were used for test.

Model Architecture

The architecture of the multi-resolution CNN is shown in Fig. 5.2. Each of the three streams contained three Convolutional (C) layers, three Pooling (P) layers, and a Fully Connected (FC) Layer. FC layers from three streams are then fused using another FC layer, which is followed by a logistic regression layer at the

end to perform classification (fixation and non-fixation). Each C layer across three streams share parameters to learn scale-invariant features. The first C layer used 96 convolutional kernels with 7×7 size, while the next two C layers used 160 and 288 convolutional kernels with 3×3 size. Convolution stride was set at 1, and pooling windows were set at the size of 2×2. The output only used 1 neuron to predict the label of inputs. Thus, the whole network has the size of I[$42 \times 42 \times 3(\times 3)$]-C[$36 \times 36 \times 96(\times 3)$]-P[$18 \times 18 \times 96(\times 3)$]-C[$16 \times 16 \times 160(\times 3)$]-P[$8 \times 8 \times 160(\times 3)$]-C[$6 \times 6 \times 288(\times 3)$]-P[$3 \times 3 \times 288(\times 3)$]-FC[$512(\times 3)$]-FC[512]-O[1], as can be seen in Fig. 5.2. The denotation ($\times 3$) means there are 3 duplicates in three streams; I and O indicates input and output.

Loss Function

This model essentially predicts the probability that each patch in a particular image $p^{(j)} \in I$ will be fixated on by the sonographer. The convolutional neural network thus classifies each patch into either a fixation point or non-fixation, and the patch-wise predictions will be reshaped into a down-sampled saliency map of the original image I during inference. For a classification problem, cross-entropy loss is used:

$$l(\theta) = \sum_{j=1}^{N} y^{(j)} \log(f(p^{(j)}; \theta) + (1 - y^{(j)}) \log(1 - f(p^{(j)}; \theta)) \tag{5.1}$$

where N represents the batch number, $y^{(j)} \in \{0, 1\}$ represents the label of patch p_j where 1 represents a fixation and 0 non-fixation, $f(\cdot; \theta)$ represents the a convolutional neural network with parameters θ.

Training Details

All layers were initialized using a zero-mean Gaussian distribution with standard deviation 0.01. The network was trained for 100 epochs with Stochastic Gradient Descent with an initial learning rate 2×10^{-4}. Batch size was set to 64.

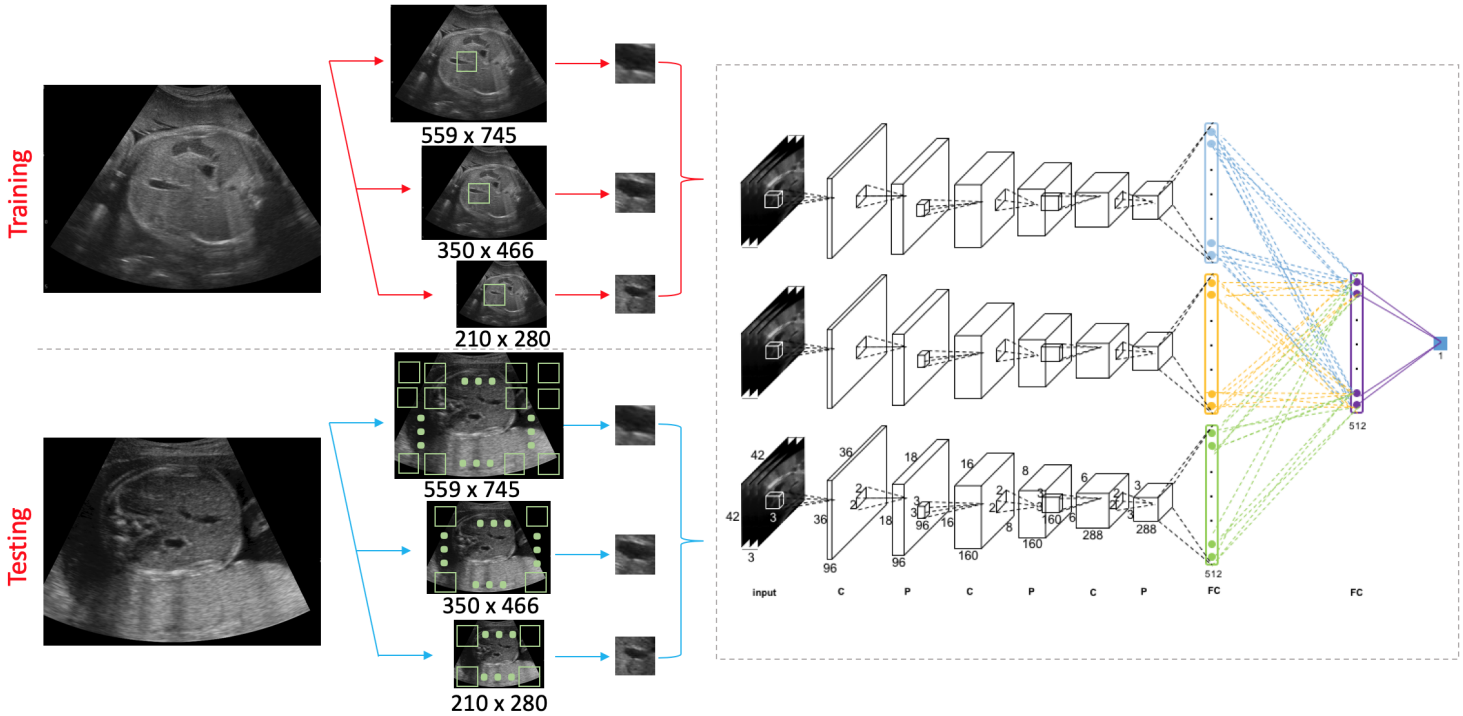


Figure 5.2: The whole process of building the saliency prediction model using CNN and the architecture of the network. During training, images were first rescaled to three resolutions (559×745, 350×466, 210×280), and 42×42 patches around the same fixation/non-fixation points are cropped out and fed into the CNN. Convolutional layer, Pooling layer and Fully Connected layers are denoted as C, P and FC. First three C layers in three streams share parameters.

Performance Metrics

In order to quantitatively measure the quality of predicted visual attention map as compared to ground truth visual attention maps, several metrics are available and for the purpose of this thesis, 6 commonly used metrics were chosen for visual attention map quality measurement.

Area Under ROC Curve (AUC). The ground truth of human visual attention A is represented as a binary map where the value of a pixel represents whether it is a fixation point (1) or not (0). The predicted attention map $\hat{A}$ is treated as a binary classifier, and by varying a discriminative threshold on pixel values this map, a Receiver Operating Characteristic (ROC) curve is generated, showing the trade-off between true and false positives at different thresholds. The Area Under ROC Curve (AUC) measures the quality of the predicted attention map [**bylinskii2018different**]. Higher AUC indicates better quality visual attention map.

Normalized Scanpath Saliency (NSS). NSS [**peters2005components**] is calculated as the average normalized saliency at fixed locations. For ground truth binary fixation map A and predicted visual attention map $\hat{A}$:

$$\begin{aligned} NSS(\hat{A}, A) &= \frac{1}{N} \sum_i \bar{\hat{A}}_i \times A_i \\ \text{where } N = \sum_i A, \quad \bar{\hat{A}}_i &= \frac{\hat{A} - \mu(\hat{A})}{\sigma(\hat{A})} \end{aligned} \tag{5.2}$$

where i is the index of the i^{th} pixel, N is the total number of fixated pixels. μ and σ represents the mean and standard deviation, respectively. Higher NSS score indicates better quality visual attention map.

Information Gain (IG). Given a ground truth binary map of fixations A and predicted visual attention map $\hat{A}$, the Information Gain score [**kummerer2014close**] is calculated based on a baseline map B, which represents the center prior of human observers:

$$IG(\hat{A}, A) = \frac{1}{N} \sum_i A_i [log_2(\epsilon + \hat{A}_i) - \log_2(\epsilon + B_i)] \tag{5.3}$$

where i is the index of the i^{th} pixel, N is the total number of fixated pixels, and ϵ is a regularization term for numerical stability. A score above zero indicates the predicted visual attention map $\hat{A}$ predicts fixation locations better than the center prior baseline B.

Similarity (SIM). SIM [**rubner2000earth**] is a distribution-based metric and views both the ground truth visual attention map A and predicted visual attention map $\hat{A}$ as distributions. It measures the similarity between these two distributions:

$$\begin{aligned} SIM(\hat{A}, A) &= \sum_i min(\hat{A}_i, A_i) \\ \text{where} \sum_i \hat{A}_i &= \sum_i A_i = 1 \end{aligned} \tag{5.4}$$

where i is the index of the i^{th} pixel. A SIM of one indicates the distributions are the same; SIM of a zero indicates no overlap.

Pearson's Correlation Coefficient (CC). CC [**le2007predicting**] evaluates the linear relationship between two distributions, the ground truth visual attention map A and predicted visual attention map $\hat{A}$:

$$CC(\hat{A}, A) = \frac{\sigma(\hat{A}, A)}{\sigma(\hat{A}) \times \sigma(A)} \tag{5.5}$$

where $\sigma(\hat{A}, A)$ is the covariance of $\hat{A}$ and A. High positive CC values occur at locations where both the ground truth and predicted visual attention map have values of similar magnitudes.

5.3.2 Results

Examples of predicted visual attention maps generated by the patch-based saliency prediction model on test set US video frames can be seen in Fig. 5.3. Each column represents the input ultrasound images, ground-truth fixation maps and visual attention maps predicted by the patch-wise saliency model, respectively. Example results of two standard AC planes as well as two background planes are shown. As observed from the examples presented, generated attention maps displays high level of uncertainty with regards to predicted visual fixation. Attention spreads out across the visual scene in each image. Though attention covers major anatomical land

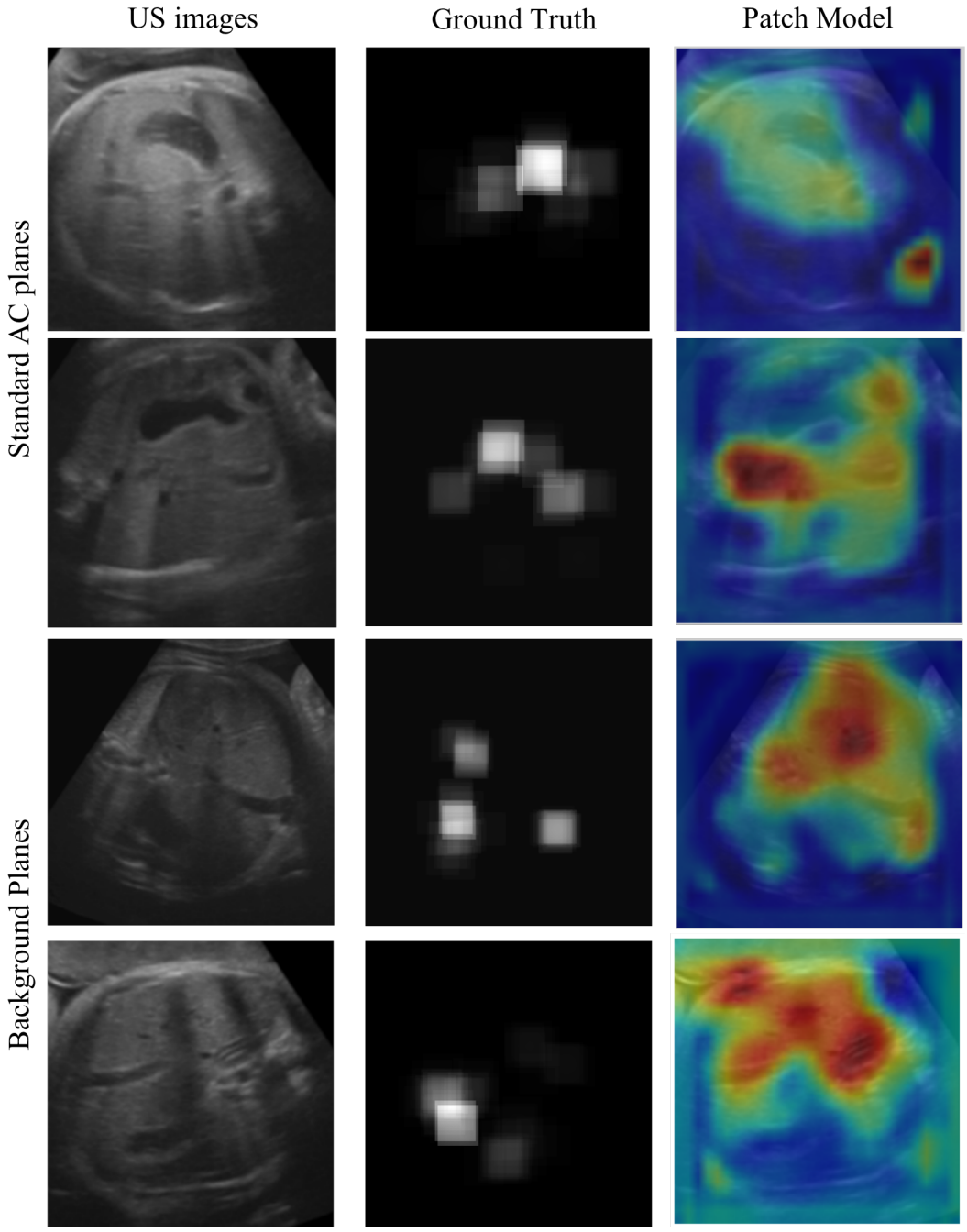


Figure 5.3: Attention maps generated by the patch-based saliency prediction model. From left to right: US image, ground truth (Sonographer's actual attention map), and attention maps predicted by the patch-basd saliency prediction model.

marks, *i.e.* stomach bubble, umbilical vein, and spine, attention peaks predicted by this model do not overlap with the ground-truth fixation points. In order to quantitatively measure the quality of predicted visual attention maps, 5 different metrics were used and the results can be seen in Table. 5.1.

Table 5.1: *Quantitative Analysis of the predicted visual attention maps.*

Models	IG	CC	NSS	SIM	AUC
Patch Model	0.082	0.403	0.741	0.273	0.712
Rand.	-1.410	-0.031	-0.321	0.105	0.619

The saliency scores of the heatmaps generated by the patch model was compared to the results of heatmaps with randomly generated values (*Rand.*) drawn from a standard normal distribution at each pixel. As can be seen in Table. 5.1, the Patch-based model's scores are higher than those of the randomly generated heatmaps on every metric. However, these scores are still relatively low. For example, the IG score is smaller than 0.1, AUC score below 0.8 and SIM score below 0.3 all indicates the poor quality of the predicted attention maps.

5.3.3 Discussions

Even though the patch-based saliency prediction model outperforms randomly generated heatmaps on every saliency metrics, the value for each metric is very low. For example, an IG score of 0.082 indicates that the predicted saliency maps barely outperform central prior, meaning the predicted visual attention is not much better than simply predicting the center of each image as the fixation point.

Possible reasons for the limited visual attention capability are two-fold. First, visual attention is very task-specific, and simply predicting visual attention without any knowledge of the task is difficult. Second, small patch size limits the receptive field of the model and the network only processes a small region on the image. As mentioned in Chapter 2, a very important step in human visual search is the *global impression* [**Nodine1987**], a quick summary of the global context in the image that provides the "top-down" information before a person decides which areas to focus on in a sequential manner. Without global information, it is hard to model human visual attention. Though the patch-based model is computationally efficient and can be trained on a CPU, this gain in computation speed was at the expense of global information contained in the whole image.

Thus, two key problems should be solved to predict sonographer visual attention:

1. how to utilize task-specific information in training the network,
2. how to utilize global information in an image while not to lose computational efficiency.

It is hypothesized that the first problem can be solve through training visual attention modelling together with the task of ultrasound image classification in a multi-task learning fashion, and the second problem can be solved by using GPU to accelerate training so that the whole image, rather than patches of the image, is processed by the CNN model.

5.4 Multi-task SonoEyeNet

5.4.1 Introduction: Multi-task Learning

As mentioned earlier in the chapter, human visual attention not only relies on bottom-up information contained within input images, but also highly depends on top-down knowledge such as context and tasks. In the case of sonographer visual attention prediction on fetal abdominal 2-D ultrasound images, it will be desirable to utilize not only the collected gaze tracking information but also the result of the sonographer's main task, which is to determine a frame-wise class label, *i.e.* decisions that each frame belongs to the class of standard AC plane or a background, to assist visual attention modeling. This falls into the category of a field of research called multi-task learning.

Caruana [**caruana1997multitask**] defines multi-task learning as "an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias". Here, inductive bias works as a type of network regularization. For example, a l_1 regularization is a very common type of inductive bias, which enforces the network to be sparse. In the context of multi-task learning, auxiliary tasks are used as the additional inductive bias. By acquiring a shared representation through parallel learning of

several tasks, trained model is more generalizable and less prone to over-fitting [**caruana1997multitask**].

Two types of multi-task learning are normally used in image-based CNNs [**caruana1997multitask**]. One type uses soft parameter sharing [**duong2015low**, **yang2016trace**], which uses different neural networks for different tasks, but the distance between parameters of different models are regularized to be similar, as can be seen in Fig. 5.4. Another type of multi-task learning uses hard parameter sharing, which shares hidden layers for all tasks but branches out at higher levels for different tasks, as can be seen in Fig. 5.4. This is the more common type of multi-task learning in deep image-based networks [**long2015learning**, **kendall2017multi**, **yang2016deep**].

In this chapter, we adopted the hard parameter sharing multi-task learning scheme, which is most commonly used for computer vision problems as it is computationally and memory-wise more efficient because of the shared lower-level convolutions. We build a framework called *Multi-task SonoEyeNet (M-SEN)* , which builds on SonoEyeNet and trains two tasks in parallel: frame classification as well as sonographer visual attention prediction.

5.4.2 Model and Training Details

Architecture

The *M-SEN* architecture is summarized in Fig. 5.5. It is a multi-task convolutional neural network that can be trained to both generate a visual attention map $\hat{A}$ and predict a classification score vector of the input frame $\hat{y}$:

$$\hat{y} = G(I; \theta_G)_1 \tag{5.6}$$

$$\hat{A} = G(I; \theta_G)_2 \tag{5.7}$$

where G represents the the M-SEN network, subscription indicates the index of network output, I represents the US video frame and θ_G represents weights of the M-SEN network. First, image features are extracted using the first three convolutional blocks of a pre-trained SonoNet [**Baumgartner2017**]. All layers use

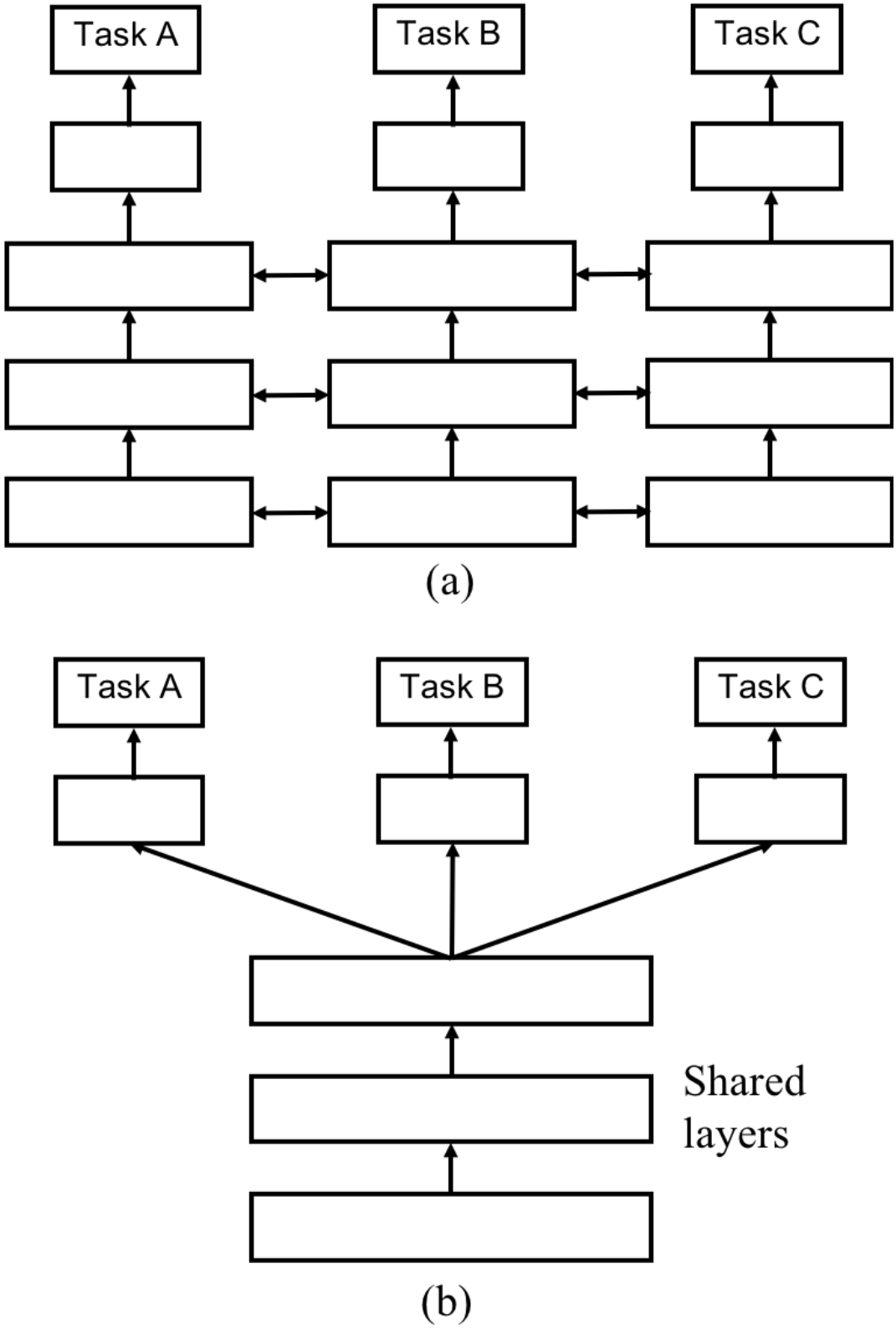


Figure 5.4: (a) Soft parameter sharing in a neural network for multi-task learning; (b) Hard parameter sharing in a neural network for multi-task learning. This figure was reproduced from [**caruana1997multitask**].

3×3 convolutional kernels and the number of kernels used can be seen in Fig. 5.5. Feature maps are down-sampled by a factor of 2 after each block using max pooling. The network separates into two branches after the third convolutional block: one branch for the auxiliary task of saliency prediction to mimic sonographer visual attention on a US video frame; the other for frame classification. Feature maps ϕ_{c3} are first spatially down-sampled by 2 and then passed through 3 convolutional layers with 3×3 kernels in each branch. This produces ϕ_{c4S} and ϕ_{c4C} for saliency prediction and classification, respectively. Convolution with 1×1 kernels is performed on ϕ_{c4S} to generate $\hat{A}$.

The attention map $\hat{A}$ is then fused with ϕ_{c4C} through element-wise multiplication. The resultant feature maps are passed through another convolutional block, and then through two adaptation layers [**Baumgartner2017**] which used 256 and two 1×1 kernels respectively. Global average pooling on the two resultant feature maps is performed before softmax so as to predict class scores for the standard AC plane and background. Classification loss L_C is defined between the predicted class scores and actual label, and saliency loss L_S between the predicted and the actual sonographer visual attention maps.

