Advancing Ultrasound Image Analysis by Capturing Operator Gaze Patterns

Richard Droste

Balliol College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2021

Abstract

Obstetric ultrasound scanning is a safe and effective tool for the early detection of fetal abnormalities and therefore crucial for determining the necessity of clinical intervention. However, ultrasound relies on operator expertise, which is a scarce resource globally. Moreover, there are large geographical and inter-observer variations of clinical outcomes. To address this, the PULSE project aims to develop a new generation of ultrasound scanning capabilities based on big data and machine learning models which capture the knowledge of experienced sonographers. To this end, the project team acquired a first-of-its-kind large-scale dataset of routine clinical ultrasound scanning with gaze-tracking data.

In this thesis, we first examine shortcomings of the operator-machine interaction. We find that sonographers adjust the biometric measurements of fetuses with potential growth abnormalities towards the healthy expected value, providing a possible explanation for the known deficiencies of these measurements. Moreover, we study the adherence to safety recommendations regarding thermal energy emission and find that, while sonographers keep within the appropriate limits, they rarely check the safety indices. We provide suggestions for the modification of the ultrasound machine interface to address these two issues.

Second, we develop the first model that predicts sonographer gaze-tracking data on ultrasound video through the method of *visual saliency prediction*. In addition, we propose the first unified visual saliency model for the prediction of gaze on both images and videos. Besides unifying the two modalities, the model obtains state-of-the-art performance on both tasks for all relevant computer vision benchmarks.

Third, we show that sonographer gaze-tracking data is a powerful supervision signal for *ultrasound image feature representation learning*. We develop a general framework for representation learning and transfer of the trained neural network to the downstream tasks of *standard plane detection* and *automatic biometry plane annotation*. We also show that the learned representations, in combination with the sonographer gaze prediction, can be used to discover and localize visually salient anatomical landmarks, *i.e.*, landmarks that sonographers use for visual navigation.

Finally, we provide an overarching discussion and an extended outlook chapter which describes a system for guiding sonographers during *standard plane acquisition*.

Advancing Ultrasound Image Analysis by Capturing Operator Gaze Patterns

Richard Droste

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Acknowledgements

Institutional

This work is supported by the ERC (project PULSE, reference ERC-ADG-2015 694581).

Personal

First and foremost, I would like to thank my supervisor Prof. Alison Noble for her constant support, encouragement and guidance. She provided me with both the fine-grained feedback and the birds-eye overview which shaped this thesis, all while giving me the freedom to explore my ideas. I owe it to her that I can present this thesis that I am truly proud of.

Next, I would like to thank my close collaborators Lior Drukker, Jianbo Jiao, Yifan Cai, Harshita Sharma, Pierre Chatelain and Aris T. Papageorghiou for the many fruitful discussions and their contributions to the papers presented here.

Special thanks go to Pierre and Lior for the data acquisition: Pierre expertly built the data acquisition setup without which the studies of this thesis would have been impossible. Lior persistently steered the data acquisition through the many inevitable obstacles and is responsible for the scale it has today.

Further, particular gratitude goes towards Lior and Jianbo with whom I had the honor of jointly writing papers with. Many thanks to Lior for his eagerness to write interdisciplinary papers and his expertise in making them great. Thanks to Jianbo for jointly writing the ECCV 2020 paper and overcoming all prior setbacks.

To all the members of the Noble lab and the IBME, thank you for your support and discussions and for creating the friendly atmosphere in which collaborations and new ideas can flourish.

On a personal note, I would like to thank my partner Monica. Thank you for bearing with me during the past three years, through easy times and difficult ones. Thank you for making me feel proud of every achievement, for your advice, and for your knowledge on statistical correlation coefficients, without which my MICCAI 2020 paper would not have been possible. Last but not least, I would like to thank my family. I owe this thesis to your endless support and encouragement in life.

Abstract

Obstetric ultrasound scanning is a safe and effective tool for the early detection of fetal abnormalities and therefore crucial for determining the necessity of clinical intervention. However, ultrasound relies on operator expertise, which is a scarce resource globally. Moreover, there are large geographical and inter-observer variations of clinical outcomes. To address this, the PULSE project aims to develop a new generation of ultrasound scanning capabilities based on big data and machine learning models which capture the knowledge of experienced sonographers. To this end, the project team acquired a first-of-its-kind large-scale dataset of routine clinical ultrasound scanning with gaze-tracking data.