Loss Functions

The loss function is set as a linear combination of L_S and L_C, with different weightings λ_1 and λ_2 assigned to two losses:

$$\mathcal{L} = \lambda_1 L_C + \lambda_2 L_S \tag{5.8}$$

where L_S represents the saliency prediction loss and L_C the classification loss.

Classification Loss. Similar to the previous chapter, *cross-entropy loss* was used for L_C. For a batch of images $\{I^{(1)}, ..., I^{(N)}\}$ with corresponding class labels $\{y^{(1)}, ..., y^{(N)}\}$:

$$L_C = -\sum_{j=1}^{N} y^{(j)} \log(G(I^{(j)}; \theta_G)_1 + (1 - y^{(j)}) \log(1 - G(I^{(j)}; \theta_G)_1 \tag{5.9}$$

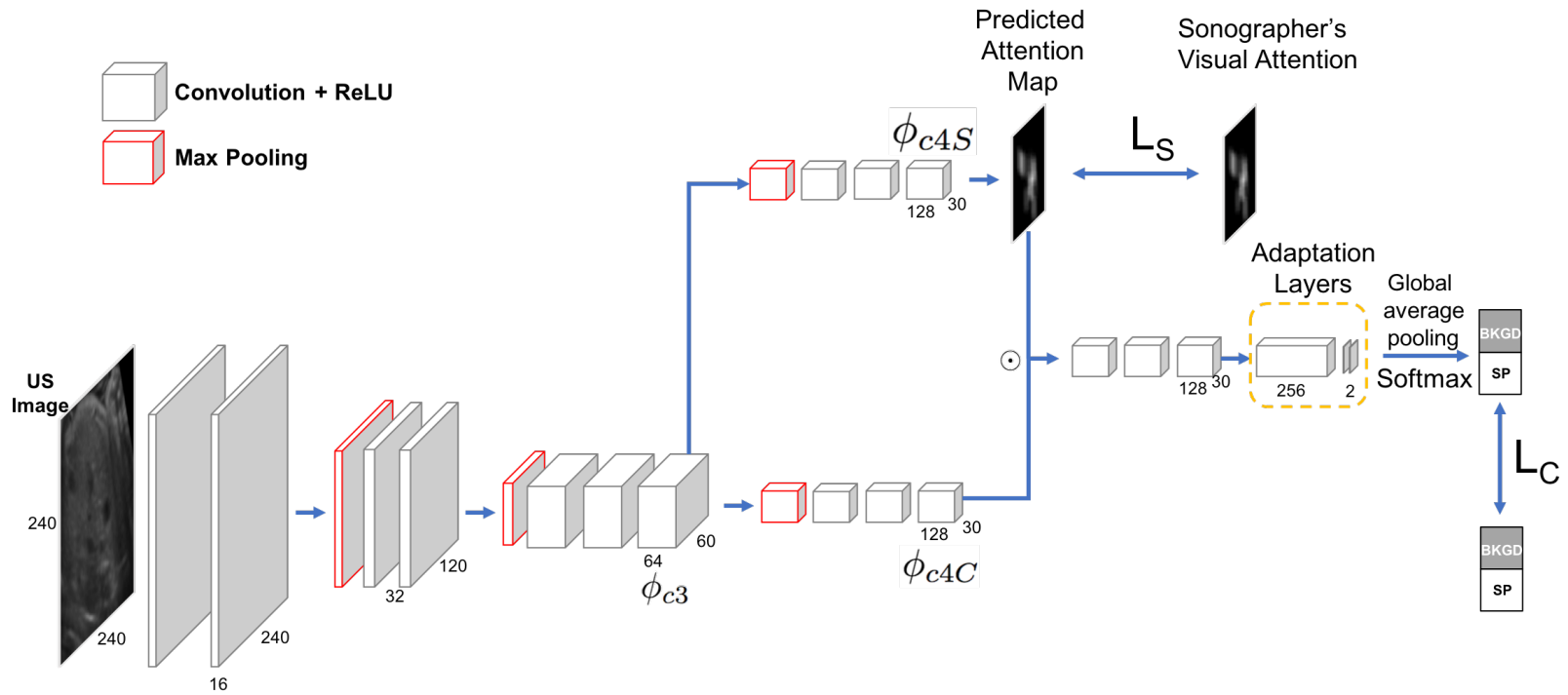


Figure 5.5: Architecture of the multi-task SonoEyeNet (*M-SEN*). The network is trained on two tasks: a primary task to classify frames (bottom) and an auxiliary task to predict visual attention map ($\hat{A}$). The dotted circle $\odot$ indicates element-wise multiplication. L_S and L_C represent the losses of saliency prediction and frame classification.

where G represents the the M-SEN network, $G(\cdot;\theta_G)$ indicates the class prediction vector, and θ_G represents weights of the generator network.

Saliency Prediction Loss. We tried three different forms of loss function for the saliency prediction loss L_S: *Mean Square Error (MSE)*, *Intersection over Union (IoU)* loss and *Kullback-Leibler Divergence (KLD)*. Let $\hat{A}^{(j)}$ be the predicted attention map $G(I^{(j)};\theta_G)_2$.

For a batch of images $\{I^{(1)},...,I^{(N)}\}$ with corresponding visual attention ground truth $\{A^{(1)},...,A^{(N)}\}$, the MSE loss is defined by:

$$L_{MSE} = \frac{1}{N}\sum_{j=1}^{N}\left(A^{(j)} - \hat{A}^{(j)}\right)^2 \tag{5.10}$$

For the same batch of images, the intersection of a ground-truth attention map $A^{(j)}$ and $\hat{A}^{(j)}$ is:

$$Intersection(A^{(j)},\hat{A}^{(j)}) = \sum_a\sum_b (A^{(j)}_{a,b}\cdot\hat{A}^{(j)}_{a,b}) \tag{5.11}$$

The union of the two can be calculated as:

$$Union(A^{(j)},\hat{A}^{(j)}) = \sum_a\sum_b A^{(j)}_{a,b} + \sum_a\sum_b \hat{A}^{(j)}_{a,b} - Intersection(A^{(j)},\hat{A}^{(j)}) \tag{5.12}$$

so the IoU Loss is defined as:

$$L_{IoU} = 1 - \frac{Intersection(A^{(j)},\hat{A}^{(j)})}{Union(A^{(j)},\hat{A}^{(j)})} \tag{5.13}$$

In addition, it is most common to use the Kullback-Leibler divergence to measure the difference between two distributions $A^{(j)}$ and $\hat{A^{(j)}}$:

$$\begin{aligned} D_{KL}(A||\hat{A}) &= -\sum_a\sum_b A^{(j)}_{a,b} log\frac{\hat{A}^{(j)}_{a,b}}{A^{(j)}_{a,b}} \\ &= -\sum_a\sum_b A^{(j)}_{a,b} log\hat{A}^{(j)}_{a,b} + \sum_a\sum_b A^{(j)}_{a,b} logA^{(j)}_{a,b} \end{aligned} \tag{5.14}$$

As can be observed, the first term is the cross-entropy, while the second term the entropy of $A^{(j)}$, which is not related to $\hat{A}^{(j)}$. In this case, optimizing a Kullback-Leibler divergence is equivalent to optimizing the cross-entropy between $A^{(j)}$ and $\hat{A}^{(j)}$. Each pixel can be modeled as a binary random variable, with $\hat{A}^{(j)}_{a,b}$ representing

the predicted probability that the pixel will be fixated on, and $1-\hat{A}^{(j)}_{a,b}$ the probability that it will not. Thus optimizing the KLD loss is equivalent to optimizing a Binary Cross Entropy (BCE) Loss:

$$L_{BCE} = -\sum_{a}\sum_{b} A^{(j)}_{a,b} log \hat{A}^{(j)}_{a,b} + \sum_{a}\sum_{b}(1 - A^{(j)}_{a,b}) log(1 - \hat{A}^{(j)}_{a,b})) \tag{5.15}$$

Training Details

The network was initialized using the first three convolutional blocks of SonoNet as a feature extractor; all other layers were initialized using a zero-mean Gaussian distribution with standard deviation 0.01. Batch normalization and dropout (rate = 0.2) were used for each convolutional layer before the adaptation layers. The weight λ_1 was dynamically changed from 2 to 1 over epochs so as to allow the generator to focus on learning attention maps first and then frame classification. The weight λ_2 was set to 1, as classification was the primary task of the network. The network was trained using Stochastic Gradient Descent (SGD) for 100 epochs with an learning rate 2×10^{-4} to avoid over-fitting. Batch size was set to 64. Five-fold cross-validation was used.

5.4.3 Results

Saliency Prediction Performance

Examples of predicted visual attention maps are presented in Fig. 5.6. All *M-SEN* models trained with different loss functions showed a higher level certainty in terms of visual attention prediction. Predicted visual attentions are to different extents more focused to key anatomical structures compared to the patch-based model, as shown at the bottom of the figure. Among all the *M-SEN* models, *M-SEN BCE* and *M-SEN MSE* models show higher level of attention overlap with the ground truth as compared to the *M-SEN IoU* model, which predicts fixations with high certainty to 50 to 90 percent of the entire area within the abdominal wall. The *M-SEN MSE* model predicts more realistic attention maps, as the areas around fixation points show Gaussian-like distribution of attention, which is more similar to the ground truth. *M-SEN BCE* arguably overlaps more with the ground truth

Table 5.2: *Quantitative Analysis of the predicted visual attention maps.*

Models	IG	CC	NSS	SIM	AUC
MSEN-BCE	0.429	0.615	2.144	0.469	0.726
MSEN-MSE	0.288	0.556	2.253	0.310	0.603
MSEN-IoU	0.275	0.584	1.679	0.305	0.714
Patch Model	0.082	0.403	0.741	0.273	0.712
Rand. Init.	-1.410	-0.031	-0.321	0.105	0.619

Table 5.3: *Comparative evaluation of classification performance. In column "Inputs", "I"refer to US images.*

Models	Inputs	Precision	Recall	F1-score
M-SEN BCE	I	96.7	90.5	93.5
M-SEN MSE	I	92.4	75.6	83.2
M-SEN IoU	I	86.5	80.4	83.3

but the attention maps are not very realistic. A quantitative analysis of the quality of attention maps generated by different model can be seen in Table. 5.2.

As can be seen from the table, all *M-SEN* models performed significantly better in all metrics comparing to the previous patch-based saliency model, which corresponds to the conclusion reached through visual inspection. *M-SEN BCE* models scores highest in 4 out of 5 saliency metrics except for NSS, on which *M-SEN MSE* model performs the best. *M-SEN IoU* performs better than *M-SEN MSE* on CC and AUC, but it performs the worst in the other three metrics among all *M-SEN* models.

Frame Classification Performance

All *M-SEN* models are trained to both predict visual attention and classify frames into standard AC planes or background, and the frame classification results can be seen in Table. 5.3. As can be seen, *M-SEN BCE* is the best at frame classification, reaching an F1-score of 93.5. *M-SEN MSE* and *M-SEN IoU* have similar F1-scores. *M-SEN MSE* has higher precision for standard AC planes, while *M-SEN IoU* has higher recall.

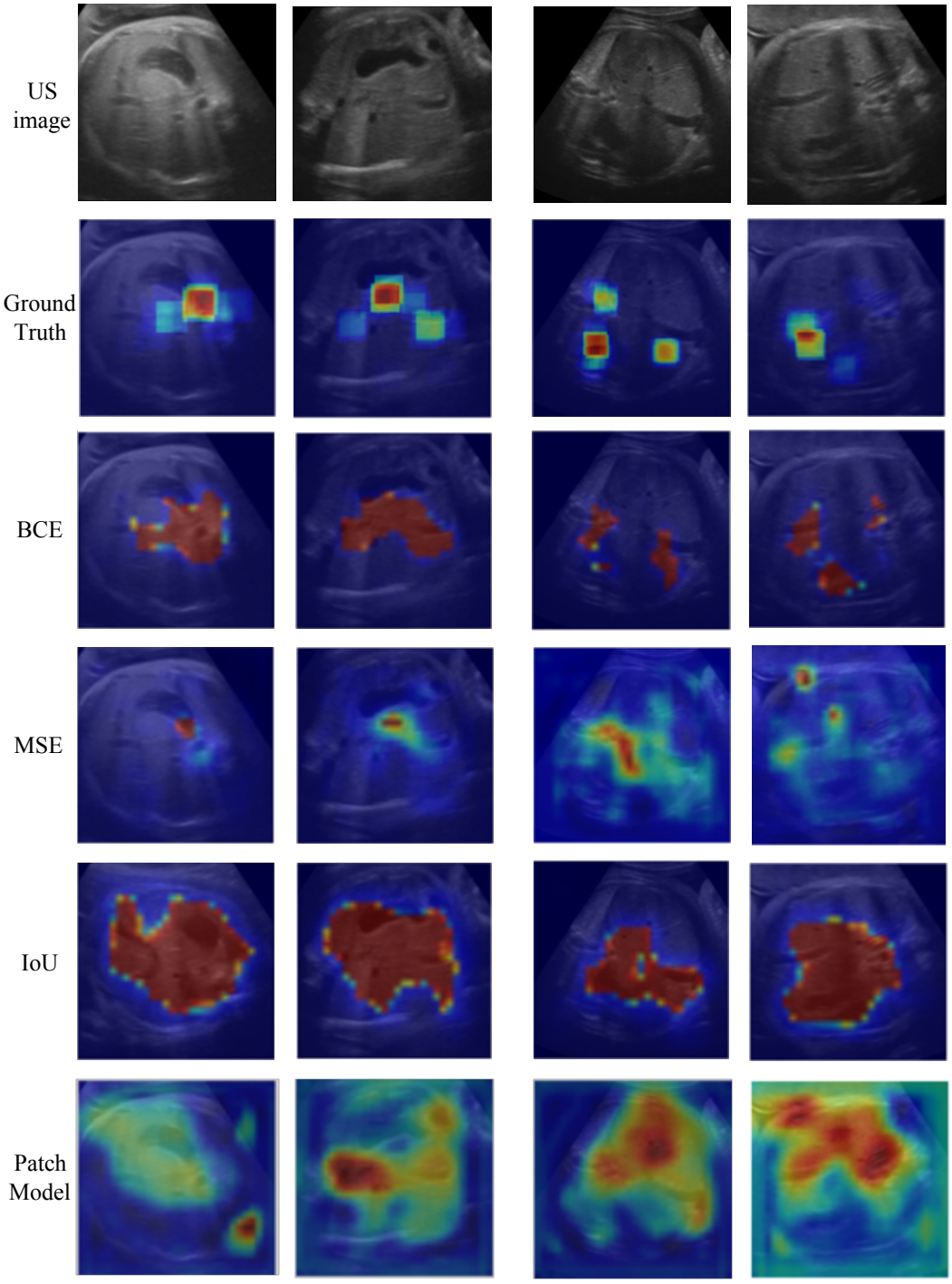


Figure 5.6: Attention maps generated by different variations of the Generator. From top to bottom: US image, ground truth (Sonographer's actual attention map), *M-SEN BCE*, *M-SEN MSE*, *M-SEN IoU*, *Patch Model*.

5.4.4 Discussions

M-SEN models demonstrated the power of multi-task training: all models outperformed previous patch-based saliency prediction model in the saliency prediction task by training visual attention prediction and frame classification together. Not only were the predicted attention more focused on key anatomical structures, they were more similar to the ground truth. *M-SEN IoU* predicts on all regions within the abdominal wall; *M-SEN BCE* and *M-SEN MSE* predicts saliency maps that overlaps more with the ground truth visual attention map, as shown in Table.5.2. Based on the information from Table. 5.3 and Table.5.2, it is observed that the model that can mimick human visual attention maps better (*i.e.* achieves higher saliency scores in Table.5.2), *M-SEN BCE*, also performs better on frame classification task. It is thus hypothesized that by further improving model performance on saliency prediction, thus generating visual attention maps that assign higher weights to regions of US image relevant to standard planes detection, the model performance on frame classification will also be improved. The next section is inspired by generative adversarial network [**goodfellow2014generative**] and introduces an adversarial regulariser network to regularise the predicted visual attention map to enforce it to become more similar to human visual attention (thus have higher saliency scores). It is very interesting to see if the introduced adversarial regulariser, which only optimises visual attention maps, will have any positive impact on frame classification task.

5.5 Multi-task SonoEyeNet with Adversarial Regulariser

5.5.1 Introduction

Generative Adversarial Network

The basic idea of a generative adversarial network (GAN) is to set up a mini-max game between two differentiable functions: the **generator (G)**, parameterised by $\theta^{(G)}$, and the **discriminator (D)**, parameterised by $\theta^{(D)}$. A vanilla GAN's generator

takes latent variables Z as input; the discriminator takes both the observed variables x and the generator's output $\hat{x}$ as inputs. The discriminator learns using traditional supervised learning techniques by classifying inputs into two classes: real or fake. The generator learns to generate samples $\hat{x}$ as close to observed variables x as possible; the discriminator learns to distinguish $\hat{x}$ from x.

The loss function of D is a standard binary cross-entropy loss that's minimized when training a binary binary classifier with a sigmoid output:

$$J^{(D)}(\theta^{(D)}, \theta^{(G)}) = -\frac{1}{2}\,\mathbb{E}_{x \in p_{data}} \log D(x) - \frac{1}{2}\,\mathbb{E}_z \log(1 - D(G(z))) \tag{5.16}$$

The simplest version of loss function for G is assuming a **zero-sum game**:

$$J^{(G)} = -J^{(D)} \tag{5.17}$$

However, this formulation of generator loss function does not perform well in practice. Specifically, when the discriminator minimises a cross-entropy loss (*i.e.* successfully rejects generator samples with high confidence), the generator's gradient vanishes. In practice, generator loss function could be constructed by flipping the target used so that the generator maximises the log-probability of the discriminator being mistaken:

$$J^{(G)}(\theta^{(D)}, \theta^{(G)}) = -\frac{1}{2}\,\mathbb{E}_z \log(D(G(z))) \tag{5.18}$$

Since the generator and discriminator are two networks trained sequentially and iteratively in an adversarial manner, losses are volatile during training. Previous literature suggests using optimization algorithms with adaptive learning rates can accelerate learning; Adam optimizer [**kingma2014adam**] appeared to be the most commonly optimizer used for GAN training.

In this work, the M-SEN model plays the role of a generator that generates predicted attention map. An adversarial regulariser, a discriminator network, is added to further fine-tune visual attention map. In each training step, the discriminator and generator are trained sequentially.

5.5.2 Model and Training Details

The *M-SEN* architecture consists of two CNN modules: the generator (G) and the discriminator (D). It is summarized in Fig. 5.7.

Generator Architecture

G is a multi-task module that can be trained independently without discriminator D to generate both a predicted visual attention map $\hat{A}$ and a classification score vector of the input frame $\hat{y}$: $(\hat{A}, \hat{y}) = G(I; \theta_G)$, where I represents the US video frame and θ_G represents weights of the generator network. First, image features are extracted using the first three convolutional blocks of a pre-trained SonoNet [**Baumgartner2017**]. All layers use 3×3 convolutional kernels and the number of kernels used can be seen in Fig. 5.7. Feature maps are down-sampled by a factor of 2 after each block using max pooling. The network separates into two branches after the third convolutional block: one branch for the auxiliary task of saliency prediction to mimic sonographer visual attention on US video frames; the other for frame classification. Feature maps ϕ_{c3} are first spatially down-sampled by 4 and then passed through 3 convolutional layers with 3×3 kernels in each branch. This produces ϕ_{c4S} and ϕ_{c4C} for saliency prediction and classification respectively. Convolution with 1×1 kernels is performed on ϕ_{c4S} to generate $\hat{A}$. The attention map $\hat{A}$ is then fused with ϕ_{c4C} through element-wise multiplication. The resultant feature maps are passed through another convolutional block, and then through two adaptation layers [**Baumgartner2017**] which used 256 and two 1×1 kernels respectively. Global average pooling on the two resultant feature maps is performed before softmax so as to predict class scores for the standard AC plane and background. Classification loss L_C is defined between the predicted class scores and actual label, and saliency loss L_S between the predicted and the actual sonographer visual attention maps.

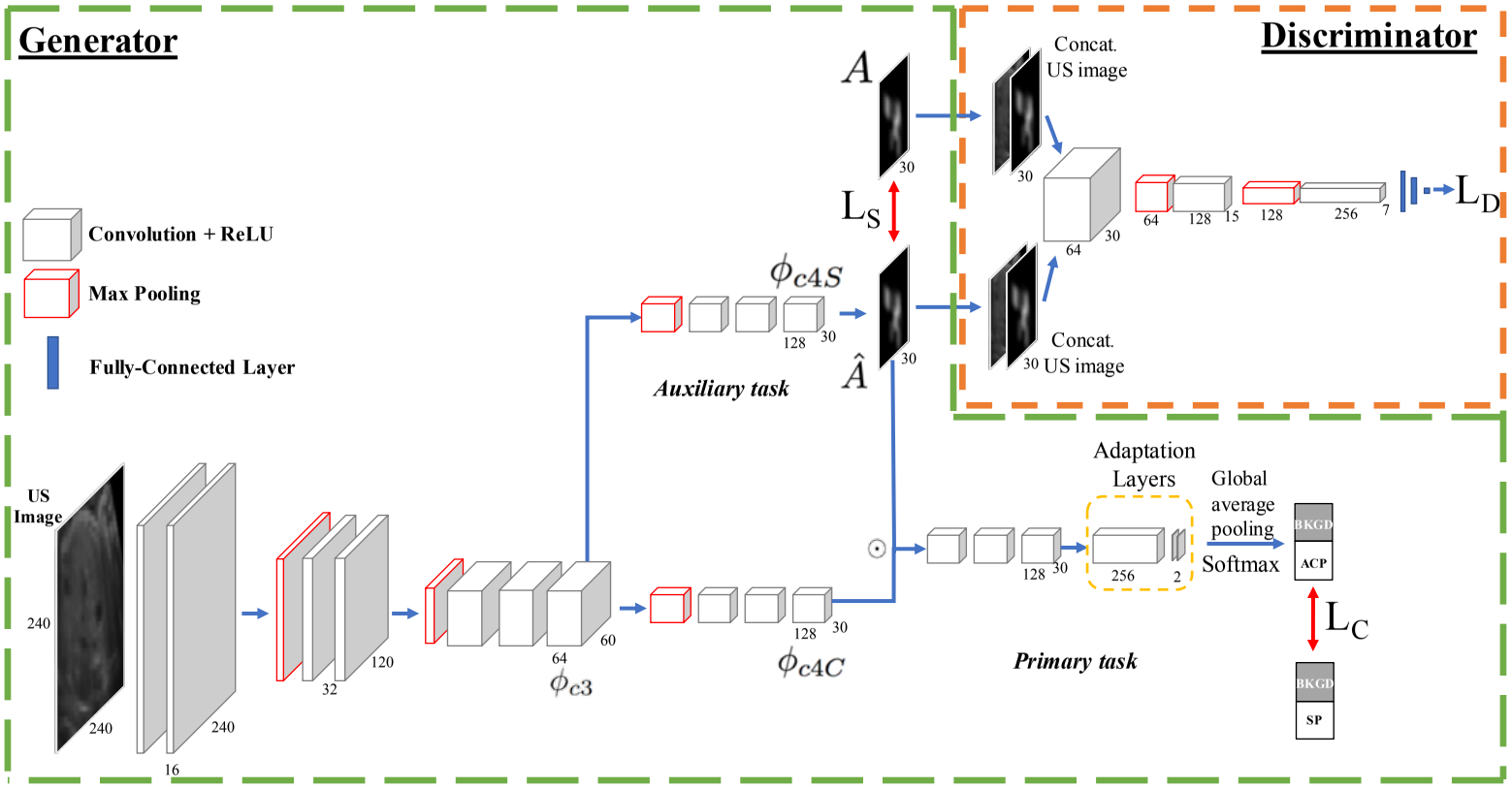


Figure 5.7: Architecture of the multi-task SonoEyeNet (*M-SEN*). It has two modules: the generator (in Green-dashed polygon) and the discriminator (Orange-dashed box). The generator has two tasks: a primary task to classify frames (bottom) and an auxiliary task to predict visual attention map ($\hat{A}$). The discriminator differentiates between real (A) and predicted ($\hat{A}$) attention maps. The dotted circle $\odot$ indicates element-wise multiplication. L_S, L_C and L_D represent the losses of saliency prediction, frame classification, and the discriminator, respectively.

Discriminator architecture

The discriminator module D is a CNN with three convolutional layers with 64, 128 and 256 3×3 kernels respectively, each with max pooling and leaky rectified linear unit (leaky ReLU) activation, followed by three fully-connected (FC) layers. Hyperbolic tangent (*tanh*) activation was used for the first two FC layers and sigmoid activation for the last FC layer, following the choice in literature [**radford2015unsupervised**]. As sonographer attention is conditional on the US video frame, I is concatenated to A or $\hat{A}$ as inputs into D.

The discriminator loss L_D and the generator loss L_G for each mini-batch with m samples are defined [**goodfellow2014generative**] as:

$$L_D = -\frac{1}{m}\sum_{i=1}^{m} \log(D(I_i, A_i; \theta_D)) + \log(1 - D(I_i, \hat{A}_i; \theta_D)) \tag{5.19}$$

$$L_G = \lambda_1 L_S + \lambda_2 L_C - \frac{\lambda_3}{m}\sum_{i=1}^{m} \log(D(I_i, \hat{A}_i; \theta_D)) \tag{5.20}$$

where $D(I, A; \theta_D)$ is the probability that the discriminator successfully recognizes the real attention map, while $D(I, \hat{A}; \theta_D)$ is the probability that the discriminator is fooled. θ_D represents the weights of D. The generator loss is designed to include both classification and saliency losses L_C and L_S as well as an adversarial regulariser by using the discriminator loss on $\hat{A}$; this regulariser was not used during generator pre-training. Hyper-parameters λ_1, λ_2, and λ_3 determine the relative contributions of the three losses. The saliency loss L_S is defined as the pixel-level content loss between $\hat{A}$ and A, which is used to train the generator of the attention maps. Two loss functions were experimented with in the models shown in Table. 5.4: *M-SEN MSE* uses the mean squared error (MSE) loss, a baseline loss as it has been used in many visual saliency prediction works [**kruthiventi2017deepfix**]; *M-SEN BCE* uses binary cross-entropy (BCE) loss, which is mathematically equivalent to Kullback-Leibler divergence, arguably the best metric to measure saliency prediction performance [**huang2015salicon**]. For the classification task, cross-entropy loss was used as L_C, the same as in [**yifan2018eye1**].

Training details

The generator was independently pre-trained for 30 epochs before adding the discriminator as the adversarial regulariser. The network was initialized using the first three convolutional blocks of SonoNet [**Baumgartner2017**]; all other layers were initialized using a zero-mean Gaussian distribution with standard deviation 0.01. Batch normalization and dropout (rate = 0.2) were used for each convolutional layer before the adaptation layers. The weight λ_1 was dynamically changed from 2 to 1 over epochs so as to allow the generator to focus on learning attention maps first and then frame classification. The weight λ_2 was set to 1, as classification was the primary task of the network. After 30 epochs, the network was further fine-tuned for 2000 steps using an adversarial training scheme by training the discriminator and the generator once per step in an alternating manner. When training the discriminator, one-sided label smoothing [**radford2015unsupervised**] was used; when training the generator, the weights of discriminator were not updated. The network was trained using adaptive moment estimation (Adam) with an initial learning rate 2×10^{-4}. Batch size was set to 64. Five-fold cross-validation was used.

5.5.3 Results

Frame classification performance

Classification results of all models are presented in Table. 5.4. Two observations can be made. First, all *M-SEN* models that use learned saliency maps to assist frame detection outperform the *SonoNet* models [**Baumgartner2017**], which are supervised only by image-level labels. The classification precision of 79.3% for the *SonoNet-32* model was increased to 96.8% ($p = 0.008$) in *M-SEN BCE + GAN*, which uses BCE as saliency loss and is further fine-tuned using adversarial regulariser. Recall increased from 82.1% to 96.2% ($p = 0.014$). Second, models that adopted adversarial regularisers achieve better results: performances of both BCE and MSE models are improved by training with an adversarial discriminator. For example, introducing adversarial regulariser to *M-SEN MSE* increased its precision from 92.4% to 94.8% ($p = 0.037$), recall from 75.6% to 91.9% ($p = 0.021$), and F-1 score

Table 5.4: *Comparative evaluation of classification performance. In column "Inputs", "I"and "A" refer to US images and attention maps, respectively. "SS-cls Net" refers to single-stream network trained only on attention maps to classify US video frames.*

Models	Inputs	Precision	Recall	F1-score
M-SEN BCE + GAN	I	96.8	96.2	96.5
M-SEN BCE	I	96.7	90.5	93.5
M-SEN MSE + GAN	I	94.8	91.9	93.3
M-SEN MSE	I	92.4	75.6	83.2
SonoNet-32 [**baumgartner2017sononet**]	I	79.3	82.1	80.7
SonoNet-16 [**baumgartner2017sononet**]	I	73.6	74.1	73.8
SS-cls Net	A	71.5	76.4	73.9
SonoEyeNet-Late FT [**yifan2018eye1**]	I and A	96.5	99.0	97.8

Table 5.5: *Quantitative metrics of saliency prediction on the test set. "SS-att" indicates those single-stream models for saliency prediction without a classification branch. Saliency metrics include information gain (IG), Pearson's Cross-Correlation (CC), normalized saliency scan path (NSS), similarity (SIM), and area under curve (AUC)* [**bylinskii2016different**].

Models	IG	CC	NSS	SIM	AUC
M-SEN BCE + GAN	**0.543**	0.693	**2.525**	0.512	0.775
M-SEN BCE	0.429	0.615	2.144	0.469	0.726
M-SEN MSE + GAN	0.307	0.634	2.327	0.309	0.616
M-SEN MSE	0.288	0.556	2.253	0.310	0.603
SS-att BCE	0.192	**0.708**	1.480	**0.570**	**0.801**
SS-att MSE	0.152	0.546	1.329	0.532	0.788

from 83.2% to 93.3% ($p = 0.026$). The best performing *M-SEN BCE + GAN* model achieves performance competitive to that of *SonoEyeNet-Late FT* [**yifan2018eye1**] ($p = 0.692$), which uses both the US frame and sonographer visual attention map for inference. *SS-cls Net* that attempts to classify US video frames solely on sonographer visual attention map achieved a performance similar to that of *SonoEyeNet-Late FT*.

Saliency prediction performance

Examples of predicted visual attention maps generated by variations of *M-SEN* models on the test set US video frames can be seen in Fig. 5.8. All independently

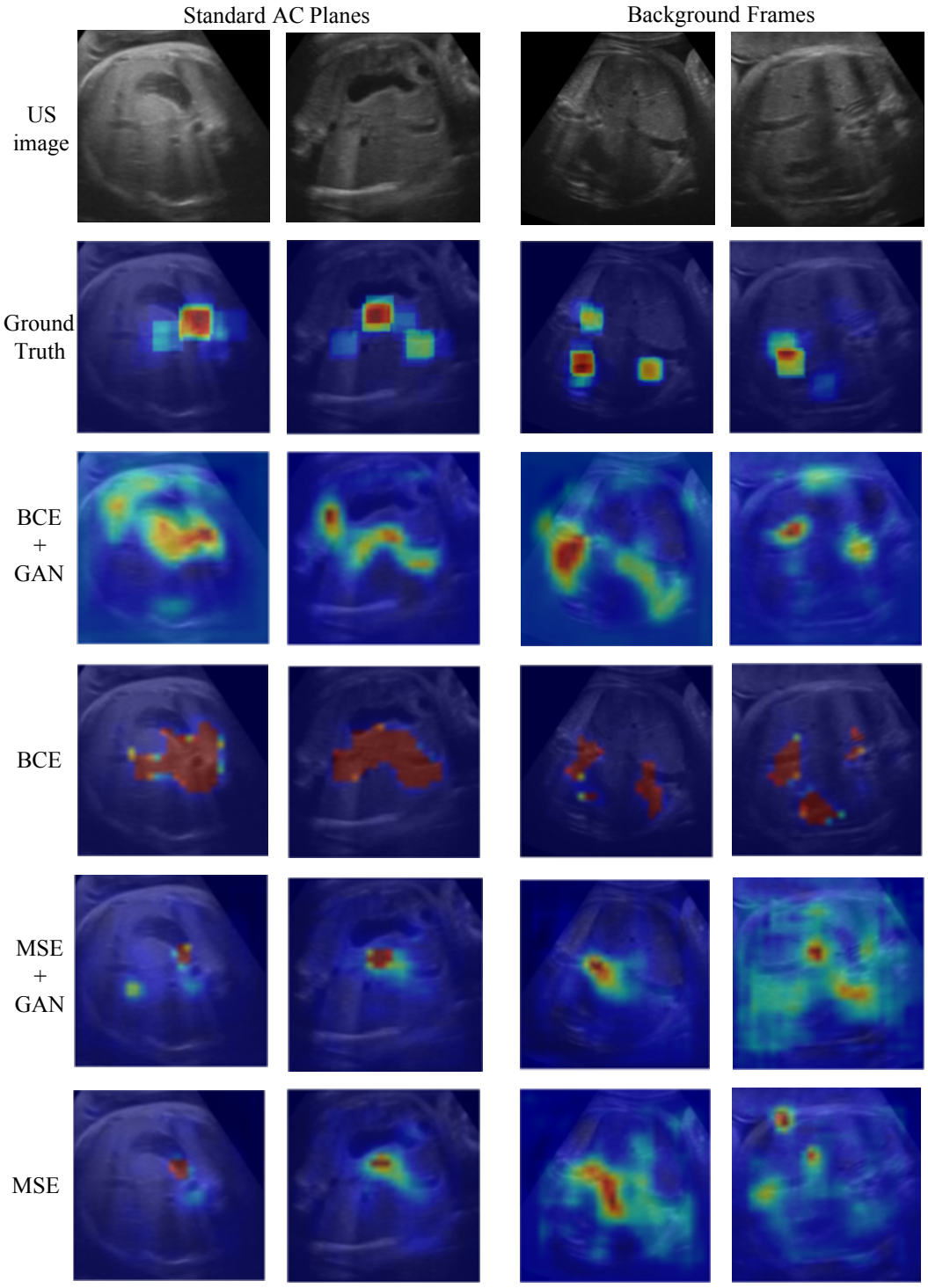


Figure 5.8: Attention maps generated by different variations of the Generator. From top to bottom: US image, Ground-truth (Sonographer's actual attention map), *M-SEN BCE + GAN*, *M-SEN BCE*, *M-SEN MSE + GAN*, *M-SEN MSE*.

trained *M-SEN* models generate visually good quality attention maps that capture the salient regions fixated by sonographers, e.g. edges of the stomach bubble. Interestingly, *M-SEN BCE* extends beyond the constraint of the ground-truth salient regions and explores other key anatomical structures, e.g. the umbilical veins, that the sonographers had not necessarily looked at during examination. In addition, we observe a similar trend previously observed in Table.5.4 that adding an adversarial regulariser improves model performance. *M-SEN BCE + GAN* is able to learn a more realistic visual attention map while retaining the ability to assign confidence to other key anatomical structures. Adding an adversarial discriminator regularises the predicted saliency map by reducing confidence in its false-positive points, as can be observed in Fig. 5.8; this can also be observed in Table. 5.5, where saliency scores measured by all five metrics increase.

5.5.4 Discussions

The best performing *M-SEN BCE + GAN* model outperforms baseline models (*SS-att BCE* and *SS-att MSE*) in IG and NSS, but not in CC, SIM and AUC. It shows that tasks in the multi-task network influence each other, which strikes a balance between mimicking a ground truth attention map as close as possible and generating an attention map that includes clinically useful information on the US video frame.

In general, MSE models generate attention maps with more false-negatives (predicted as non-salient while fixated by sonographer), while BCE models generate more false-positives (predicted as salient but not fixated by sonographer). A change of loss function from *M-SEN MSE* to *M-SEN BCE* improves saliency performance measured by four out of five metrics. The performance improvement is consistent with the way gaze-tracking data was collected.

MSE corresponds to computing the Euclidean distance between the predicted saliency and the ground truth and it aims at maximizing the peak signalto-noise ratio (PSNR) [**Pan_2017_SalGAN**]. Empirically, MSE tends to filter high spatial frequencies in the output, favoring blurred contours. In our experiment, sonographers were able to view the images for as long as they wanted, there could be multiple

points where sonographers attended to. Thus, each pixel in the ground-truth attention map can be interpreted as estimates of the probability that a particular pixel is attended by an observer. It is thus appropriate to apply an element-wise sigmoid to each output in the final layer so that the pixel-wise predictions can be thought of as probabilities for independent binary random variables. BCE is natually the more appropriate loss function for this task compared to MSE. As confirmed by other literature [**Pan_2017_SalGAN**], BCE loss performs better than MSE loss. The only exception is in NSS, where *M-SEN MSE* outperforms *M-SEN BCE*. This can be attributed to the fact that NSS is extremely sensitive to false positives, which *M-SEN BCE* exploits to cover non-fixated anatomical structures to benefit classification task.

Since sonographers view each frame for as long as they want, fixations on background frames, where there's no relevant anatomical structure, are non-specific; frames closer to the standardized planes exhibit a more consistent gaze pattern. Thus, attention maps provide a coarse distinction between backgrounds and frames that contain relevant anatomical structures. On the other hand, for frames close to the standardized plane, attention maps can look similar, and this is where image features (i.e. intensity values) become more important for this task.

5.6 Conclusions

This chapter discussed several architectures to model and fine-tune sonographer visual attentions. A novel multi-task convolutional neural network called *Multi-task SonoEyeNet* (MSEN) that learns to generate clinically relevant visual attention maps in order to assist standard AC plane detection achieved the best performance, both in terms of saliency prediction and frame classification. It expands the potential clinical usefulness of SonoEyeNet, the classification model discussed in Chapter 4, by eliminating the requirement of input gaze tracking data during inference without compromising plane detection performance.

Two possible extensions to the current model should be explored. First is to model spatio-temporal visual attention. It has been shown that spatio-temporal

visual attention can improve frame classification performance using bag-of-visual-words (BoVW) framework [**AhmedThesis**]. It is thus worth exploring methods to model spatio-temporal attention under a deep-learning framework. Second is to extend to the detection of all standard biometry planes: standard AC planes, standard Head Circumference (HC) planes, and standard Femur Length (FL) planes. Recently a large scale US fetal anomaly full scan video dataset with simultaneous gaze tracking collected under the project PULSE became available. These two extensions are studied using PULSE dataset, and presented in the next chapter.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe

6

Spatio-temporal visual attention modelling for standard biometry planes detection

Contents

6.1 Introduction

In clinical practice, a sonographer performs routine anomaly scans by moving and adjusting the US probe on a pregnant woman's abdomen while viewing a US video stream displayed on a monitor, navigating through different fetal anatomies and

structures, to identify the views upon which to make biometric measurements. This is very different from the experimental setting used in the previous two chapters, where sonographers (1) were not manipulating the US probe, but rather using the key-board to navigate through previously recorded US video frames; and (2) were free to view every single frame for as long as they wanted. A crucial step towards automating the acquisition of standardised views of the fetus in a real clinical setting is to learn from real clinical data.

A novel dataset under the project *Perception Ultrasound by Learning Sonographic Experience* (PULSE) is being collected in the Oxford University NHS Trust Department of Obstetrics. This dataset contains real-time screen recording of US scans, coupled with simultaneous recording of gaze-tracking data as well as probe movement data. This chapter's contents is based on a subset of the PULSE dataset.

In the previous chapter, **Multi-task SonoEyeNet** (MSEN) was trained on both an input fetal ultrasound image's class label (standardized abdominal circumference plane or background) and its corresponding sonographer visual attention map, to predict sonographer visual attention on unseen test images. Building on this work, in this chapter we propose Temporal SonoEyeNet (T-SEN) to (1) learn spatio-temporal information in short video clips to model sonographer visual attention; (2) using the predicted sequence of visual attention maps to assist standard frame detection task; and (3) to expand from Abdominal Circumference Plane (ACP) detection to the detection of all three standard biometry planes: ACP, HCP (Head Circumference Plane), and FLP (Femur Length Plane). A key component of this work is to study the most suitable network architectures to model spatio-temporal information in US sequences, and explore loss functions that provide temporal regularisation for visual attention modelling.

6.2 Originality and Individual Role

The PULSE dataset, including video frames as well as gaze-tracking data, was collected as a part of the PULSE project. An image-based text-recognition algorithm

for the recognition of standard biometry planes was written by Mr. Richard Droste, a D.Phil student in the Institute of Biomedical Engineering at University of Oxford.

Independently, I selected a subset of the PULSE data, including video clips of fetal abdomen, brain, and femur from 89 pregnant women, for processing. I manually annotated the anatomy of these videos at frame-level; I also used aforementioned gaze-tracking data to generate ground-truth visual attention heatmaps for each video frame. Using Pytorch [**paszke2017automatic**], an open source Python library, I wrote a video frame sampler for this dataset; I defined, trained, and tested the Temporal SonoEyeNet (T-SEN) to model sonographer temporal visual attention as well as to detect standard biometry planes from US video clips.

6.3 PULSE Data Collection and Processing

A noval dataset of clinical ultrasound exams with real-time gaze-tracking data were collected on 10 experts (clinicians and sonographers). By the time this thesis was written, more than 700 pregnant women have participated in the study. Ethics approval was obtained for data recording and data were stored according to local data governance rules. Even though data collection and pre-processing were designed and performed by Dr. Pierre Chatelain and Dr. Harshita Sharma, this process is written in detail here so that it can be compared with the dataset used in the previous chapters.

Experimental Setup

All free-hand ultrasound exams were performed on a GE Voluson E8 scanner (General Electric, USA); its LCD monitor has a resolution of 1920×1080 pixels and refreshes at a frequency of 60 Hz, while the video signal is recorded lossless at 30 Hz. The gaze-tracking data were recorded using a Tobii Eye Tracker 4C (Tobii, Sweden) that records the point-of-gaze (relative x and y coordinates with corresponding timestamps) and 3-D eye position of each eye at a rate of 90 Hz, effectively recording 3 gaze points per frame. The eye tracker was rigidly attached under the display area with a magnetic mounting bracket as per the instruction

Figure 6.1: The experimental setup of gaze tracking experiment of PULSE project, consisting of a screen that displays stimuli, an Tobii eye tracker attached to the bottom of a monitor, and a GE Voluson E8 scanner. The figure was reproduced from [**chatelain2018evaluation**] ©IEEE.

of the product. The eye tracker was calibrated for each sonographer following a 9-point calibration protocol as described in [**chatelain2018evaluation**]. The experiment setup is presented in Fig. 6.1.

Data Collection Protocol

Calibration of the eye tracker was conducted for each sonographer. During calibration, 9 circular targets appeared randomly on the 4 corners, 4 edge mid-points as well as the center of the screen. The calibration procedure was repeated until the error between all recorded gaze positions and their corresponding target positions fell below the threshold of 1.5° visual angle. Before each ultrasound exam the sonographer-specific configuration was selected.

Sonographers were free to adjust the height of the chair and the inclination of the monitor, and operate the ultrasound probe in order to perform ultrasound examinations without being affected by the existence of an eye tracker so that real and clinically-relevent gaze data were recorded. During ultrasound examination, sonographers identify key fetal anatomies and make corresponding measurements according to the UK FASP guidelines for mid-pregnancy ultrasound examinations

[**kirwan2010nhs**]. Video signals of the scanner and sonographers' gaze tracking data are stored as raw data for later processing. Video signals were stored as .mp4 video files, and each video file was processed and broken down to individual video frames stored as .png image files. A normal examination takes 34 ± 14 minutes.

6.4 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) [**pearlmutter1989learning**, **giles1994dynamic**] are commonly used to encode temporal or spatio-temporal information. In order to introduce several important components used in the next section, foundations for several key concepts are introduced here.

RNNs are a family of neural networks that process sequential data: a sequence of words that form a sentence, or a sequence of video frames that forms a video clip. The sequential data can be denoted as $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(\tau)}$, where τ denotes the length of the sequence. An internal state of a system $\mathbf{h}^{(t)}$ at time step t is a representation of all data in the past time steps:

$$\mathbf{h}^{(t)} = g^{(t)}(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t-2)}, ..., \mathbf{x}^{(2)}, \mathbf{x}^{(1)}) \tag{6.1}$$

where $g^{(t)}$ represents a function that calculates the current state of the system taking into consideration all past sequences. RNN performs recurrent computations that involves the internal state of the system $\mathbf{h}^{(t)}$ and external input of the previous time step $\mathbf{x}^{(t-1)}$, at time t, by using the same transition function f at every time step:

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta) \tag{6.2}$$

where θ represents the parameters of the transition function. Learning a single shared transition function allows generalization to sequence lengths and avoids the need to learn a separate model $g^{(t)}$ for all possible time steps.

Recurrent Neural Networks architectures can generate output on every time step or generate one output at the final time step, depending on the purpose of the network. For the purpose of this study where it is desired to classify every frame from a video sequence of certain length, a RNN with recurrent connections between

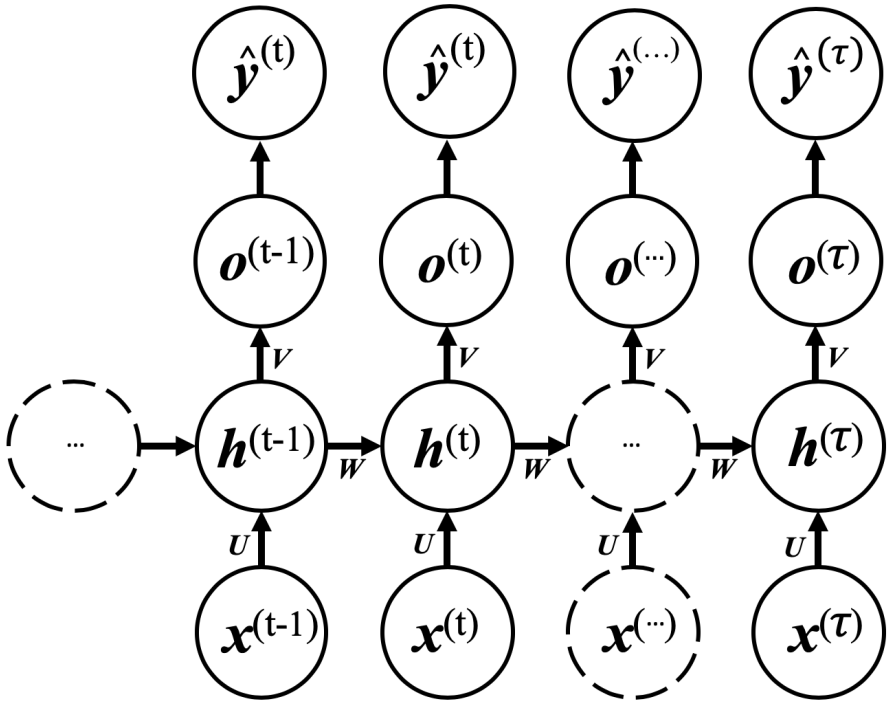


Figure 6.2: General architecture of a unrolled RNN that produces a single output.

hidden units that reads an entire sequence and then produces an output for each time step is built and analysed, as can be seen in Fig. 6.2.