In this thesis, we first examine shortcomings of the operator-machine interaction. We find that sonographers adjust the biometric measurements of fetuses with potential growth abnormalities towards the healthy expected value, providing a possible explanation for the known deficiencies of these measurements. Moreover, we study the adherence to safety recommendations regarding thermal energy emission and find that, while sonographers keep within the appropriate limits, they rarely check the safety indices. We provide suggestions for the modification of the ultrasound machine interface to address these two issues.

Second, we develop the first model that predicts sonographer gaze-tracking data on ultrasound video through the method of *visual saliency prediction*. In addition, we propose the first unified visual saliency model for the prediction of gaze on both images and videos. Besides unifying the two modalities, the model obtains state-of-the-art performance on both tasks for all relevant computer vision benchmarks.

Third, we show that sonographer gaze-tracking data is a powerful supervision signal for *ultrasound image feature representation learning*. We develop a general framework for representation learning and transfer of the trained neural network to the downstream tasks of *standard plane detection* and *automatic biometry plane annotation*. We also show that the learned representations, in combination with the sonographer gaze prediction, can be used to discover and localize visually salient anatomical landmarks, *i.e.*, landmarks that sonographers use for visual navigation.

Finally, we provide an overarching discussion and an extended outlook chapter which describes a system for guiding sonographers during *standard plane acquisition*.

Contents

List of Figures

List of Tables

List of Abbreviations

3VT . . . . . Three vessel and trachea view

4CH . . . . . Four chamber view

AC . . . . . Abdominal circumference

ACP . . . . . Abdominal circumference plane

AUC . . . . . Area under the ROC curve

AHRS . . . . . Attitude and heading reference system

BN . . . . . Batch-normalization

CC . . . . . Cross-correlation

CRL . . . . . Crown-rump length

CSP . . . . . Cavum septum pellucidi

FASP . . . . . Fetal Anomaly Screening Programme

EDD . . . . . Expected date of delivery

EFW . . . . . Estimated fetal weight

FCN . . . . . Fully convolutional network

FGR . . . . . Fetal growth restriction

FL . . . . . Femur length

FLP . . . . . Femur length plane

GN . . . . . Group Normalization

GRU . . . . . Gated-recurrent-unit

HC . . . . . Head circumference

HCP . . . . . Head circumference plane

IMU . . . . . Inertial measurement unit

IoU . . . . . Intersection over union

KLD . . . . . Kullback-Leibler Divergence

LV . . . . . Lateral ventricle

LVOT . . . . . Left ventricle outflow tract

LSTM . . . . . Long short-term memory

MI . . . . . Mechanical Index

NGA . . . . . Normal-for-gestational-age

NHS . . . . . National Health Service

NSS . . . . . Normalized scanpath saliency

ODS . . . . . Output Display Standard

OCR . . . . . Optical character recognition

PULSE . . . . Perception Ultrasound by Learning Sonographic Experience

RCN . . . . . Recurrent convolutional network

RVOT . . . . . Right ventricle outflow tract

SB . . . . . Stomach bubble

SGA . . . . . Small-for-gestational-age

SGD . . . . . Stochastic gradient descent

SIM . . . . . Similarity index

SOB . . . . . Suboccipitobregmatic plane

TC . . . . . Transcerebellar

TCD . . . . . Transcerebellar diameter

TCP . . . . . Transcerebellar plane

TI . . . . . Thermal Index

TIb . . . . . Thermal Index for bone

TIs . . . . . Thermal Index for soft tissue

TFF . . . . . Time-to-first-fixation

TV . . . . . Transventricular

TVP . . . . . Transventricular Plane

UV . . . . . Umbilical vein

US . . . . . Ultrasound

VA . . . . . Ventricular atrium

VAM . . . . . Visual attention models

WS . . . . . Weight Standardization

1 Introduction

1.1 Thesis Outline and Contributions

The human visual system has the remarkable ability to automatically direct our gaze towards the most important information [KU85]. For the acquisition and interpretation of obstetric ultrasound images, this ability is of particular importance: the operator manipulates the position of a transducer (probe) in order to find desired anatomical views of the mother and fetus via the B-mode ultrasound signal that is displayed on the machine's monitor. This requires skilled hand-eye coordination since only one cross-section of the structures of interest is visible at any time and since the visual signal (monitor) is removed from the observed object (mother and fetus) [AHM13]. In addition, ultrasound images are usually difficult to read since they are acoustic maps that can contain considerable levels of noise and artifacts [KT86]. Finally, the operators need to perform accurate biometric measurements on certain standard views in order to determine related abnormalities [NHS18].