6.4.1 Classic RNN

The forward propagation of a **vanilla RNN** begins with a random initialisation of initial state $\mathbf{h}^{(0)}$. Then for each time step from $t = 1$ to $t = \tau$, $\mathbf{h}^{(t)}$ is computed by linear combination of $\mathbf{h}^{(t-1)}$ and $\mathbf{x}^{(t)}$ using input-to-hidden weight matrices $\boldsymbol{U}$, hidden-to-output weight matrices $\boldsymbol{V}$, and hidden-to-hidden weight matrices $\boldsymbol{W}$, together with bias vectors $\boldsymbol{b}$ and $\boldsymbol{c}$, followed by a non-linear operation, which is commonly a hyperbolic tangent.

$$\begin{aligned} \boldsymbol{a}^{(t)} &= \boldsymbol{b} + \boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{U}\boldsymbol{x}^{(t)}, \\ \boldsymbol{h}^{(t)} &= \tanh(\boldsymbol{a}^{(t)}). \end{aligned} \tag{6.3}$$

At each time step, the predicted class probability $\hat{\boldsymbol{y}}^{(t)}$ is calculated from the output $\boldsymbol{o}^{(t)}$ using the *softmax* function:

$$\begin{aligned} \boldsymbol{o}^{(t)} &= \boldsymbol{c} + \boldsymbol{V}\boldsymbol{h}^{(t)}, \\ \hat{\boldsymbol{y}}^{(t)} &= \mathrm{softmax}(\boldsymbol{o}^{(t)}). \end{aligned} \tag{6.4}$$

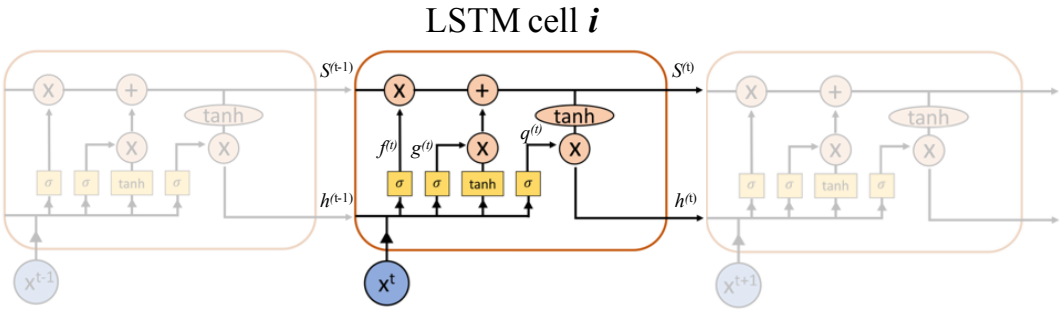


Figure 6.3: Schematic of a Long Short-term Memory (LSTM) Cell.

In the case of using a vanilla RNN to classify sequences of video frames, the input-to-hidden, hidden-to-output, and hidden-to-hidden weight matrices are modeled using fully connected layers. The loss function used here is the same as the normal classification loss, *i.e.* cross-entropy loss.

The greatest challenges for a vanila RNN is the problem of long-term dependencies: if the input sequence is relatively long, gradients propagated over many steps tend to either vanish or explode. The most effective way of dealing with the long-term dependency problem is to use gating mechanisms that control how much information from a previous hidden state and input should be used. Most successful application of a gated RNNs are **long short-term memory (LSTM)** and networks based on a **gated recurrent unit (GRU)**.

Long-Short Term Memory

LSTM recurrent networks have "LSTM cells" (Fig. 6.3) that have an internal recurrence (inner loop) in addition to the outer recurrence of the vanilla RNN as mentioned before. This inner loop controls the flow of information using several gates: **forget gate (f), input gate (g),** and the **output gate (q)**. Each gate outputs a value between 0 and 1 via a sigmoid unit to control information flow. For a LSTM cell i:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right) \tag{6.5}$$

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right) \tag{6.6}$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right) \tag{6.7}$$

where $\boldsymbol{b}$, $\boldsymbol{U}$ and $\boldsymbol{W}$ respectively denote the biases, input weights, and recurrent weights of the forget gates; $\boldsymbol{x}^{(t)}$ and $\boldsymbol{h}^{(t)}$ respectively denote the current input vector and the current hidden layer vector.

The LSTM cell internal state, $s_i^{(t)}$, is updated using the forget gate and input gate:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \tanh\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right) \tag{6.8}$$

where $\boldsymbol{b}$, $\boldsymbol{U}$ and $\boldsymbol{W}$ respectively denote the biases, input weights, and recurrent weights into the LSTM cell. The hidden layer vector is updated:

$$h_i^{(t)} = \tanh\left(s_i^{(t)}\right) q_i^{(t)} \tag{6.9}$$

Gated Recurrent Units

A **GRU** is a simpler type of gated RNN, where instead of using 3 gates, it only uses an "update" gate ($\boldsymbol{u}$) and a "reset" gate ($\boldsymbol{r}$) as can be seen in Fig. 6.4. The gate values are defined similar to the gates mentioned for a LSTM:

$$u_i^{(t)} = \sigma\left(b_i^u + \sum_j U_{i,j}^u x_j^{(t)} + \sum_j W_{i,j}^u h_j^{(t-1)}\right) \tag{6.10}$$

and

$$r_i^{(t)} = \sigma\left(b_i^r + \sum_j U_{i,j}^r x_j^{(t)} + \sum_j W_{i,j}^r h_j^{(t-1)}\right) \tag{6.11}$$

Using these two gates, the internal state $h_i^{(t)}$ is update:

$$h_i^{(t)} = u_i^{(t)} h_i^{(t-1)} + (1 - u_i^{(t)}) \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} r_j^{(t-1)} h_j^{(t-1)}\right) \tag{6.12}$$

6.4.2 Convolutional RNNs

Convolutional LSTM

Although a LSTM layer was proven powerful for modelling temporal information in a sequence [**sutskever2014sequence**, **graves2013generating**, **gregor2015draw**], its dot product operation is especially redundant for spatial information such as

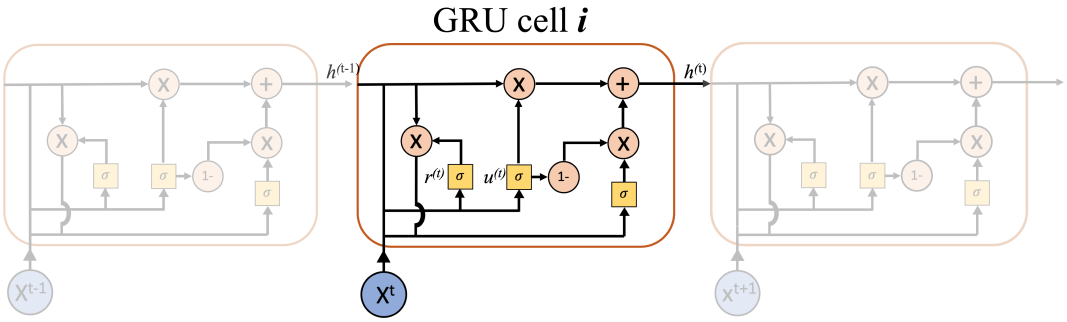


Figure 6.4: Schematic of Gated Recurrent Unit (GRU).

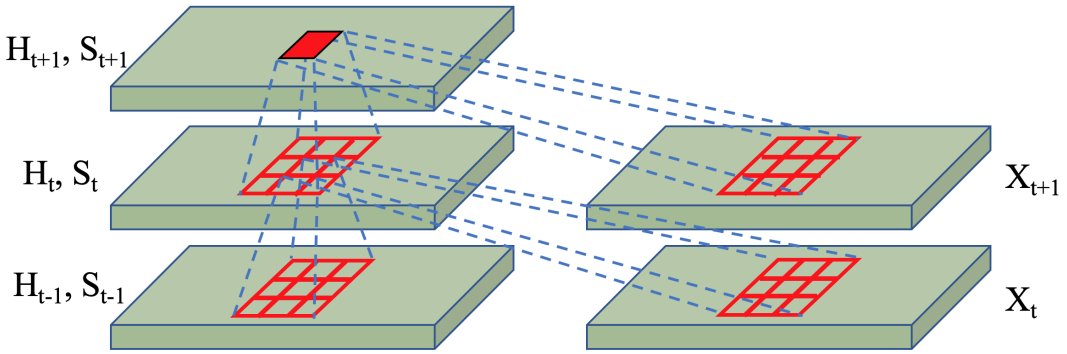


Figure 6.5: Schematic of convolution operation for input-to-state and state-to-state transitions. This figure was reproduced from [**xingjian2015convolutional**].

images with high dimensional feature representations. In input-to-state and state-to-state transitions, no spatial information is encoded. To address this problem, Shi *et al.* [**xingjian2015convolutional**] extended the LSTM concept by replacing the dot product operation by a convolution operation in input-to-state and state-to-state transitions, as can be seen in Fig. 6.5. For a convolutional LSTM (**ConvLSTM**) cell i, the forget gate, input gate and output gate are:

$$\mathbf{f}^{(t)} = \sigma\left(\mathbf{b}^{f} + \mathbf{U}^{f} * \mathbf{x}^{(t)} + \mathbf{W}^{f} * \mathbf{h}^{(t-1)}\right) \tag{6.13}$$

$$\mathbf{g}^{(t)} = \sigma\left(\mathbf{b}^{g} + \mathbf{U}^{g} * \mathbf{x}^{(t)} + \mathbf{W}^{g} * \mathbf{h}^{(t-1)}\right) \tag{6.14}$$

$$\mathbf{q}^{(t)} = \sigma\left(\mathbf{b}^{o} + \mathbf{U}^{o} * \mathbf{x}^{(t)} + \mathbf{W}^{o} * \mathbf{h}^{(t-1)}\right) \tag{6.15}$$

where $\mathbf{b}$, $\mathbf{U}$ and $\mathbf{W}$ respectively denote the bias, convolutional kernel for the input gate and convolutional kernel for the forget gate. $\boldsymbol{x}^{(t)}$ and $\boldsymbol{h}^{(t)}$ respectively

denote the current input tensor and the current hidden layer tensor. The symbol $*$ represents a convolution operator.

The convolutional LSTM cell internal state, $\boldsymbol{s}^{(t)}$, is updated using the forget gate and input gate:

$$\mathbf{s}^{(t)} = \mathbf{f}^{(t)} \odot \mathbf{s}^{(t-1)} + \mathbf{g}^{(t)} \odot \tanh\left(\mathbf{b} + \mathbf{U} * \mathbf{x}^{(t)} + \mathbf{W} * \mathbf{h}^{(t-1)}\right) \tag{6.16}$$

where $\mathbf{b}$, $\mathbf{U}$ and $\mathbf{W}$ respectively denote the biases, input convolutional kernel, and recurrent convolutional kernel into the LSTM cell. The symbol $*$ represents a convolution operator and $\odot$ represents a Hadamard product. The hidden layer vector is updated:

$$\mathbf{h}^{(t)} = \mathbf{q}^{(t)} \odot \tanh\left(\mathbf{s}^{(t)}\right) \tag{6.17}$$

Convolutional GRU

Similarly, an extension to GRUs was proposed by Siam *et al.* [**siam2017convolutional**] to incorporate convolutional operations. For a convolutional GRU (**ConvGRU**), the modified update gate ($\mathbf{u}$) and the reset gate ($\mathbf{r}$) are:

$$\mathbf{u}^{(t)} = \sigma\left(\mathbf{b}^{u} + \mathbf{U}^{u} * \mathbf{x}^{(t)} + \mathbf{W}^{u} * \mathbf{h}^{(t-1)}\right) \tag{6.18}$$

and

$$\mathbf{r}^{(t)} = \sigma\left(\mathbf{b}^{r} + \mathbf{U}^{r} * \mathbf{x}^{(t)} + \mathbf{W}^{r} * \mathbf{h}^{(t-1)}\right) \tag{6.19}$$

The internal state $h^{(t)}$ is updated:

$$\hat{\mathbf{h}}^{(t)} = \sigma\left(\mathbf{b} + \mathbf{U} * \mathbf{x}^{(t)} + \mathbf{W} * (\mathbf{r}^{(t)} \odot \mathbf{h}^{(t-1)})\right) \tag{6.20}$$

$$\mathbf{h}^{(t)} = (\mathbf{1} - \mathbf{u}^{(t)}) \odot \mathbf{h}^{(t-1)} + \mathbf{u}^{(t)} \odot \hat{\mathbf{h}}^{(t)} \tag{6.21}$$

6.5 Temporal SonoEyeNet

Multi-task SonoEyeNet, as discussed in the previous chapter, was able to capture spatial information in 2-D US images by predicting sonographer visual attention

maps to assist a binary classification task of detecting standard Abdominal Circumference Planes (ACP) from background abdominal images, but it fails to utilize temporal information inherent in 2-D US videos, which is hypothesized to help standard biometry plane detection. This following section describes Temporal SonoEyeNet (T-SEN), which expands on the idea of Multi-task SonoEyeNet in two ways. First, it models temporal visual attention variations of a sonographer on consecutive frames of 2-D US videos using a Temporal Attention Module (TAM); second, it expands detection targets from ACP to all three standard biometry planes: ACP, HCP (Head Circumference Planes) and FLP (Femur Length Planes).

6.5.1 Data Processing

Video Frame Processing

A subset of the PULSE dataset containing anomaly scans performed by a single sonographer was selected. An image-based text recognition algorithm was used to capture the standard biometry planes on which sonographers made measurements, including standard ACP, standard FLP and HCP. Since it is common for sonographers to hold the probe still or make very small movements when a standard plane is found, multiple consecutive frames could contain the same standard plane. It is also common for a sonographer to move the probe away from the already found standard plane to confirm that it is the best available plane before entering freeze frame, a static video frame on which the sonographer makes biometry measurement. Thus, in each anomaly scan, an experienced biomedical engineer manually annotated all such standard biometry planes. Each frame in this dataset was assigned one of the following *frame-level* labels: standard ACP, background abdomen (bg Ab), standard HCP, background Head (bg Head), standard FLP, background Femur (bg Femur), and others. A video clip C starting from a video frame containing discernable non-standard anatomy up until the frame before freeze frame capturing the standard biometry plane of that anatomy was sampled from a scan to reflect sonographer navigation and plane-finding decision-making for a particular standard biometry plane. For each anomaly scan, C_A, C_H and C_F are sampled to reflect

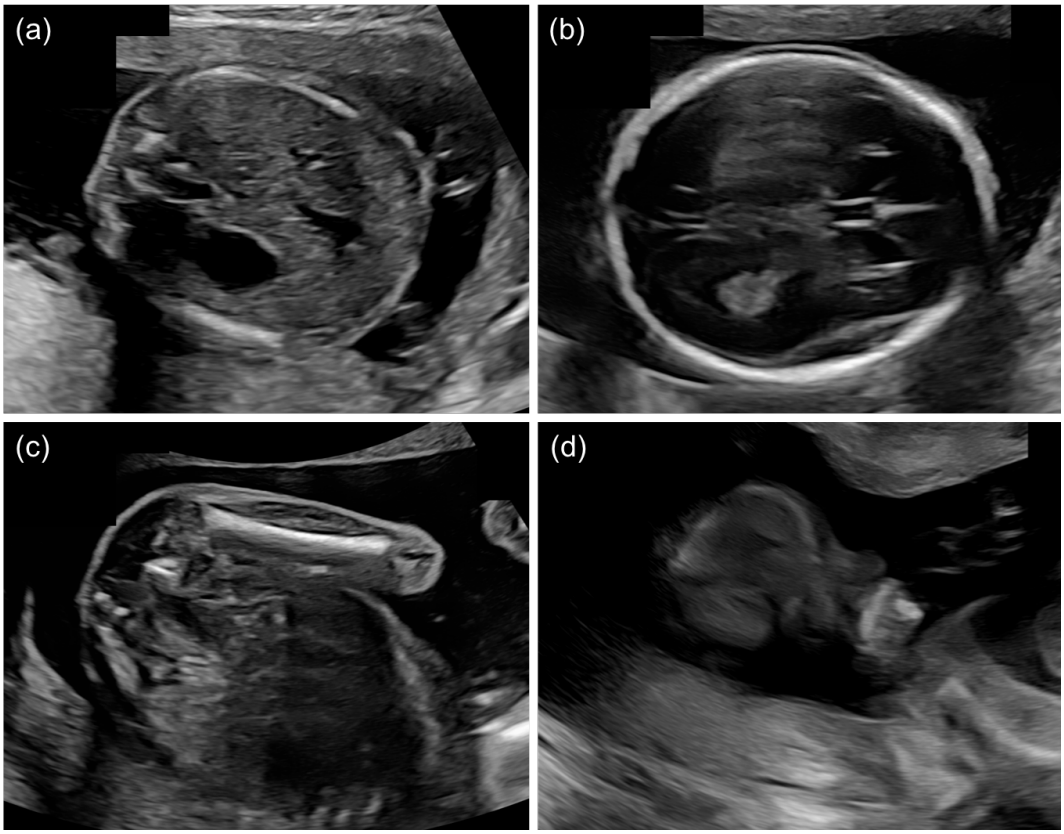


Figure 6.6: Standard Biometry Planes and an example of background frame. (a) Abdominal Circumference Plane; (b) Brain Transventricular (tv.) Plane; (c) Femur Plane; (d) others

the navigation process of finding ACP, FLP and HCP. An "other" class video clip C_O that doesn't contain any clear anatomical structures was also sampled from the scans. A total of 89 C_A, 71 C_H, 76 C_F, and 44 C_O were sampled. Each of these clips contains between 200 and 500 frames, of which between 10 and 40 are standard biometry planes. In total, this dataset contains 1910 ACPs, 2359 HCPs and 2151 FLPs from 22927 Abdominal frames, 24437 Head frames, 12762 Femur frames, and 8982 other frames. All frames containing Doppler overlay, 3-D/4-D, split-screens, or freeze frames containing bounding boxes/circles were excluded from the dataset. Text information and the Graphical User Interface on each frame was cropped out. Examples of the standard planes are shown in Fig. 6.6. All frames were resized to 240×240 and randomly cropped into size 224×224 on-the-fly for data augmentation during training. In addition, all frames were normalised to zero-mean and unit-variance.

Gaze Data Processing

Using the gaze data, binary maps B of the same dimensions as the corresponding frames are generated, with pixels corresponding to fixation points labeled as 1 and others labeled as 0. A sonographer visual attention map A is generated for each binary map by convolving it with a truncated Gaussian Kernel $G(\sigma_{x,y})$: $A = B * G(\sigma_{x,y})$, where $G(\sigma_{x,y})$ has 30 *pixels* along x, y-dimemsions corresponding to visual angle of 1.5° with an observer-to-screen distance of 0.5 m. A is further normalized such that all pixel values add up to 1. Examples of songrapher visual attention maps overlaid on their corresponding examples of 6 consecutive frames of C_A, C_F, C_H and C_O can be seen in Fig. 6.7.

Training Sample Generation

Short sequences of video frames with defined time depth and skip size are sampled from the aforementioned video clips, as can be seen in Fig. 6.8. Time depth is defined as the number of frames in a video sequence that forms an input to the network, and skip size is the number of frames skipped in the original video clip C between consecutive sampled frames. Arrows of different colors and line styles indicate 4 possible ways of sampling from a clip C with the same skip size (*left*), while the stacked frames with different colors and line styles shows the 4 resultant training sample with the same time depth (*right*). In order to model temporal attention of different time scales, time depth of 5, 10, 15, and 20 frames were tested. Sampling with skips of every 5th, 10th, 15th, or 20th frames was also experimented with. This study uses time depth of 10 for most efficient use of GPU memory, and a skip size of 10, as this skip size strikes a good balance in terms of reducing temporal redundancy and not losing temporal information. Each video sequence was coupled with frame-wise anatomy labels as well as corresponding sonographer visual attention maps. In addition, a *sequence-level* label (one label per sequence) is assigned for each input sequence: if the sequence contains one or more instances of a standard biometry plane, it is labeled as a standard sequence; otherwise a background sequence of that anatomy.

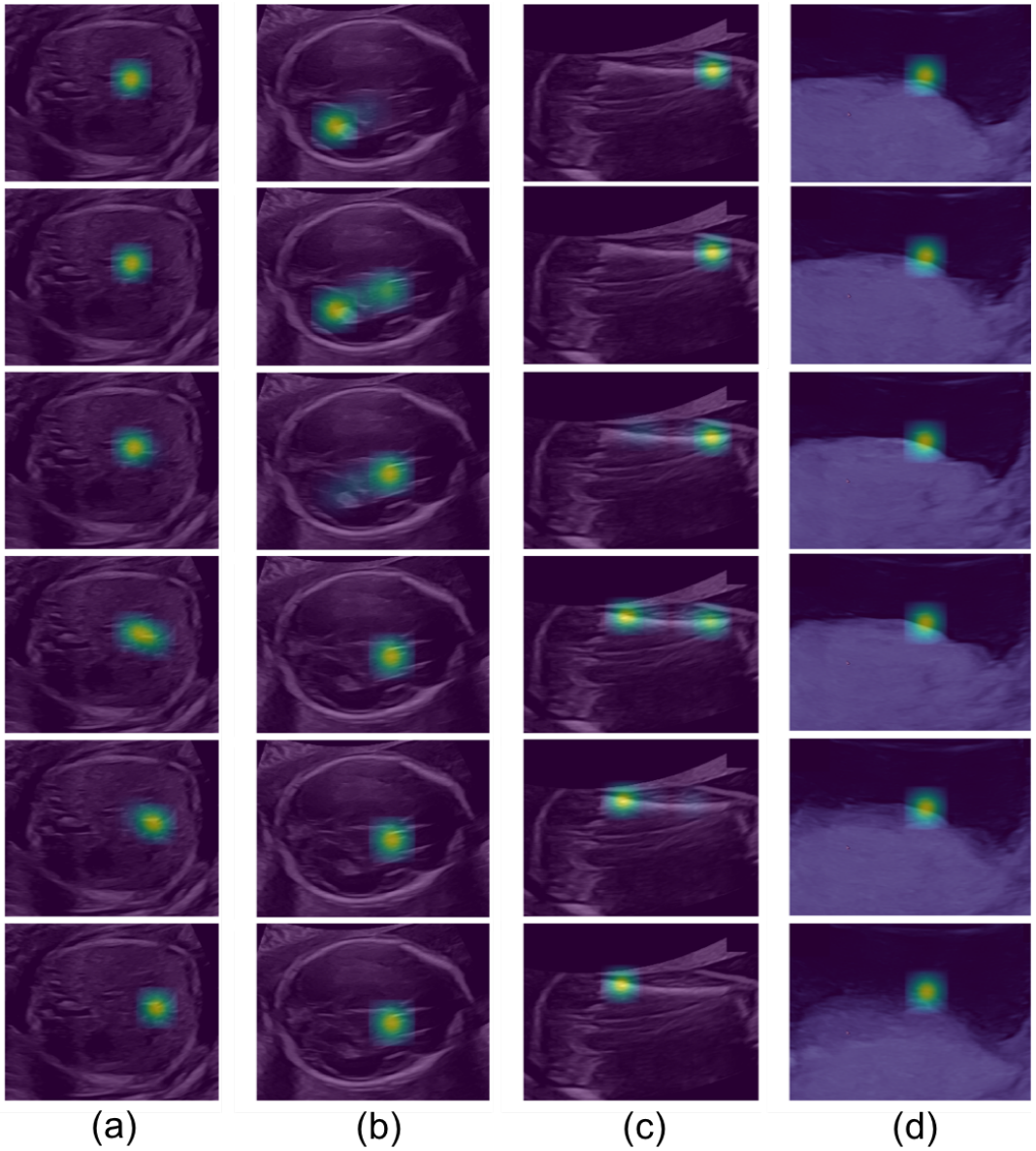


Figure 6.7: Sonographer visual attention maps on 6 consecutive frames of (a) Standard Abdomen sequence (b) Standard Head sequence (c) Standard Femur sequence and (d) background sequence.

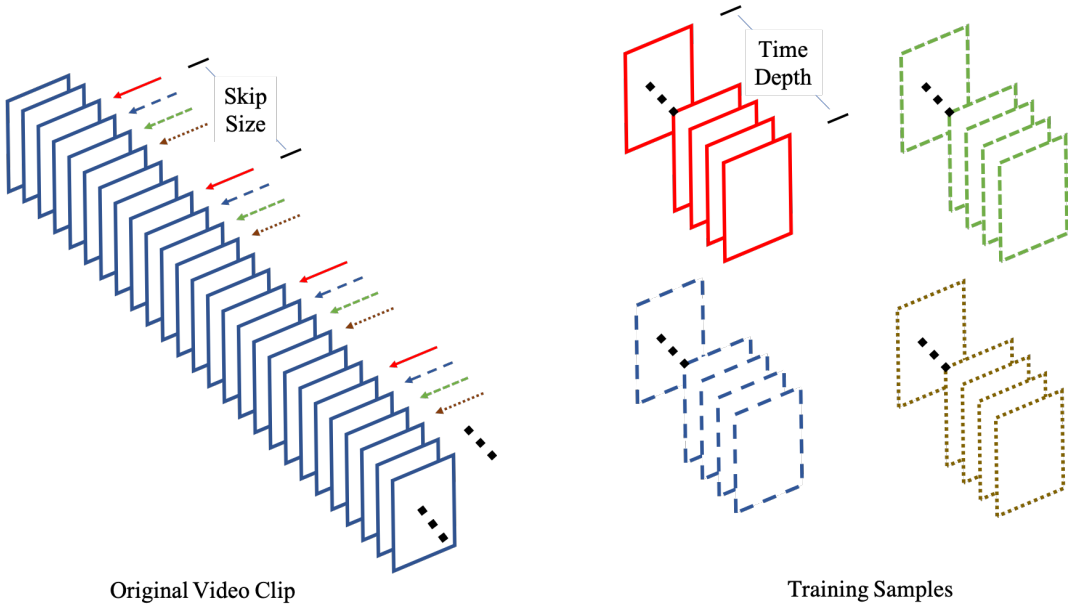


Figure 6.8: Cartoon showing sampling video frames with defined Time Depth and Skip Size from the original video clip. Different colors and line styles indicate 4 different training samples.

It is worth noticing that not all sampled sequences have corresponding sonographer visual attention maps for every frame in the sequence. The reason is sonographers are not guaranteed to be looking at the screen and they have to frequently check the position of ultrasound probe, talk to patients, or look at GUI to conduct measurements. In these cases, no gaze-data are recorded, thus no valid heatmaps will be generated; the sampled sequence containing frames with no corresponding heatmaps is discarded. It is worth noting that for future works if the time gap between missing gaze point or points is small (e.g. smaller than $100ms$) then it is possible to fill in the gap with linear interpolation.

6.5.2 Architecture

The architecture of T-SEN is described in Fig. 6.9. The network takes a sample input $\mathbf{X} = \{\mathbf{X}^1, ..., \mathbf{X}^T\} \in [|0, 255|]^{h \times w \times T}$, where H, W represent the height and width of input frames, and T the number of frames in a sample. For each US video frame, a CNN feature extractor is used to extract spatial feature representations ϕ, which are subsequently fed into a Temporal Attention Module (TAM), a recurrent module

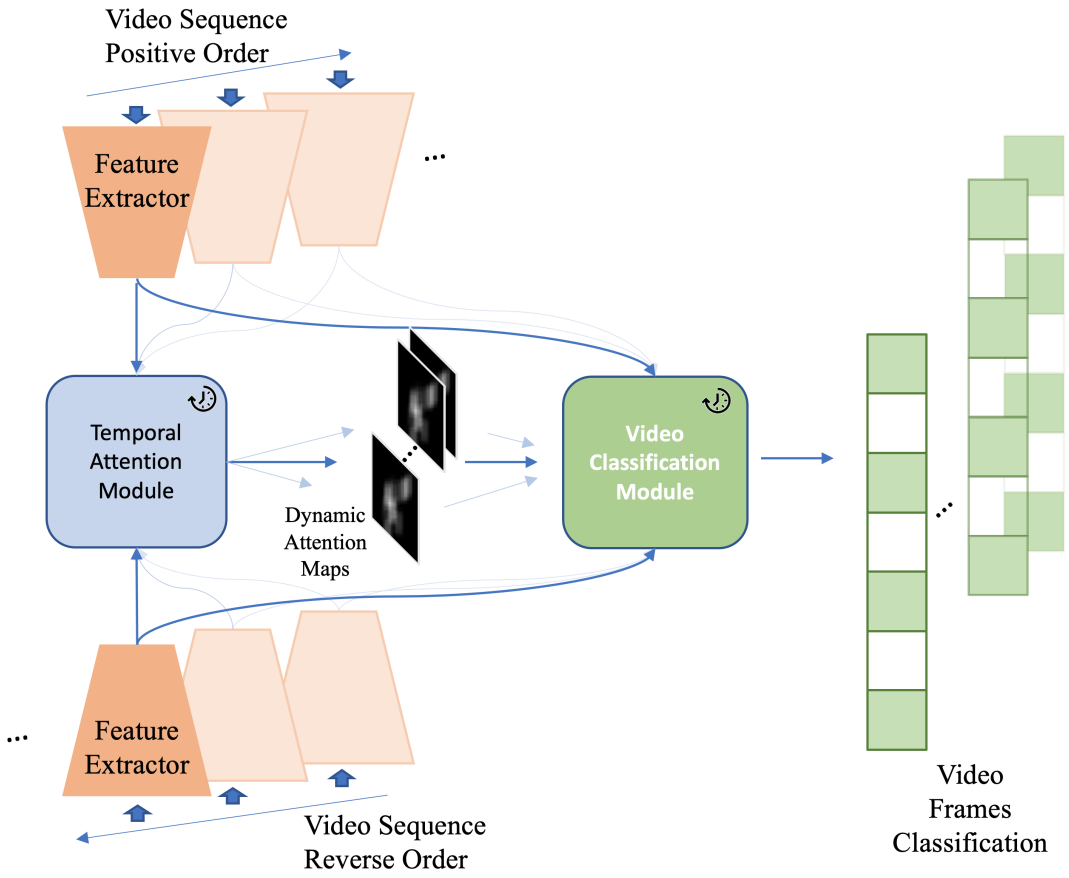


Figure 6.9: Architecture of Temporal-aware Multi-task SonoEyeNet, consisting of a Feature Extractor, Temporal Attention Module (TAM), and a Video Classification Module (VCM). The clock symbol in the figure indicates a recurrent module.

that produces Dynamic Attention Maps (DAM) $\mathbf{M} = \{\mathbf{M}^1, ..., \mathbf{M}^T\} \in [0,1]^{h \times w \times T}$ for each input video frame, where h, w represents the height and width of the predicted attention maps. The generated attention maps are then fed into the Video Classification Module (VCM), which is also recurrent, that predicts frame-wise class label $k \in [|1, K|]$ for each frame $\mathbf{X}^t, t \in [|1, T|]$, where $k = 7$ (bg Ab, ACP, bg Head, HCP, bg Femur, FLP, and an additional "other" class). The labels are one-hot encoded so that for a class $k \in [|1, K|]$, the corresponding target is $\mathbf{y} = (y_i)_1^K$ with $y_k = 1$ and $\forall i : i \neq k, y_i = 0$. In any trained module compared below, TAM and VCM use the same recurrent module, both in terms of bi-direction/uni-directional, and Convolutional LSTM (CLSTM)/ convolutional GRU (CGRU). In the following sections, schematics of the TAM and VCM are presented in the case of bi-directional RNNs.

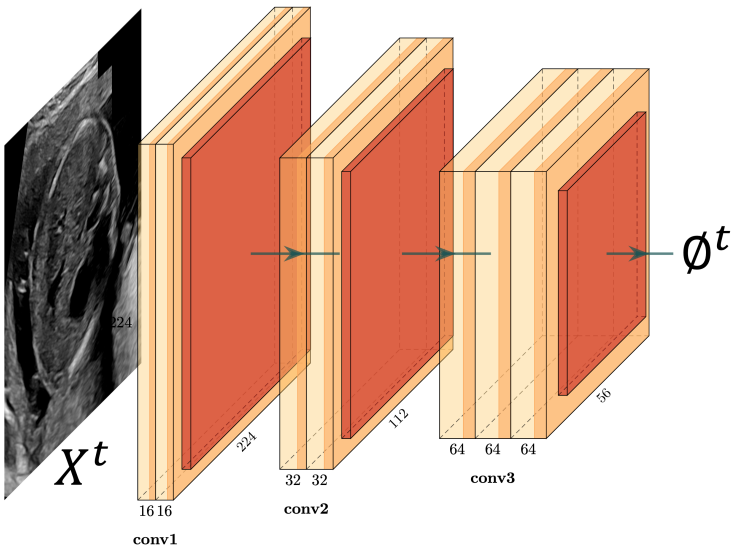


Figure 6.10: Schematic of the Feature Extractor used in T-SEN. $\mathbf{X}^t$ represents the t^{th} frame in an input US video clip; ϕ^t represents a tensor of feature representations from $\mathbf{X}^t$.

Feature Extractor

The Feature Extractor is derived from VGG-16 [**simonyan2014very**] with quartered number of convolutional kernels in each layer, as can be seen in Fig. 6.10. The Feature Extractor consists of three convolutional blocks, first two of which consist of two convolutional layers, and the third block consists of three convolutional layers. All convolutional layers use 3×3 convolutional kernels; the number of convolutional kernels used in three blocks are 16, 32, and 64, respectively.

Temporal Attention Module

A detailed architecture of the Temporal Attention Module (TAM) can be seen in Fig. 6.11, where yellow cubes represent tensors of feature maps from convolution operations, orange stripes represent activation function (ReLUs), and blue cubes represent tensors of hidden state of recurrent neural networks (Convolutional GRU or Convolutional LSTM). Since the recurrent modules are bi-directional, feature maps extracted by CNN $\phi^t, t \in [|1, T|]$ from sample video clips are fed into TAM in both positive and reverse order, as demonstrated in the Fig. 6.11. ϕ^t is passed through several convolutional layers to generate a Static Attention Map (SAM) $\tilde{\mathbf{M}}^t$,

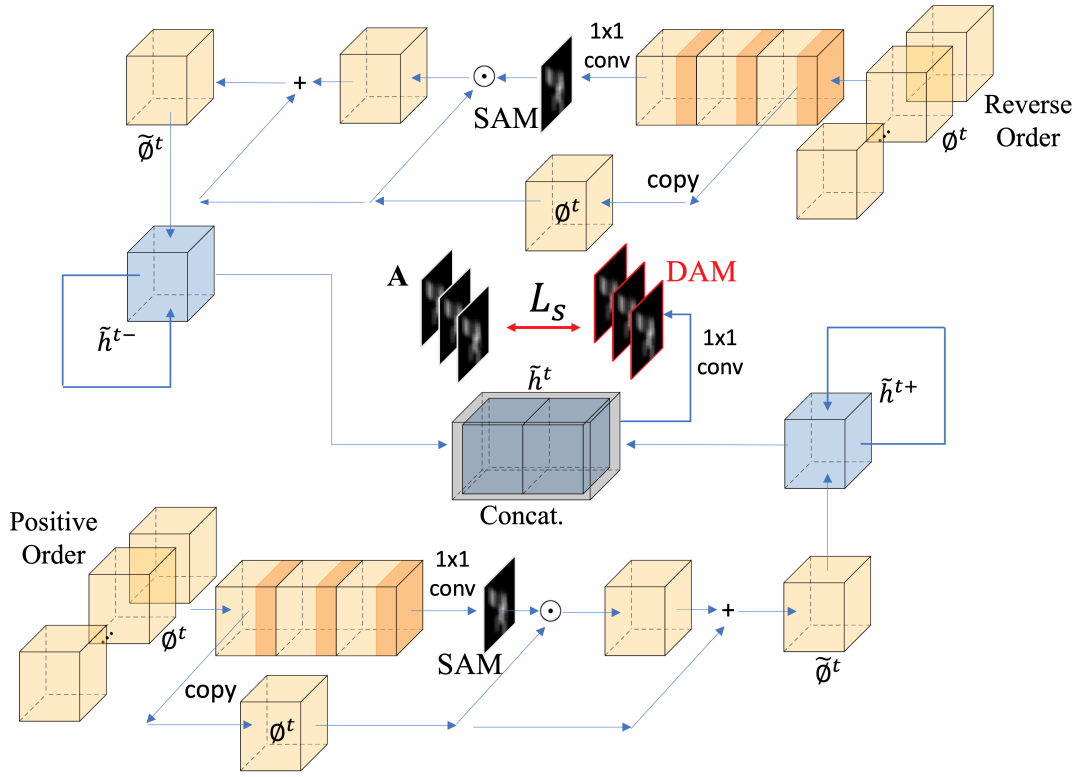


Figure 6.11: Schematic of the Temporal Attention Module (TAM).

which is then processed through a residual operation to generate $\tilde{\phi}^t$:

$$\tilde{\phi}^t = \phi^t \odot \tilde{\mathbf{M}}^t + \phi^t \tag{6.22}$$

$\tilde{\phi}^t$ from both the positive order ($\tilde{\phi}^{t+}$) and reverse order ($\tilde{\phi}^{t-}$) are each fed into a convolutional recurrent neural network to generate $\tilde{h}^{t+}$ and $\tilde{h}^{t-}$, respectively. These two hidden-states are concatenated to generate $\tilde{h}^t$, which, after further convolution and sigmoid activation, generates Dynamic Attention Map $\mathbf{M}^t$. Loss function between ground-truth visual attention maps $\mathbf{A}$ and $\mathbf{M}$ are defined in the Loss Functions section.

In the case when a uni-directional RNN is used, the branch that processes reverse order feature maps is discarded. Dynamic attention maps are generated from $\tilde{\phi}^{t+}$ directly.

Video Classification Module

A detailed architecture of the Video Classification Module (VCM) can be seen in Fig. 6.12. Feature maps ϕ^t from the t^{th} frame are fed into a bi-directional RNN to

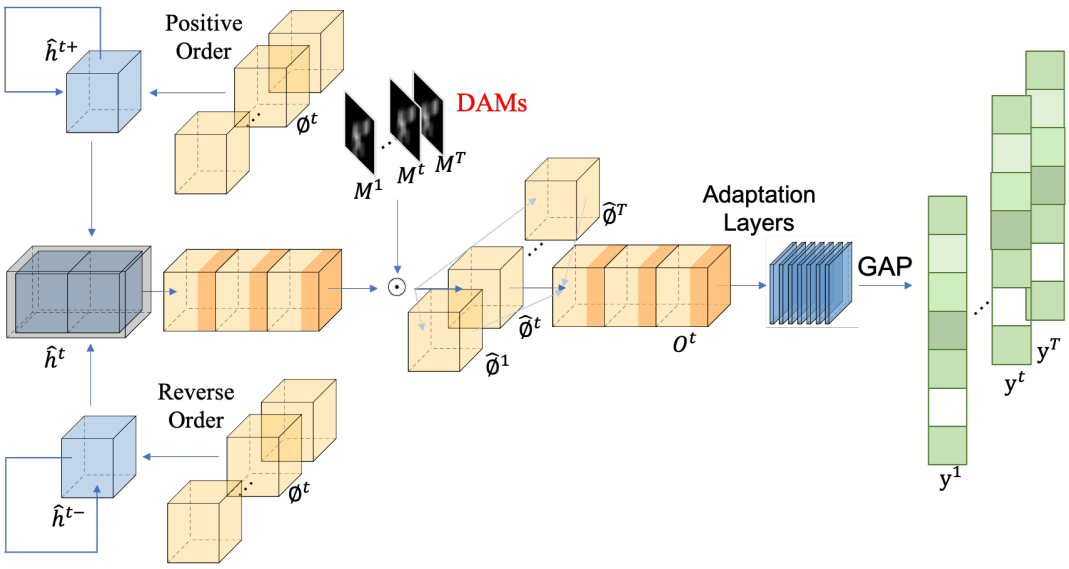


Figure 6.12: Schematic of the Video Classification Module (VCM).

generate $\hat{h}^{t+}$ and $\hat{h}^{t-}$ from the positive and reverse order, which are subsequently concatenated to form $\hat{h}^t$. After three convolution layers, the resultant feature maps are merged with $\mathbf{M}^t$ through element-wise production to produce $\hat{\phi}^t$. One class prediction is performed on each $\hat{\phi}^t$ through three convolutional layers, two adaptation layers, and a global average pooling layer before producing video-wise class prediction $\hat{y}$. The loss function for classification is discussed in the next section.

Similarly, in the case when a uni-direcional RNN is used, the branch that processes reverse-order feature maps is discarded. $\hat{h}^{t+}$, instead of $\hat{h}^t$, is used for further processing to predict the input classes.

6.5.3 Loss Functions

The network is trained with three losses. For an input video clip $\mathbf{X} = \{\mathbf{X}^1, ..., \mathbf{X}^T\} \in [|0, 255|]^{h \times w \times T}$, the first is a **classification loss** L_c between class-prediction $\hat{\mathbf{y}} = \{\hat{\mathbf{y}}^1, ..., \hat{\mathbf{y}}^T\} \in [0, 1]^{C \times T}$ and ground-truth class label $\mathbf{y} = \{\mathbf{y}^1, ..., \mathbf{y}^T\} \in \{0, 1\}^{C \times T}$, with C being the number of classes. Cross-entropy (CE) loss is traditionally used; to tackle class-imbalance problem, a variant of CE, the Focal Loss [**lin2017focal**] was used instead. The second is a **saliency loss** L_s, a loss between each pair of ground-truth visual attention map $\mathbf{A}^t$ and DAM $\mathbf{M}^t$. Kullback-Leibler divergence is used. The third is a **temporal regularization loss** L_t that encourages alignment

of dynamic attention maps over time. Soft Dynamic Time Warping Loss (sDTW) [**cuturi2017soft**] is used. The total loss is represented as:

$$L = \alpha L_c + \beta L_s + \lambda L_t \tag{6.23}$$

where α, β, λ are hyper-parameters that control the contribution of each loss to the total loss.

Classification Loss

Focal Loss [**lin2017focal**] is a variant of cross-entropy loss that allows hard samples to contribute more to total loss, and down-weight the contribution from easy samples. In a multi-class classification problem with C classes, a standard multi-class cross-entropy (CE) loss between $\mathbf{y}^t$ and $\hat{\mathbf{y}}^t$ is defined as:

$$CE = -\sum_{i}^{C} \mathbf{y}_i^t \log(\hat{\mathbf{y}}_i^t) \tag{6.24}$$

In a deep learning context, $\mathbf{y}$ is normally a one-hot encoded vector with "1" at the correct class index and "0" elsewhere, while a score vector $\mathbf{s}$ predicted by a neural network is passed through a *softmax* layer f in order to sum up to 1 so as to approximate the probability distribution of the input, so that $\hat{\mathbf{y}} = f(\mathbf{s})$.

$$CE = -\sum_{i}^{C} \mathbf{y}_i^t \log(f(\mathbf{s}_i)) \tag{6.25}$$

$$\hat{y}_n = f(\mathbf{s}_n) = \frac{e^{\mathbf{s}_n}}{\sum_i^c e^{\mathbf{s}_i}} \tag{6.26}$$

Due to one-hot encoding, only the positive class keeps its term in the loss. Let p be the index of the true class, then $\mathbf{y}_p^t = 1$ and $\forall i : i \neq p, \mathbf{y}_i^t = 0$. Let $\mathbf{s}_p$ be the output score of the true class and $\hat{\mathbf{y}}_p^t = f(\mathbf{s}_p)$ be the posterior probability that the input belongs to the true class. Then, cross-entropy loss can be rewritten as:

$$CE = -\log\left(\frac{e^{\mathbf{s}_p}}{\sum_i^c e^{\mathbf{s}_i}}\right) = -\log(f(\mathbf{s}_p)) = -\log(\hat{\mathbf{y}}_p) \tag{6.27}$$

In practice, class imbalance is a problem for classification models by overwhelming training with easy negatives that contribute no useful learning signals, leading to

degenerate models. To tackle this, an α-balanced CE loss can be used. A weighting factor $\alpha \in [0, 1]$ is introduced:

$$\alpha_{tr} = \frac{1}{\text{freq}_p} \tag{6.28}$$

where $freq_p$ represents the frequency of true class. An α-balanced CE loss can be written as:

$$CE = -\alpha_{tr} \log(\hat{\mathbf{y}}_p^t) \tag{6.29}$$

While α is able to balance between different classes, it does not differentiate between hard/easy examples. Instead, Lin *et al.* [**lin2017focal**] developed Focal Loss that down-weights the loss from easy samples and focus training on hard negatives by adding a modulating factor $(1 - \hat{y}_p)^{\xi}$ to cross-entropy loss:

$$FL(p_t) = -(1 - \hat{\mathbf{y}}_p^t)^{\xi} \log(\hat{\mathbf{y}}_p^t) \tag{6.30}$$

Thus, for $\mathbf{X}$ the classification loss L_c can be written as:

$$L_c = -\sum_{t=1}^{T} (1 - \hat{\mathbf{y}}_p^t)^{\xi} \log(\hat{\mathbf{y}}_p^t) \tag{6.31}$$

Saliency Loss

It is most common to use the Kullback-Leibler Divergence (KLD) to measure the difference between two distributions. In the context of visual attention prediction, the KLD between predicted dynamic visual attention maps $\mathbf{M} = \{\mathbf{M}^1, ..., \mathbf{M}^T\} \in [0, 1]^{h \times w \times T}$ and ground-truth sonographer visual attention map $\mathbf{A} = \{\mathbf{A}^1, ..., \mathbf{A}^T\} \in [0, 1]^{h \times w \times T}$ can be written as:

$$L_s = D_{KL}(\mathbf{A}||\mathbf{M}) = -\sum_{t=1}^{T} \sum_{i=1}^{h} \sum_{j=1}^{w} \mathbf{A}_{i,j}^t log \frac{\mathbf{M}_{i,j}^t}{\mathbf{A}_{i,j}^t} \tag{6.32}$$

Temporal Regularisation Loss

Soft Dynamic Time Warping (sDTW) [**cuturi2017soft**] is a differentiable function, derived from Dynamic Time Warping (DTW) [**sakoe1990dynamic**], that can be used as a loss function to enforce alignment of two time series. **Dynamic Time Warping (DTW)** measures the discrepancy between two time series by computing the best possible alignment between the two time series x and y with lengths n and m. It first computes the $n \times m$ pairwise cost matrix $\mathbf{\Delta}(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{n \times m}$ between points (Fig. 6.13(A)), and then finds an Alignment Matrix Ψ that defines a path on a n $\times$ m matrix that connect the upper-left (1, 1) matrix entry to the lower-right (n, m) one using only down $\downarrow$,right $\rightarrow$ and down-right $\searrow$ moves. The minimised sum of cost found by solving a dynamic program (DP) problem using Bellman's recursion [**bellman1952theory**] with a quadratic (nm) cost is the DTW score.

Specifically, for time series $\mathbf{a}$ and $\mathbf{b}$ with length n and m, let $\mathbf{\Psi}_{n,m} \subset \{0, 1\}^{n \times m}$ be a set of binary alignment matrices describing the path connecting top-left to bottom-right of cost matrix $\mathbf{\Delta}$ using only right, down, or right-down connections, as presented in Fig. 6.13(B). The DTW score is represented by:

$$\mathrm{DTW}(\mathbf{a}, \mathbf{b}) = \min_{\Psi \in \mathbf{\Psi}_{n,m}} \langle \Psi, \mathbf{\Delta}(\mathbf{a}, \mathbf{b}) \rangle \tag{6.33}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product operator.

Algorithmically, the optimal alignment matrix Ψ^* that minimises $\mathrm{DTW}(\mathbf{a}, \mathbf{b})$ is not explicitly calculated; rather, an intermediary alignment cost matrix $\mathbf{R}$ (Fig. 6.13(D)) is constructed using Bellman's Recursion (Fig. 6.13(C)) to calculate the minimum summed cost (*i.e.* sDTW) achieved by Ψ^*. Bellman's recursion to construct $\mathbf{R}$ is calculated through the following equations:

$$r_{i,j} = \delta_{i,j} + \min(r_{i-1,j}, r_{i,j-1}, r_{i-1,j-1}) \tag{6.34}$$

$$\delta_{i,j} = MSE(\mathbf{a}_i, \mathbf{b}_j) \tag{6.35}$$

where $r_{i,j}$ represents an element in $\mathbf{R}$ and $\delta_{i,j}$ an element in $\mathbf{\Delta}$. When $\mathbf{R}$ is complete through Bellman's Recursion,

$$\mathrm{DTW}(\mathbf{x}, \mathbf{y}) = r_{n,m} \tag{6.36}$$

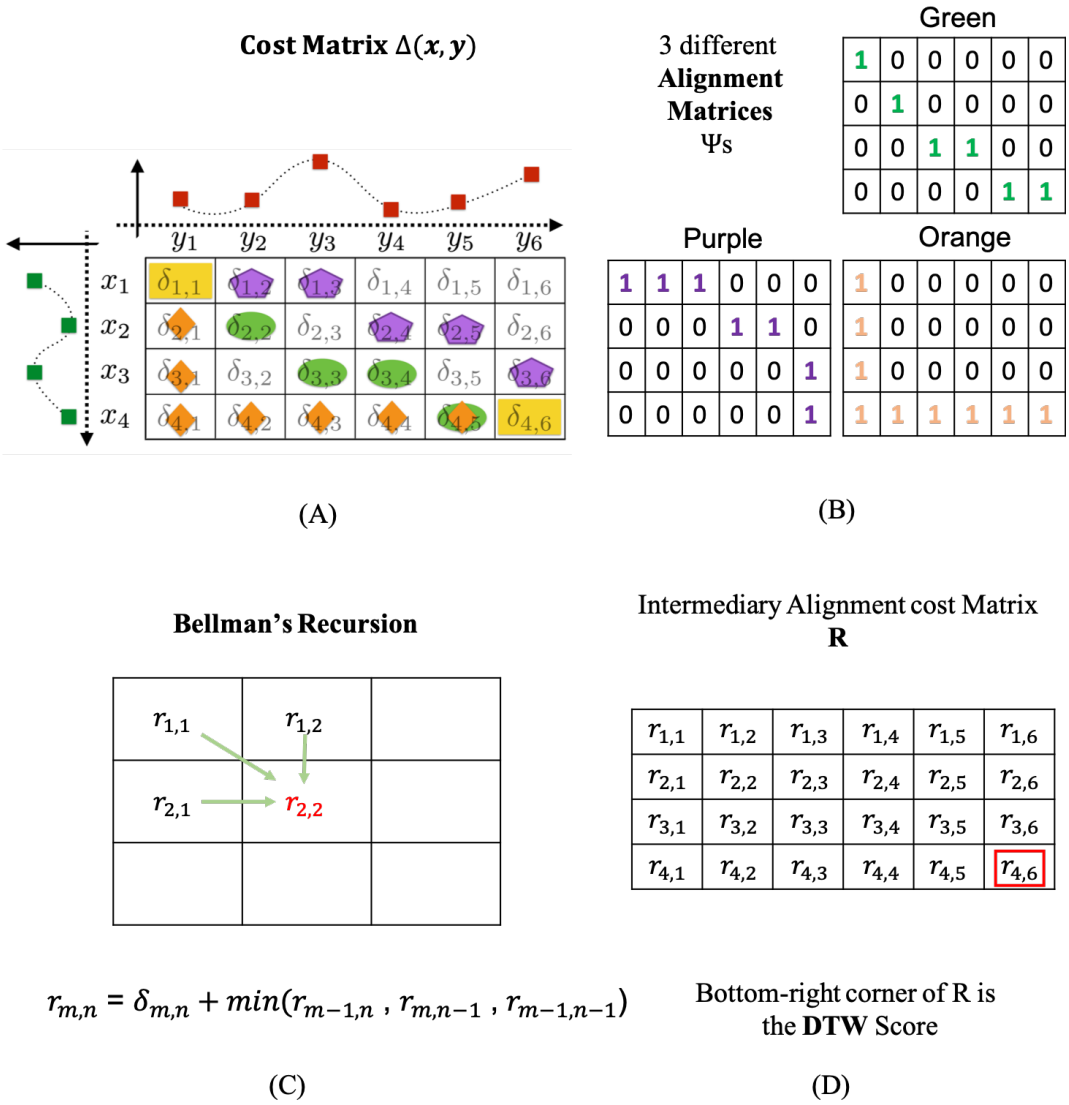


Figure 6.13: Schematic showing how Dynamic Time Warping (DTW) works. (A) A cost matrix Δ between time series $\mathbf{x}, \mathbf{y}$ with orange, green and purple as color codes for three different possible connections between top-left and bottom right elements. (B) Binary Alignment matrices with corresponding color codes. (C) Bellman's Recursion (D) A complete Intermediary Alignment cost Matrix R. The figure is adapted from [**cuturi2017soft**].

Though an excellent measure for discrepancies between time series, it is not a differentiable function that can be used as a loss function. A smoothed formulation of DTW, called **sDTW** [**cuturi2017soft**], is differentiable and can be used specifically for this purpose. A soft min operator with a smoothing parameter $\gamma > 0$ is used, instead of the normal minimum function:

$$\min^{\gamma}\{a_1, ..., a_n\} = -\gamma \log \sum_{i=1}^{n} e^{\frac{-a_i}{\gamma}} \tag{6.37}$$

In the context of visual attention map prediction, the sDTW between predicted dynamic visual attention maps $\mathbf{M} = \{\mathbf{M}^1, ..., \mathbf{M}^T\} \in [0,1]^{h \times w \times T}$ and ground-truth sonographer visual attention map $\mathbf{A} = \{\mathbf{A}^1, ..., \mathbf{A}^T\} \in [0,1]^{h \times w \times T}$:

$$\mathbf{dtw}_{\gamma}(\mathbf{M}, \mathbf{A}) = \min^{\gamma}\{\langle \Psi, \mathbf{\Delta}(\mathbf{M}, \mathbf{A}) \rangle, \Psi \in \mathbf{\Psi}_{T,T}\} \tag{6.38}$$

and Bellman's Recursion can be written as:

$$r^{\gamma}_{i,j} = \delta_{i,j} + \min^{\gamma}(r^{\gamma}_{i-1,j}, r^{\gamma}_{i,j-1}, r^{\gamma}_{i-1,j-1}) \tag{6.39}$$

The final sDTW score is:

$$\mathbf{dtw}_{\gamma}(\mathbf{M}, \mathbf{A}) = r^{\gamma}_{T,T} \tag{6.40}$$

The algorithm for calculating $\mathbf{dtw}_{\gamma}(\mathbf{M}, \mathbf{A})$ as well as the intermediate alignment cost matrix R^{γ} is seen below.

6.5.4 Performance Metrics

Classification Metrics

Similar to the previous two chapters, *Precision, Recall* and *F1-score* are used to measure classification performances of all models. For each model, these metrics are reported on a per-anatomy basis. In addition, the overall performance across all anatomies is reported for each metric using *macro average*, *i.e.* un-weighted average of the performance across anatomies due to heavy class imbalance.

Algorithm 2 Forward recursion to compute $\mathbf{dtw}_\gamma(\mathbf{M}, \mathbf{A})$ and $\mathbf{R}^\gamma$

Require:
predicted dynamic visual attention maps $\mathbf{M} \in [0,1]^{h \times w \times T}$
ground-truth sonographer visual attention maps $\mathbf{A} \in [0,1]^{h \times w \times T}$
smoothing temperature term $\gamma > 0$
empty intermediary alignment cost matrix $\mathbf{R}^\gamma \in \mathbb{R}^{T \times T}$

for $j = 1, ..., T$ **do**
  for $i = 1, ..., T$ **do**
    $\delta_{i,j} = MSE(\mathbf{M}^i, \mathbf{A}^j)$
  end for
end for
$r^\gamma_{0,0} = 0$; $r^\gamma_{0,i} = r^\gamma_{j,0} = \infty$; $i \in [|1, T|], j \in [|1, T|]$ ▷ Initialisation
for $j = 1, ..., T$ **do**
  for $i = 1, ..., T$ **do**
    $r^\gamma_{i,j} = \delta_{i,j} + \min^\gamma(r^\gamma_{i-1,j}, r^\gamma_{i,j-1}, r^\gamma_{i-1,j-1})$
  end for
end for
return $(r^\gamma_{T,T}, \mathbf{R}^\gamma)$

Static Saliency Metrics

Similar to the previous chapter, five static saliency metrics were used to quantify the similarity between ground-truth and predicted visual attention maps: Area Under ROC Cuvre (*AUC*), Normalized Scanpath Saliency (*NSS*), Information Gain (*IG*), Similarity (*Sim*), and Person's Correlation Coefficient (*CC*). In addition, the Kullback-Leibler divergence (*KLD*) is also reported. All metrics are reported on a per-anatomy basis as well as on a macro-average basis.

Scanpath Metrics

Following MultiMatch [**dewhurst2012depends**, **jarodzka2010vector**], a set of metrics used to measure scanpath similarities, four different metrics are calculated for two scanpaths: Vector Similarity (*VecSim*), Length Similarity (*LenSim*), Direction Similarity (*DirSim*), and Position Similarity (*PosSim*). A scanpath is defined as a set of fixation points on consecutive video frames and the saccadic transitions between each pair of fixation points, as can be seen in Fig 6.14(A) where red dots represent fixation points and the dotted arrows represent saccadic transitions; different color-coding for visual attention maps indicate different time points. The

MultiMatch metrics represents the scanpath as a set of saccadic vectors on a 2-D plane, as can be seen in Fig 6.14(B); each saccadic vector originates from the fixation point at time point t and points to the fixation point at time point $t+1$. For the purpose of this study, the fixation point on a visual attention map is defined as the (x, y)-coordinate of its global maxima.

Specifically, for two scanpaths $\mathbf{P}_1 = \{\mathbf{p}_1^1, ..., \mathbf{p}_1^T\}$ and $\mathbf{P}_2 = \{\mathbf{p}_2^1, ..., \mathbf{p}_2^T\}$ where each element in the scanpath is a fixation point x, y in a 2-D space, their corresponding saccadic vector representations $S_1 = \{\mathbf{v}^1, ..., \mathbf{v}^{T-1}\}$ and $S_2 = \{\mathbf{u}^1, ..., \mathbf{u}^{T-1}\}$ are calculated such that $\mathbf{v}^t = \mathbf{p}_1^{t+1} - \mathbf{p}_1^t$ and $\mathbf{u}^t = \mathbf{p}_2^{t+1} - \mathbf{p}_2^t$.

Four metrics are calculated, as can be seen in Fig. 6.14(C):

$$VecSim = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\|\mathbf{v^t} - \mathbf{u^t}\|}{2 \times d} \tag{6.41}$$

$$LenSim = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \|\mathbf{v^t}\| - \|\mathbf{u^t}\| \tag{6.42}$$

$$DirSim = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{\pi} cos^{-1} \frac{\mathbf{v^t} \cdot \mathbf{u^t}}{\|\mathbf{v^t}\| \|\mathbf{u^t}\|} \tag{6.43}$$

$$PosSim = 1 - \frac{1}{T} \sum_{t=1}^{T} \frac{\|\mathbf{p}_1^t - \mathbf{p}_2^t\|}{2 \times d} \tag{6.44}$$

where d represents the diagonal size of the visual attention maps. It is worth noting that Duration Similarity ($DurSim$) in the original MultiMatch algorithm is not calculated in this study, because the input video frames are sampled at 30 Hz and the duration of each fixation is thus approximately 0.033 seconds. The scanpath *simplification* and *alignment* processes in the MultiMatch algorithm are not performed here, as they are pre-processing steps only relevent to raw gaze-tracking data.

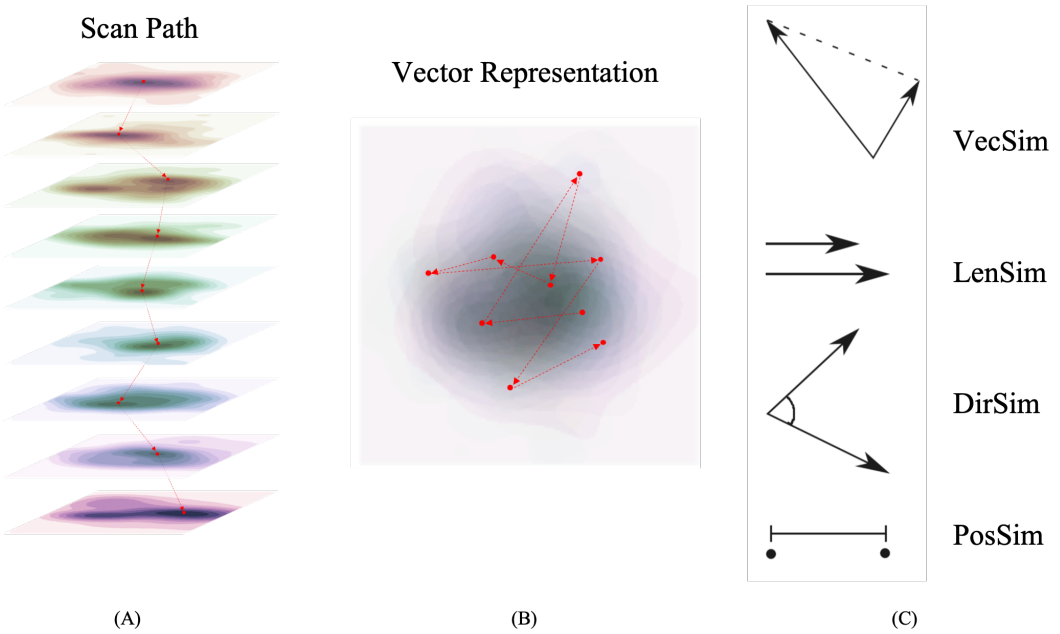


Figure 6.14: Schematic outlining the MultiMatch scanpath similarity metrics. (A) Cartoon representation of a scanpath through 9 visual attention maps on consecutive US video frames. (B) Vector representation of the scanpath on 2-D plane (C) Cartoon of the 4 similarity metrics. The figure is adapted from [**dewhurst2012depends**].

6.5.5 Training Details

All T-SEN variants are initialised using the pre-trained weights of the first three convolutional blocks of SonoNet-16; all other layers were initialized using a zero-mean Gaussian distribution with standard deviation of 0.01. They are all trained using adaptive moment estimation (Adam) with an initial learning rate of 2×10^{-4} for 100 epochs. All input samples $\mathbf{X}$ were sampled from video clips with a skip size of 10 and time depth T of 10 with a mini-batch size of 16. T-SEN models were trained with three losses: Classification loss L_c, Saliency loss L_s, and temporal regularization loss L_t. α and β for L_c and L_s were set to 0.5 to give the two tasks equal weighting. λ for L_t was set at 0.01 after searching through 5 logarithmically spaced weights between 10^{-4} and 1. The best performing model used focal loss for L_c and the weight ξ was set at a value of 2 to suppress gradients of easy negative samples; other variants used cross-entropy loss for L_c. The dataset was split at scan-level into 5 folds for cross-validation.

In order to tackle severe class imbalance, the frequency of sequence-level labels for standard and non-standard sequences for each anatomy (Abdomen, Head, Femur) and "other" class is calculated. During training, each sequence is drawn with a probability equal to the inverse of its sequence-level label's frequency.

The SonoNet models and MSEN were fine-tuned as baseline comparison; SonoNet was initialized with pre-trained weights published by the authors and fine-tuned for 50 epochs using Adam optimizer with a learning rate of 10^{-4}; MSEN was trained from randomly-initialised weights. Video frames are treated as independent images and sampled at the inverse of the frequencies of their corresponding frame-wise label.

6.5.6 Results

Temporal Visual Attention modelling

Different T-SEN variants were trained to model sonographer visual attention. The models presented in this section are named by the specifications of RNNs used in the TAM. For example, the model "biCLSTM" indicates that the model used bi-directional Convolutional LSTM in the TAM, and "biCLSTM+sDTW" indicates that it was trained with the additional Temporal Regulariser Loss L_t using sDTW; "biCLSTM" wasn't trained with L_t. As mentioned before, all variants' VCM share the same RNNs specifications with TAM; the VCM of all variants reported in this section are trained with Focal Loss as L_c.

Qualitative Assessment. Predicted visual attention maps generated by different T-SEN variants on three samples of input clips in the test set can be seen in Fig. 6.15 for the fetal abdomen, Fig. 6.16 for the fetal head, and Fig. 6.17 for the fetal femur. In general, the best performing model in all three anatomies is "biCLSTM+sDTW", demonstrating good synchronisation of saccadic transitions with the ground-truth sonographer's visual attention. In Fig. 6.15, as the ACP gradually appears, the predicted visual attention transits from the middle of the view to the area between stomach bubble and umbilical vein in the same fashion as the GT; in Fig. 6.16, the predicted visual attention follows the ground truth (GT) by scanning along the center-line of the brain before focusing on *cavum septum pellucidum*; in Fig. 6.17,

the prediction replicates the scanning behavior along the femur bone, even with the transient appearance of other structures. The model "biCLSTM", trained without temporal regularization, demonstrated less focused attention and the fixations are not synchronized with the GT. Similar improvement can be seen by comparing the prediction of "biCGRU+sDTW" with "biCGRU".

Convolutional LSTM models in general demonstrated better capacity to learn temporal visual attention transitions than those of convolutional GRU models. Overall, the predicted visual attention of convGRU models are more spread out with no clear point of fixations compared to convLSTM models. In addition, it is observed that bi-directional models are able to predict higher quality visual attention maps compared to uni-directional models.

Quantitative Assessment. In order to quantitatively assess the visual attention prediction performances of different T-SEN variants, static saliency scores (Table. 6.1) as well as scanpath similarity scores (Table. 6.2), the MultiMatch metrics, were measured on test set for each variant. Higher scores in all metrics but *KLD* show higher performance; the lower the *KLD*, the better model performs, as indicated by the downward arrow next to it in Table. 6.1.

Confirming the observation made in the qualitative assessment, "biCLSTM+sDTW" model outperforms other models in all **static saliency scores**, reaching mean scores across all classes of 64.8% for *AUC*, 69.9% for *CC*, 57.7% for *SIM*, 1.50 for *IG*, 2.76 for *NSS*, and 1.06 for *KLD*, as seen in Table. 6.1. Using sDTW as a temporal regularizer significantly improves model performances in all metrics; this improvement is more prominent on biCGRU models than on biCLSTM. "biCLSTM" outperforms "uniCLSTM" in 3 out of 6 metrics: *SIM*, *IG*, and *NSS*, while "biCGRU" outperforms "uniCGRU" in 4 out of 6 metrics: *AUC*, *CC*, *IG* and *NSS*, though the improvement in *IG* and *NSS* are not significant. All T-SEN models outperform the baseline "MSEN" on all metrics except for the cases in *IG* and *NSS*, where "MSEN" slightly outperforms "uniCGRU".

Similar to the results in static saliency scores, the performance of "biCLSTM+sDTW" exceeds those of other variants in all of the **scanpath similarities scores**.

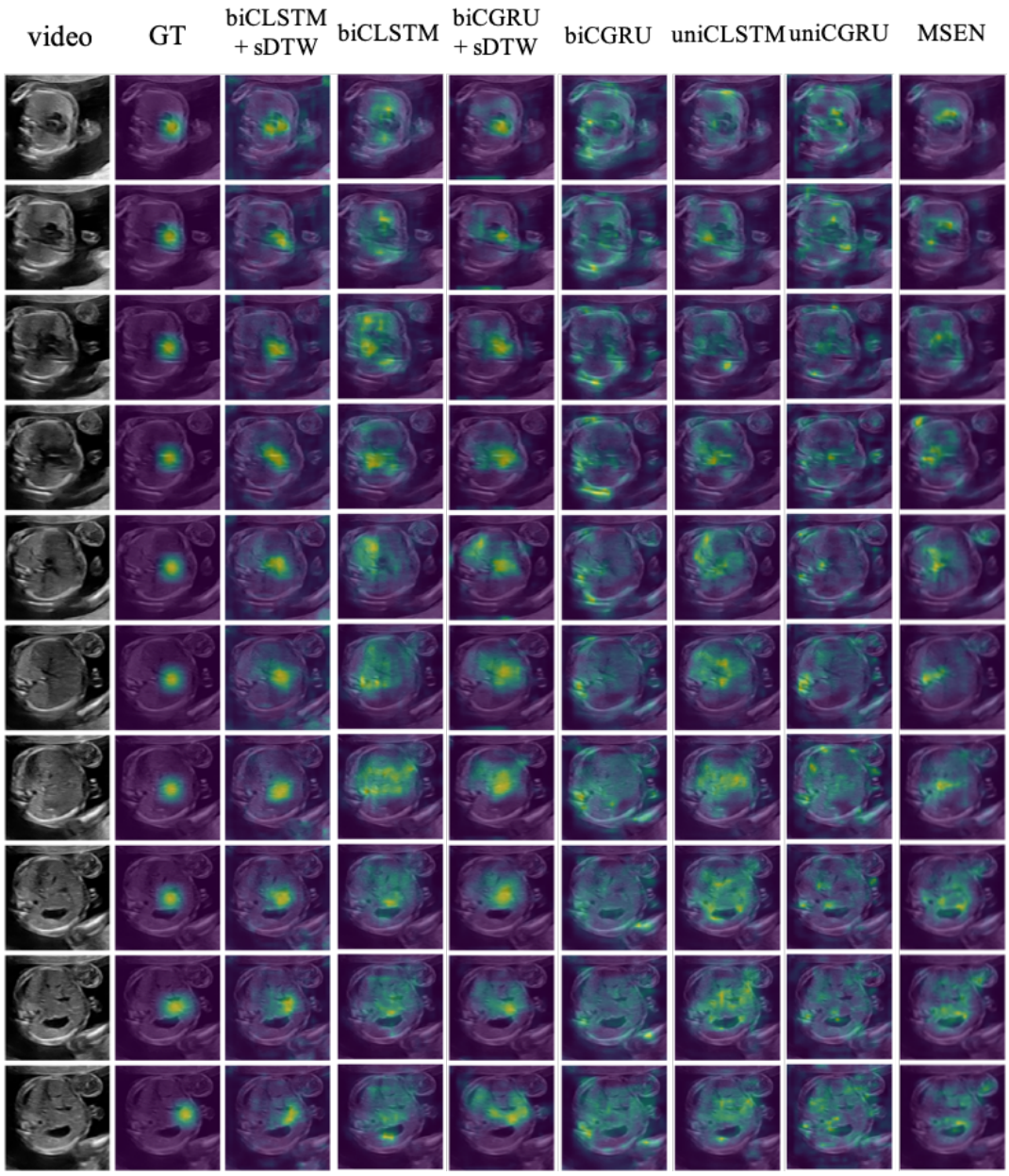


Figure 6.15: Visual attention maps generated by different T-SEN variants on an example of fetal abdomen clip. From left to right: US video frames, Ground-truth (Sonographer's actual attention map), biCLSTM+sDTW, biCLSTM, biCGRU+sDTW, biCGRU, uniCLSTM, uniCGRU, MSEN

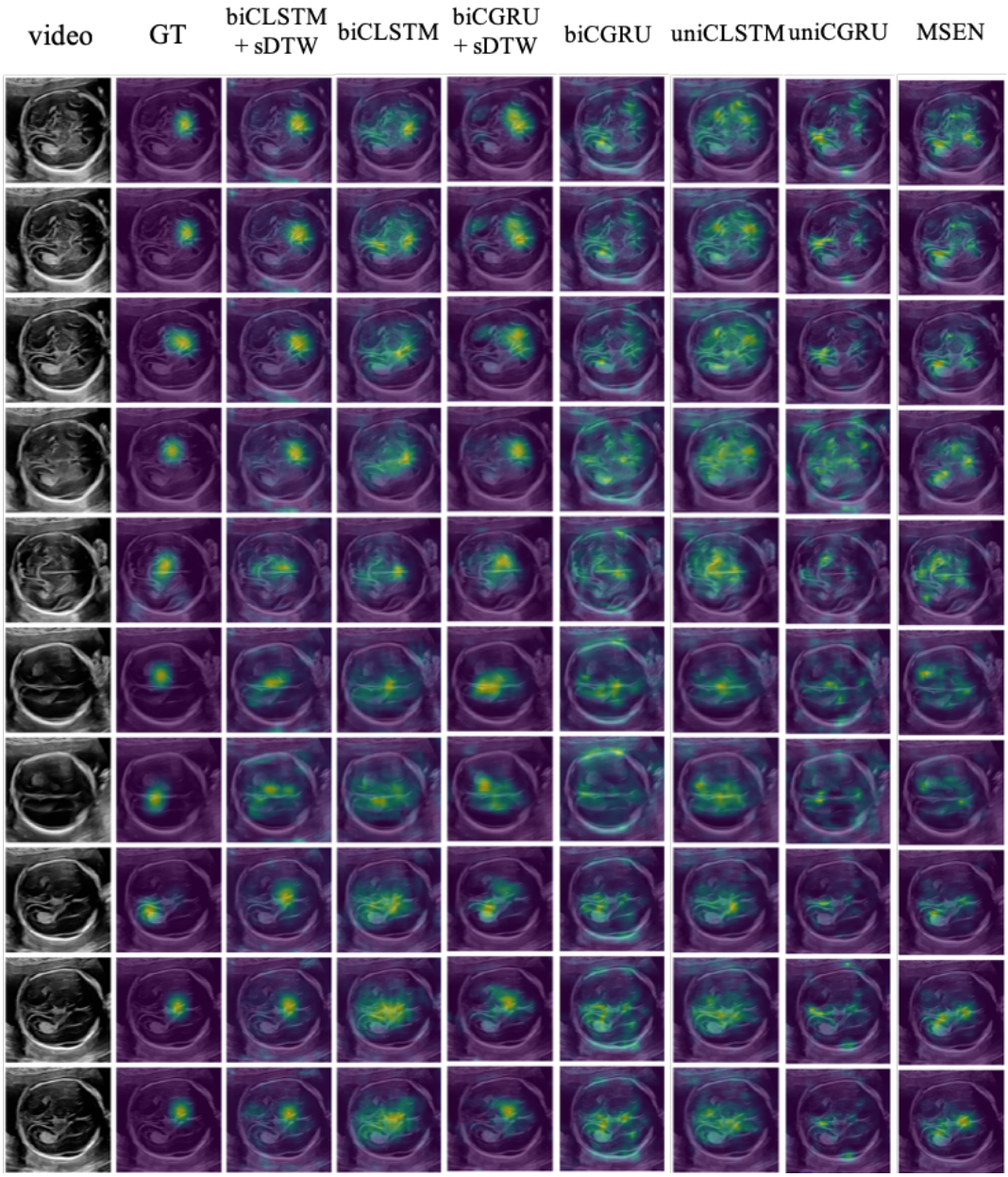


Figure 6.16: Visual attention maps generated by different T-SEN variants on an example of fetal head clip. From left to right: US video frames, Ground-truth (Sonographer's actual attention map), biCLSTM+sDTW, biCLSTM, biCGRU+sDTW, biCGRU, uniCLSTM, uniCGRU, MSEN

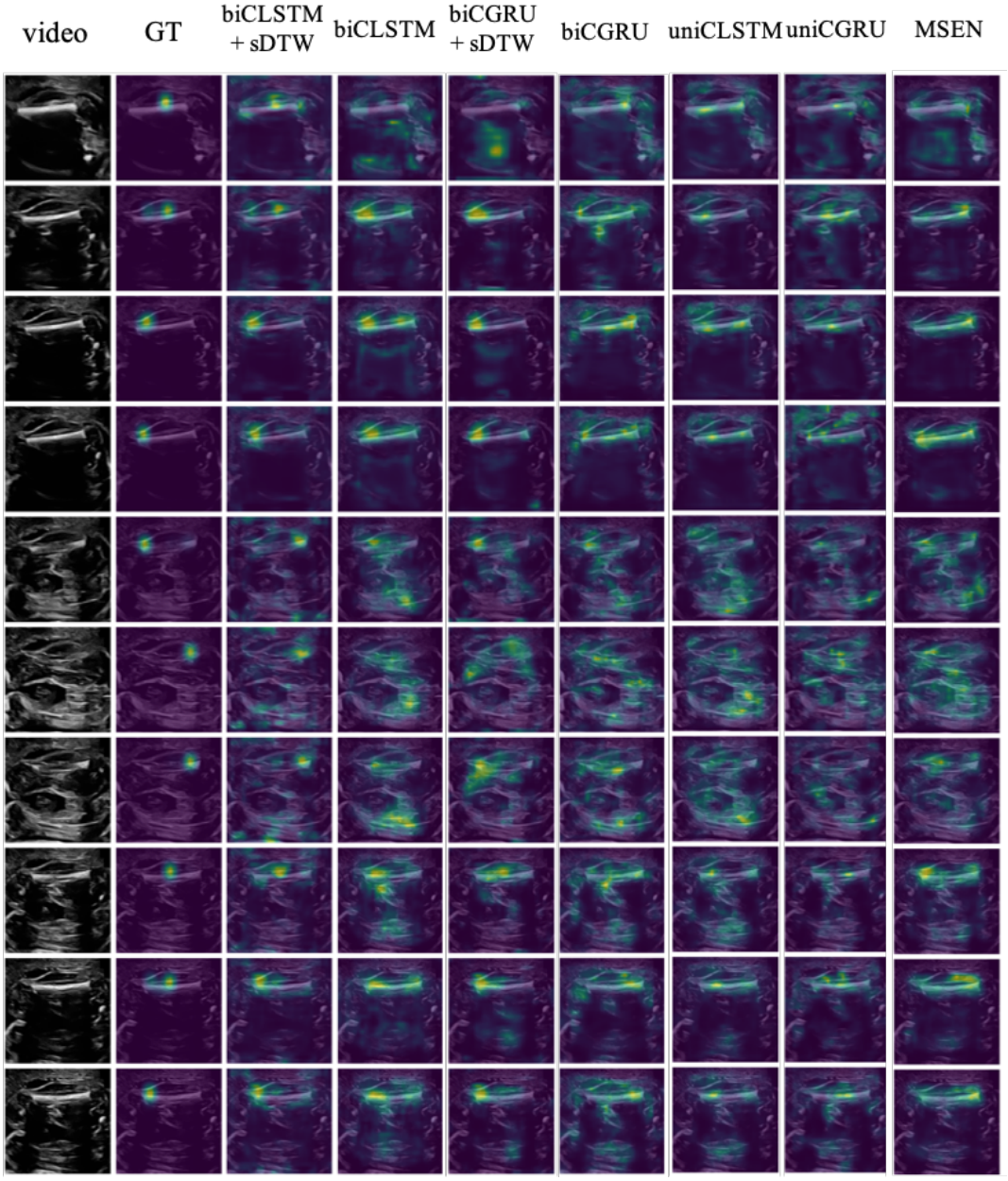


Figure 6.17: Visual attention maps generated by different T-SEN variants on an example of fetal femur clip. From left to right: US video frames, Ground-truth (Sonographer's actual attention map), biCLSTM+sDTW, biCLSTM, biCGRU+sDTW, biCGRU, uniCLSTM, uniCGRU, MSEN

Table 6.1: Static saliency scores of different models. Downward Arrow ↓ indicates that lower score refers to better performance

	AUC [%]	CC[%]	SIM [%]	IG	NSS	KLD ↓
biCLSTM+sDTW	**64.8** ± 0.6	**69.9** ± 1.1	**57.7** ± 2.6	**1.50** ± 0.02	**2.76** ± 0.05	**1.06** ± 0.08
biCLSTM	62.3 ± 0.6	41.2 ± 2.2	36.8 ± 3.1	1.27 ± 0.04	2.16 ± 0.01	1.79 ± 0.12
biCGRU+sDTW	54.2 ± 0.4	51.9 ± 2.8	36.7 ± 2.5	0.71 ± 0.02	2.18 ± 0.05	1.50 ± 0.04
biCGRU	47.5 ± 0.5	29.6 ± 2.4	19.8 ± 3.0	0.12 ± 0.04	1.32 ± 0.07	2.08 ± 0.05
uniCLSTM	61.4 ± 0.2	41.7 ± 1.0	33.0 ± 2.7	0.18 ± 0.03	1.53 ± 0.05	1.62 ± 0.07
uniCGRU	43.9 ± 0.8	27.5 ± 2.0	20.3 ± 2.9	0.10 ± 0.03	1.30 ± 0.10	2.21 ± 0.07
MSEN	40.1 ± 0.3	27.4 ± 1.8	18.4 ± 2.4	0.13 ± 0.03	1.45 ± 0.08	1.82 ± 0.05

Table 6.2: Scanpath similarity scores of different models.

	VecSim	DirSim	LenSim	PosSim
biCLSTM+sDTW	**97.6** ± 0.4	**75.9** ± 0.2	**97.1** ± 0.3	**95.9** ± 0.6
biCLSTM	96.1 ± 0.8	69.3 ± 0.3	94.5 ± 0.3	88.8 ± 0.8
biCGRU+sDTW	92.7 ± 0.7	70.6 ± 0.8	87.9 ± 0.3	79.5 ± 0.2
biCGRU	95.1 ± 0.6	68.5 ± 0.7	93.2 ± 0.4	84.5 ± 0.5
uniCLSTM	96.1 ± 0.3	69.5 ± 0.2	94.8 ± 0.6	87.8 ± 0.7
uniCGRU	92.5 ± 0.2	66.1 ± 0.6	89.6 ± 0.4	83.0 ± 0.5
MSEN	93.0 ± 0.3	69.1 ± 0.5	92.9 ± 0.5	86.7 ± 0.4

Specifically, it reaches 97.7% for *Vector Similarity*, 75.9% for *Direction Similarity*, 97.1% for *Length Similarity*, and 95.9% for *Position Similarity*. However, sDTW loss did not significantly improve biCGRU models, as "biCGRU+sDTW" only exceeds "biCGRU" in *Direction Similarity*. "biCLSTM" achieved higher scores in all metrics compared to "biCGRU", and "uniCLSTM" outperforms "uniCGRU" in all metrics, indicating that ConvLSTM models has higher capability to model saccadic transitions. Finally, there is no significant difference in performance between "biCLSTM" and "uniCLSTM" except in *Position Similarity*, while "biCGRU" outperforms "uniCGRU" in all metrics. "biCLSTM+sDTW" and "biCLSTM" exceeds the performance of the baseline, "MSEN", in all scanpath similarities metrics.

Anatomy-specific Performances. In order to gain insight on each T-SEN variant's visual attention prediction performance on standard and non-standard sequence of a certain biometry, the static saliency scores and scanpath similarity scores are broken-down according to samples' sequence-level labels, as can be seen in Fig. 6.18 (static saliency scores) and Fig. 6.19 (scanpath similarity scores). In these two figures, "std ACP" refers to the sample sequences that contain standard ACP, while "bg Abdomen" refers to those that don't contain standard ACP; similar

nomenclature is used to Head and Femur. For each type of anatomy along the $x-$axis (in the order of standard ACP, background Abdomen, standard HCP, background Head, standard FLP, and background FLP), performance of 4 different models (blue for "biCLSTM+sDTW", orange for "biCLSTM", green for "biCGRU+sDTW", and red for "biCGRU") are plotted in box plots. Each box contains the quartile of the cross-validated performance scores of a particular anatomy, and "whiskers" extend to show the rest of the distribution; diamond-shaped points indicate outliers, determined using a function of the inter-quartile range in the python plotting library Seaborn. The horizontal line inside the box indicates the median value.

"biCLSTM+sDTW" (blue boxes) remains the best-performing T-SEN model in all anatomies across all metrics, with exceptions in the *IG* score for background Abdomen (Fig. 6.18(C)), the *NSS* score for standard HCP (Fig. 6.18(E)), and *Vector Similarity* score for standard HCP (Fig. 6.19(D)), where "biCLSTM" slightly outperforms.

It can also be observed that "biCLSTM+sDTW" generally performs better on standard sequences than on background sequences of each biometry. Specifically, it performs better on standard Abdomen sequences than on background Abdomen sequences on all static saliency scores with the exception of AUC; it also performs better on standard Head sequences and standard Femur sequences on 3 out of 6 static saliency scores comparing to their respective counter-parts. Standard Head sequences achieve better scores in all scanpath similarity scores, while standard Abdomen sequences achieve better scores in 3 out of all 4 scores. Table. 6.3 summarizes comparison results based on static saliency scores and scanpath similarity scores of "biCLSTM+sDTW". Similar trends can also be observed for other model variants.

Frame Classification Performance

Frame-level classification results of all T-SEN variants are presented in Table. 6.4 and results of base-line models are presented in Table. 6.5. Macro-averaged *Precision, Recall* and *F1-score* are reported for each variant and baseline model; performances per class are reported by the F1-score. T-SEN variants are named

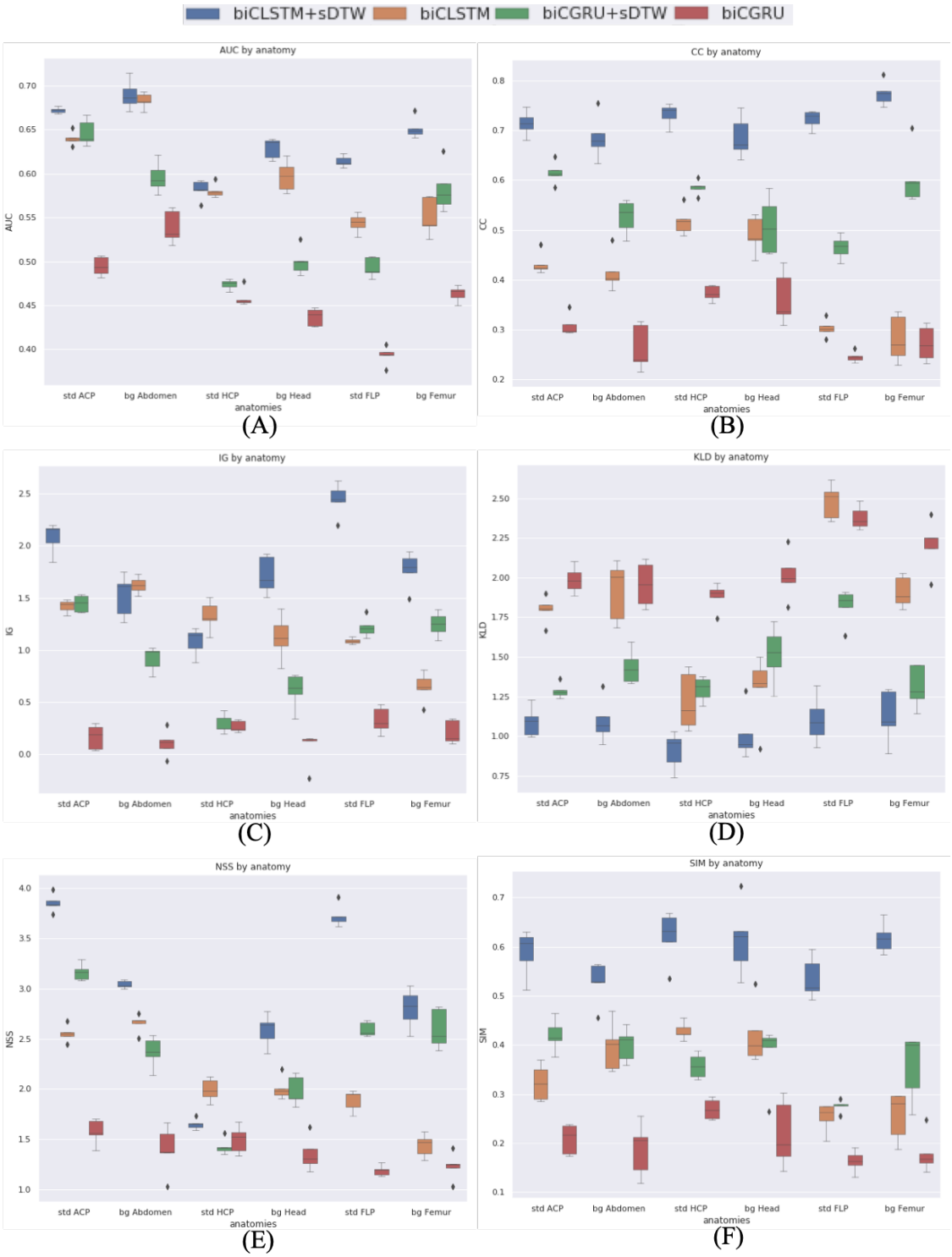


Figure 6.18: Boxplots demonstrating the mean static saliency scores by class labels on selected models.

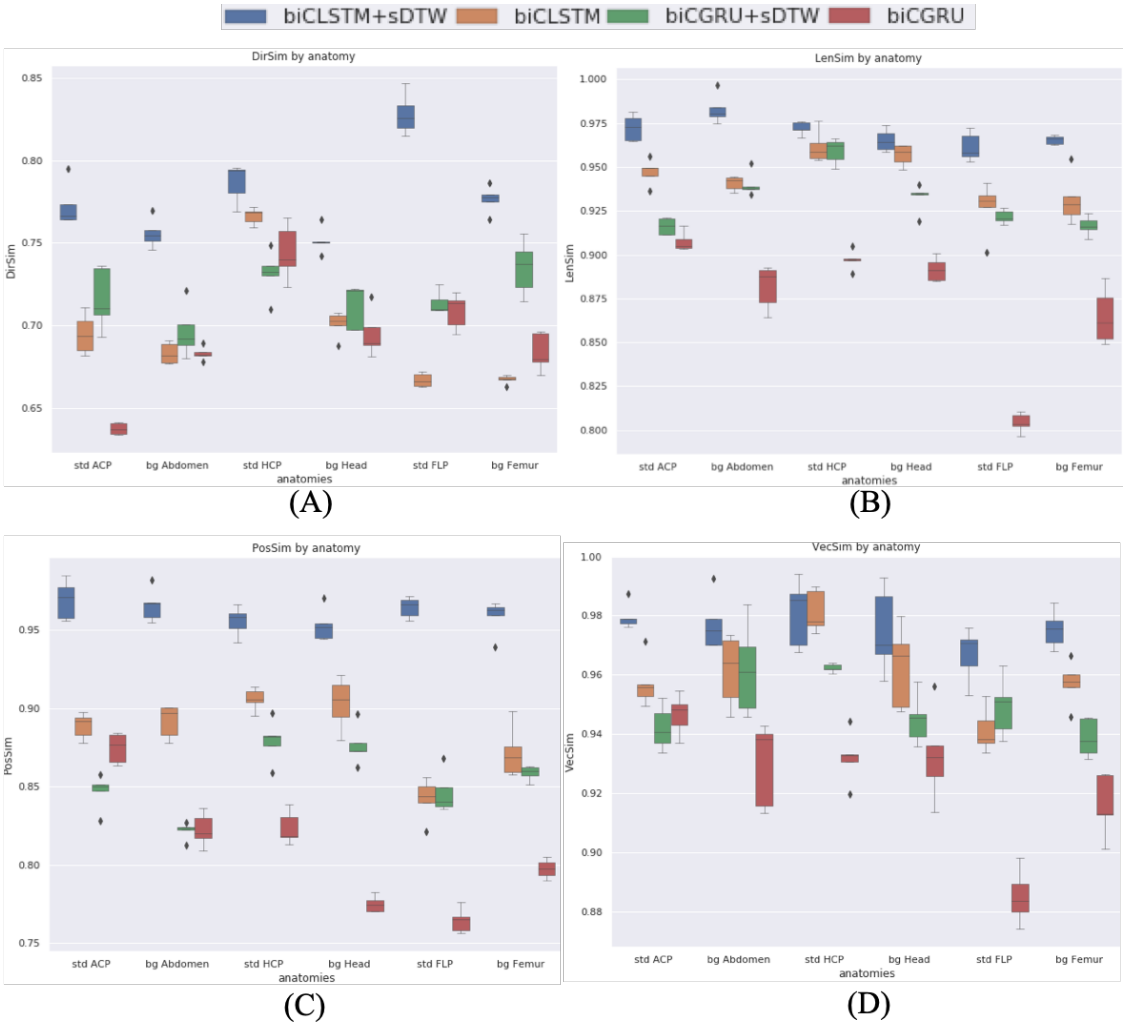


Figure 6.19: Boxplots demonstrating the mean static saliency scores by class labels on selected models.

by the specifications of the RNNs used in the VCM. Similar to the nomenclature used in the previous section, the model "biCLSTM" indicates that the model used bi-directional Convolutional LSTM in the VCM which was trained using cross-entropy loss as L_c, and "biCLSTM+FL" indicates that that the choice of classification loss L_c was Focal Loss, instead of cross-entropy. As mentioned before, all variants' TAM share the same RNNs specifications with VCM; the TAM of all variants reported in this section are trained with Kullback-Leibler Loss as L_s with additional temporal regularization loss L_t using sDTW.

It can be observed that Focal Loss is an effective loss function for improving frame classification performance: "biCLSTM+FL" model achieves the highest macro-

Table 6.3: Performance comparison on clips that contain standard biometry planes (std) *vs.* background (bg) clips based on the result of "biCLSTM+sDTW" model.

	std. better	bg. better
Static Saliency Metrics		
AUC	-	Abdomen, Head, Femur
CC	Abdomen, Head	Femur
IG	Abdomen, Femur	Head
KLD	Abdomen, Head, Femur	-
NSS	Abdomen, Femur	Head
SIM	Abodmen, Head	Femur
Scanpath Similarity Metrics		
DirSim	Abdomen, Head, Femur	-
LenSim	Head	Abdomen, Femur
PosSim	Abdomen, Head, Femur	-
VecSim	Abdomen, Head	Femur

averaged precision, recall and F1-scores in all variants compared. Its performance also exceeds those of other variants in terms of F1-scores in all classes except for background Head, on which "biCLSTM" achieves the highest score.

Convolutional LSTM is slightly more effective in encoding spatio-temporal information in input sequences for frame classification. "biCLSTM" performs better than "biCGRU" with F1-score of 82.4% compared to 80.4%; "uniCLSTM" achieved a F1-score of 82.2%, compared to 79.7% of "uniCGRU". On the other hand, using bi-directional RNNs does not improve frame classification performance.

All T-SEN models achieved higher scores in standard biometry planes comparing to baseline "MSEN", which achieved F1-scores of 68.3% for standard ACP, 68.1% for standard HCP and 60% for standard FLP. The best performing T-SEN model, "biCLSTM+FL", improved F1-scores on these standard biometry planes to 83.7%, 89.9% and 81.1%, respectively. These results are comparable to the values achieved by variants of SonoNets fine-tuned for this frame-classification task with F1-scores of 88.9%, 92.2% and 84.4% on corresponding biometry planes.

t-SNE feature visualization. A feature dimensionality reduction method t-Distributed Stochastic Neighbor Embedding (t-SNE) [**maaten2008visualizing**] was used to visualize the feature embedding generated at O^t in VCM and $\tilde{h}^t$ in the

TAM of selected variants of T-SEN models. Also, raw pixel values of the input video frames as well as the feature embedding of the last layer before adaptation layers in the classification branch and the last layer of the visual attention branch in MSEN are visualized for comparison. Compared to the t-SNE representation of the original signals (Fig. 6.20(A)), feature embedding of O^t in the "BiCLSTM+FL" model shows maximum separation between different classes. O^t of "BiCLSTM" and "BiCGRU" demonstrated lesser level of separation. However, overlaps still exist between the standard biometry planes and their corresponding background planes.

It is interesting to notice that $\tilde{h}^t$ in the TAM of "BiCLSTM+FL", though not trained for frame classification, demonstrated a good separation among different classes. Such observation can also be made for the $\tilde{h}^t$ of "BiCLSTM" and "BiCGRU", as well as "MSEN_att", indicating by learning to model human visual attention, the learnt feature embedding contain spatio-temporal information specific for different classes.

6.5.7 Discussions

It was demonstrated in the *Temporal Visual Attention modelling* section that T-SEN models successfully learns temporal visual attention and perform better than the MSEN model, both qualitatively and quantitatively, even though it was demonstrated to produce good quality visual attention maps in the previous chapter. The disparity is attributed to the difference in the nature of data used, and equally importantly, the tasks that sonographers were performing when gaze-tracking data were recorded. In the previous chapter where the Sweep dataset was used, sonographers had the freedom to view each single frame for as long as they wanted, allowing full inspection of the contents in that particular frame. Thus, gaze information recorded on one frame are less dependent on the other frames, making each frame independent from each other, which is different from the gaze-tracking data recorded in the PULSE dataset. The gaze-tracking data used here were recorded in real-time during anomaly scans, making gaze data on consecutive frames highly dependent on each other. Without RNNs to encode spatio-temporal information

Table 6.4: Classification results of different models.

	biCLSTM+FL	biCLSTM	biCGRU	uniCLSTM	uniCGRU
Precision	**89.4** ± 1.7	84.4 ± 7.2	81.8 ± 5.2	83.7 ± 7.6	79.0 ± 5.3
Recall	**85.1** ± 5.7	80.9 ± 6.0	79.2 ± 7.5	80.9 ± 6.1	80.9 ± 7.8
F1-score	**87.1** ± 3.4	82.4 ± 6.0	80.4 ± 9.3	82.2 ± 6.6	79.7 ± 4.6
F1-scores by class					
bg Ab	**90.6** ± 2.3	89.9 ± 2.4	87.1 ± 1.7	90.3 ± 1.9	85.5 ± 2.9
ACP	**83.7** ± 1.5	75.6 ± 2.5	74.4 ± 2.6	73.7 ± 3.0	72.3 ± 4.8
bg Head	90.7 ± 2.5	**91.3** ± 4.1	89.6 ± 3.1	89.4 ± 3.1	90.8 ± 2.7
HCP	**89.9** ± 1.1	79.5 ± 2.3	75.1 ± 2.7	79.3 ± 5.9	74.2 ± 2.2
bg Femur	**86.4** ± 4.0	83.3 ± 4.7	78.7 ± 2.1	82.4 ± 3.8	80.3 ± 3.7
FLP	**81.1** ± 2.3	74.4 ± 3.0	72.9 ± 2.6	73.1 ± 3.5	68.6 ± 3.3
Others	**87.1** ± 2.7	83.1 ± 4.8	84.9 ± 3.4	86.9 ± 4.2	86.3 ± 2.4

Table 6.5: F1-scores of different base-line models by anatomy.

	MSEN	SonoNet-64	SonoNet-32	SonoNet-16
bg Ab	**87.2** ± 3.5	69.5 ± 3.6	65.0 ± 4.1	65.9 ± 1.9
ACP	68.3 ± 2.6	**88.9** ± 3.1	88.3 ± 3.6	83.9 ± 1.8
bg Head	**92.1** ± 3.2	70.7 ± 3.7	58.9 ± 2.3	57.1 ± 3.0
HCP	68.1 ± 4.1	91.4 ± 1.8	**92.2** ± 4.0	89.6 ± 3.6
bg Femur	**76.7** ± 2.4	63.5 ± 5.9	64.1 ± 3.6	65.3 ± 2.1
FLP	60.0 ± 2.6	81.4 ± 2.1	**84.4** ± 2.0	80.2 ± 3.1
Others	87.1 ± 2.7	87.2 ± 2.9	85.9 ± 4.1	**88.8** ± 4.9

from consecutive frames with dependent gaze information, MSEN cannot model sonographers' behavior of sampling the visual field both spatially and temporally.

It's interesting to notice that T-SEN models generally perform better on standard sequences than on background sequences of each biometry, as demonstrated in Table. 6.3. Abdomen and Head are the two anatomies showing higher static saliency scores and scanpath similarity scores in standard sequences than background sequences. This can be attributed to the fact that Abdomen and Head show complicated structures, and sonographers follow the FASP guidelines when potential candidate standard planes appear. For example, sonographers consistently search for stomach bubble and umbilical veins, according to the study [**Ahmed2016**], with constant reference to the spine when determining the standard ACP. This consistency makes it easier to learn spatio-temporal transitions of visual attention.

As demonstrated in Table. 6.4 and Table. 6.5, the Dynamic Attention Maps

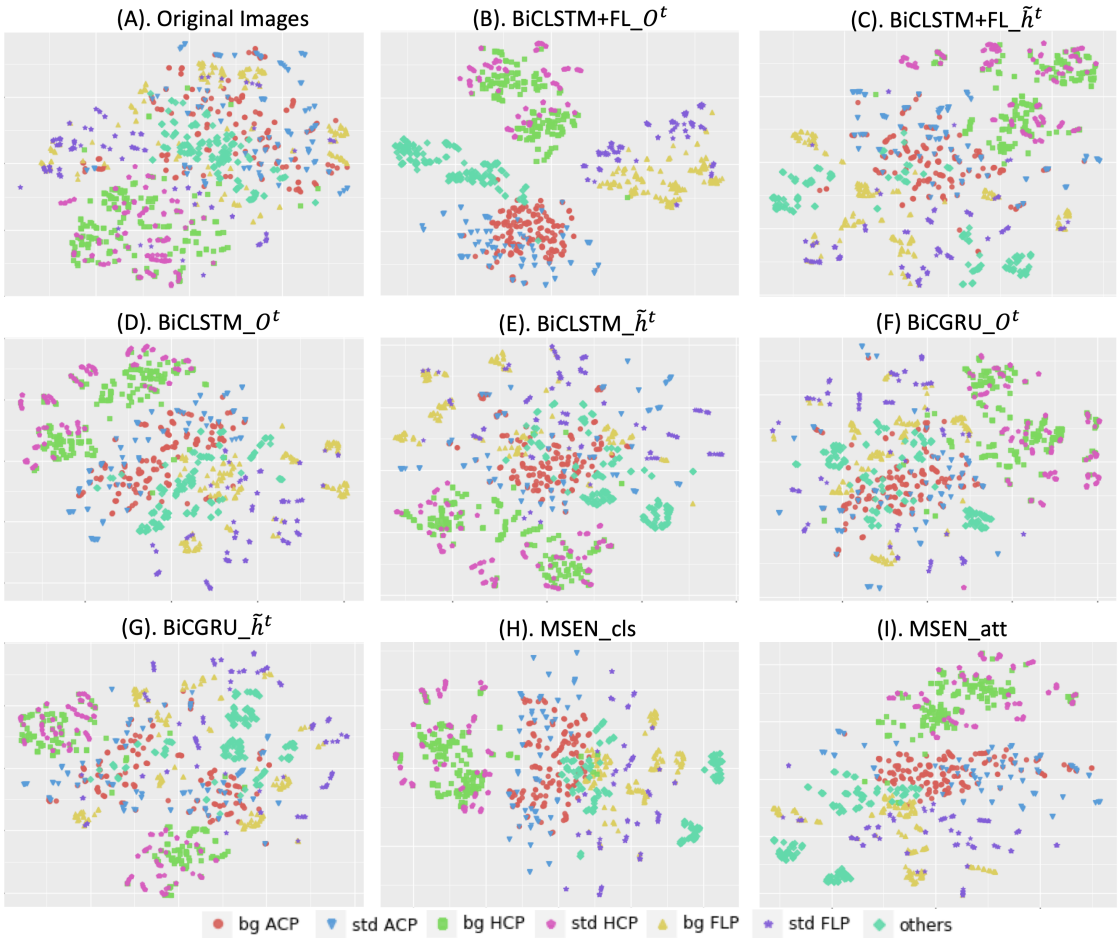


Figure 6.20: t-SNE visualization of the feature embedding of selected T-SEN variants as well as base line models SonoNets and MSEN.

predicted by TAM of T-SEN models assist frame classification tasks, supported by the fact that F1-scores of ACP, HCP and FLP all demonstrated significant improvement compared to those of MSEN. Even though the scores are lower than those achieved by heavier models of SonoNet-64 and SonoNet-32, they are higher than SonoNet-16, on which the T-SEN models were built with slightly more computational overhead. It remains interesting how models built on SonoNet-32 or even SonoNet-64 will perform both in terms of temporal visual attention modelling as well as frame classification tasks.

Even though the data set used in this work is significantly larger than the one used in the previous two chapters (89 patients vs 33 patients), the dataset used is still small and the result can still suffer from the limitations discussed in Section

4.4. However, as more data has been collected since the completion of this work, the model can be re-trained on larger dataset for comparison.

It was also pointed out that even though 5-fold cross-validation is used hyper-parameter tuning and model comparison, no independent test set was used for final comparison. Thus, the result is prone to "information bleeding". Essentially, multiple-hypothesis testing was performed and the significance level of the results could potentially be inflated. For future works, it is advisable to used a dataset for training, a validation set for hyper-parameter tuning, and a dataset for testing.

The two tasks trained in the network could potentially be useful in a clinical setting. The visual attention prediction branch of the network could be used for sonographers to retrospectively examine their decision-making procedure, and a model trained on experienced sonographers could potentially be developed into a tool to train new sonographers. The video classification branch, on the other hand, could potentially be developed into a computer-assisted diagnostic tool for standard plane detection so as to accelerate routine US examination.

Feature embedding visualization using t-SNE, as presented in Fig. 6.20, demonstrated that features learnt for visual attention prediction separates samples of different classes even though it did not receive any information regarding sample classes. As demonstrated by [**droste2019ultrasound**], feature representations learnt for visual attention modelling on 2-D US video frames are predictive for fetal anomaly standard plane detection. It will be of research interest to see if such spatio-temporal features learnt for visual attention modelling on US video sequences are more discriminative for frame prediction.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe

7 Conclusions and Future Works

Contents

7.1 Conclusions

The main contributions of this thesis are as follows:

1. **SonoEyeNet: Standardized fetal ultrasound plane detection informed by sonographer visual attention.** The first significant contribution of this thesis is a novel automated approach for detection of standardized abdominal circumference planes (ACP) in fetal ultrasound built in a convolutional neural network (CNN) framework, called SonoEyeNet. SonoEyeNet utilizes gaze-tracking data of a sonographer in automatic interpretation. Eye movement data was collected from experienced sonographers as they identified standard AC planes in fetal ultrasound video clips. A visual attention heatmap was generated from the gaze-tracking data for each video frame. A CNN model

was built using ultrasound frames and their corresponding visual attention heatmaps. Different methods of processing visual attention heatmaps and their fusion with image feature maps were investigated. We showed that with the assistance of human visual attention information, the precision, recall and F1-score of ACP detection was increased to 96.5%, 99.0% and 97.8% respectively, compared to 73.6%, 74.1% and 73.8% without using eye fixation information.

2. **Multi-task SonoEyeNet: Detection of fetal standardized planes assisted by generated sonographer attention maps.** The second significant contribution of this thesis is a novel multi-task convolutional neural network called Multi-task SonoEyeNet (MSEN) that learns to generate clinically relevant visual attention maps using sonographer gaze tracking data on input ultrasound (US) video frames to assist standardized abdominal circumference (AC) plane detection. The proposed architecture consisted of a generator and a discriminator, which were trained in an adversarial scheme. The generator learnt sonographer attention on a given US video frame to predict the frame label (standardized AC plane / background). The discriminator further fine-tuned the predicted attention map by encouraging it to mimick the ground-truth sonographer visual attention map. The novel model expanded the potential clinical usefulness of SonoEyeNet by eliminating the requirement of input gaze tracking data during inference without compromising its plane detection performance (Precision: 96.8%, Recall: 96.2%, F-1 score: 96.5%).

3. **Temporal SonoEyeNet: Spatial-temporal visual attention modelling for standard biometry planes detection.** The third significant contribution of this thesis is a novel network that learns sonographer spatio-temporal visual attention on input US video sequences to assist the detection of standard biometry planes of fetal abdomen, head and femur. The proposed architecture consists of two modules: a Temporal Attention Module (TAM) that predicts

Dynamic Attention Map, and a Video Classification Module (VCM) for video frame classification, both of which utilize bi-directional convolutional LSTM to encode spatial-temporal information from an input sequence. A soft Dynamic Time Warping (sDTW) loss was used as an excellent loss function to regularise temporal visual attention synchronisation between ground truth and predicted visual attention maps, and focal loss an effect loss function for frame classification. The best performing model out-performed MSEN in both static saliency scores and as well as scanpath similarity scores, and improves F1-scores for ACP, HCP and FLP to 83.7%, 89.9% and 81.1%, compared 68.3%, 68.1% and 60.0% from MSEN.

7.2 Future Work

There are a number of potential directions of future work that can be based on the contributions presented in this thesis. One direction is attention-assisted knowledge transfer to construct compact models that are computationally efficient with high performance. Another direction is learning feature representations from visual attention and to transfer them to other annotation-intensive tasks. The third direction is to combine gaze-tracking data with probe-tracking data collected in the PULSE project in order to progress towards automated anomaly scanning in a real clinical setting.

7.2.1 Attention-assisted knowledge transfer

Current deep learning models for medical analysis are recognized as having large memory footprints and inference costs. While there have been studies on over-parameterization of deep networks [**liu2015sparse**], efficient models have largely been defined empirically rather than using well-principled approaches. *SonoEyeNet* in Chapter 4 demonstrated that ACP classification performance is improved by combining sonographer visual attention maps with image features through element-wise production. Sonographer visual attention maps act as a strong prior that guides the network to efficiently focus its computation on key regions, rather than

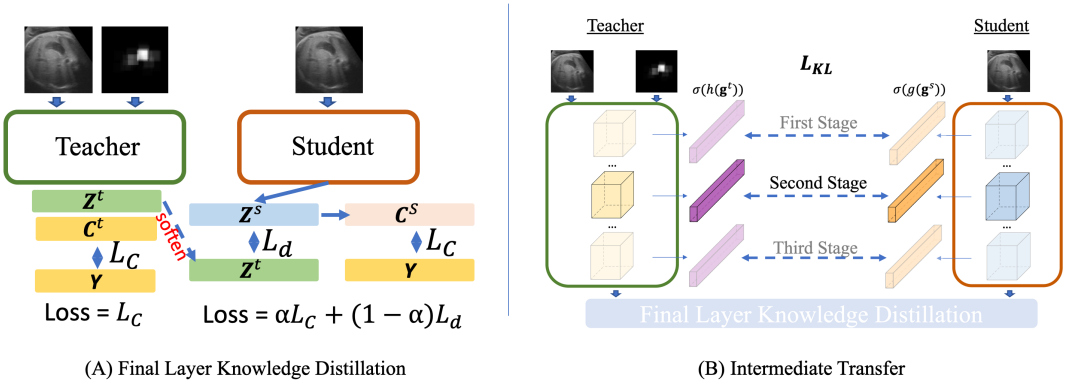


Figure 7.1: Schematic of our proposed knowledge-distillation pipeline. (A) Final Layer knowledge distillation (B) Intermediate Transfer.

the whole image. Thus it is reasonable to hypothesize that expert knowledge in the form of gaze data could filter out irrelavant information and fascilitate efficient knowledge transfer from heavy-weight high-performance models to create compact deep learning models.

Drawing inspirations from SonoEyeNet, we did a preliminary work [**cai2019efficient**] in this direction where we (1) trained compact models using both final layer (Fig. 7.1(a)) and intermediate knowledge distillation (Fig. 7.1(b)) from large models for the exemplar task of anatomy classification of fetal abdomen, head, and femur frames from a free-hand fetal ultrasound sequence; and (2) incorporate expert human knowledge in the form of gaze tracking data into a teacher model to enhance knowledge transfer. It was found that compact models trained using gaze-assisted knowledge distillation in the form of element-wise production achieve higher accuracy than the same models trained with image-only knowledge distillation. The best performing model gained an additional 16% accuracy. Compared to the teacher model, the compact student model gained 5× inference speed up and 1000× memory reduction while reaching a classification performance comparable to that of the teacher (accuracy for Abdomen, head and femur are 0.87, 0.85, 0.84 for the best performing student and 0.92, 0.90, 0.87 for the best performing teacher).

Future work in this direction can explore novel ways of incorporating gaze data into knowledge distillation pipelines for more complicated networks trained for tasks such as anatomical landmark detection or segmentation. With the

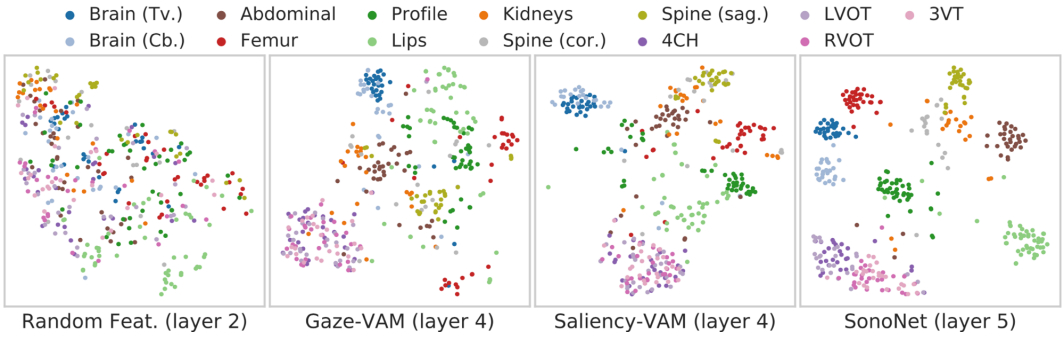


Figure 7.2: t-SNE visualization of the feature embedings at respective layers with the highest F1-scores (Background class omitted for legibility).

emergence of edge devices such as the Butterfly US probe, compact models created through gaze-assisted distillation that are deployable onto mobile devices will be increasingly needed.

7.2.2 Task Transfer

As discovered in Chapter 6, features learnt for visual attention prediction separates samples of different classes on t-SNE visualization (Fig. 6.20) even though they were not trained to distinguish frames of different classes. It is confirmed by the study that humans direct their attention towards semantically informative regions when interpreting a visual scene [**wu2014guidance**], indicating that feature representations learnt through modelling human visual attention might have already encoded semantic information, and can be used for other tasks.

In our recent work [**droste2019ultrasound**], it is demonstrated that transferable representations of US images can be learned without manual annotations by modelling human visual attention. Two CNN-based sonographer visual attention models (Gaze-VAM and Saliency-VAM) were trained on PULSE data, and a simple softmax regression was trained on the feature activations of each CNN layer in order to evaluate the representations independently of transfer learning hyper-parameters. We find that the attention models derive strong representations, approaching the precision of SonoNet-64 for all but the last layer. The feature representations of the best layer in each model plotting using t-SNE is presented in Fig. 7.2.

The visual attention models in [**droste2019ultrasound**] were trained on static visual attention without consideration of temporal information. As demonstrated in Chapter 6, on the other hand, temporal information is important to model sonographer visual attention. Thus it is reasonable to hypothesize that spatial-temporal feature representations learnt in temporal attention models could be semantially more meaningful than those trained on static visual attention. Future works in this direction can study how to transfer spatial-temporal attention features to classification tasks, and even extend to other tasks closely related to attention, *e.g.* localisation.

7.2.3 Learning Hand-Eye Coordination

The dataset collected under PULSE project is unique in that gaze-tracking data as well as US probe-tracking data are collected, providing a good opportunity to understand the decision making process of sonographers during US examinations: how did a sonographer decide in which direction they should move the probe, and based on the video monitor how should the sonogrpaher plan their next probe movement accordingly.

This process can potentially be fit nicely into a reinforcement learning framework [**sutton1998introduction**]. Reinforcement learning can be defined as a computational approach for learning by interacting with an environment to maximize cumulative reward signals. An agent interacts with an environment E at every state s through an action a from a set of allowed actions A. Each valid action results in a positive reward, and invalid action results in no or negative reward. The aim of the agent is to learn an optimal *policy* that can maximize its cumulative award. In this problem setting, the probe can be viewed as the agent, which is allowed to interact with the 3-D environment inside a pregnant woman's womb. The reward can be defined according to the specific task that needs to be achieved, *e.g.* navigate through the environment and find standard biometry planes. Preliminary work of framing medical imaging problems in the form of Reinforcement Learning problems have been attempted [**alansary2018automatic**, **alansary2019evaluating**]. These

work use features extracted only from images to describe the state s for their agent. However, gaze data in the PULSE dataset adds an additional dimension to describe the s, and this is a direction that future works could explore.

Another possible direction worth exploring is imitation learning [**ho2016generative**], a way of transferring human knowledge to the agent, but one of the biggest challenges for imitation learning is collecting expert demonstration dataset. The PULSE dataset, on the other hand, has already collected probe tracking data. Future works could explore how to transfer expert knowledge in the form of probe-tracking data to the agent.