To develop a better understanding of sonographers' interaction with the ultrasound scanner and the ultrasound signal, the *Perception Ultrasound by Learning Sonographic Experience* (PULSE) project [PUL] was recently established. Its cornerstone is the acquisition of a large-scale dataset of routine clinical obstetric ultrasound examinations, including full-length video data, gaze-tracking of several

operators and motion-tracking of their probe movements. In chapter 3 we show that the video and eye-tracking data provide interesting insights into the operator's interaction with the ultrasound machine. In Sec. 3.1 we discover that biometric measurements performed by the operators are biased towards the known expected measurement outcome. This leads to the potential misclassification of fetuses with abnormal weight. In Sec. 3.2 we perform the first comprehensive analysis of the adherence to safety guidelines. We find that the safety indices, which are displayed on the ultrasound machine's graphical interface, are gazed at in only a small fraction of scans.

Besides addressing the evident need for a better understanding of operator-machine interactions, a closely related aspect of the PULSE project is to capture the experience that sonographers acquire through years of experience. Ultrasound is a relatively low cost, portable and safe imaging modality, yet many women in developing countries do not receive a single ultrasound examination throughout their pregnancy due to a lack of skilled operators [SBA+15]. Developing assistive systems that support image acquisition, interpretation and quality assessment holds great promise to solve this issue. While prior work has proposed numerous solutions based on established computer vision algorithms, we take a new approach and design algorithms that learn from the interaction of experienced operators with the ultrasound machine.

As mentioned before, experience reflects in the way that operators direct their gaze towards task-relevant visual information. In this thesis, we capture the operator gaze patterns by training machine learning models to predict the operator gaze on unseen images, which is commonly referred to as *visual saliency prediction*. Existing visual saliency prediction methods focus on either image or video data, but ultrasound data consists of both live video and static frozen images, which we address in chapter 4. In Sec. 4.1 we find that sonographer gaze on ultrasound video can be captured most accurately with a cognition-inspired saliency model that incorporates the observers' implicit predictions of future video frames. Moreover, since ultrasound data consists of both video and image data, we develop the first

unified saliency model in Sec. 4.2. We find that it improves saliency prediction on all relevant computer vision benchmarks for both modalities and published an open-source implementation that has gained popularity in the community.

In chapter 5 we employ learned saliency models to design assistive algorithms for three main sonographer tasks. First, while operators acquire standardized views of the fetus, they can be supported through *standard plane detection*. In Sec. 5.1 we explore how expert visual saliency models can support this task. In parallel, operators need to identify the relevant anatomical landmarks and in Sec. 5.2 we present a method to automatically discover the landmarks that are salient to the expert operators, and to localize them automatically. Finally, several downstream tasks such as biometric measurements or quality assurance depend on the segmentation of relevant structures and the correct placement of measurement calipers. In Sec. 5.3 we demonstrate that visual saliency models can be adapted to effectively perform these downstream tasks with few manual annotations.

In addition, a review of the relevant literature is provided in chapter 2 and an overarching discussion and conclusion in chapter 6. Finally, chapter 7 is dedicated to the research outlook, including our work for providing automatic ultrasound probe movement guidance for standard plane acquisition.

1.2 The NHS Fetal Anomaly Screening Programme

Obstetric ultrasound scanning in England is governed by the Fetal Anomaly Screening Programme (FASP) of the National Health Service (NHS) [NHS18]. It includes two routine scans: the "*early pregnancy scan*", performed in the first trimester of pregnancy between 11^{+2} weeks to 14^{+1} weeks of gestation (weeks$^{+\text{days}}$), and the "*anomaly ultrasound scan*"performed in the second trimester between 18^{+0} weeks to 20^{+6} weeks of gestation. In addition a third trimester "*growth scan*" is usually performed at ca. 36 weeks of pregnancy.

The purpose of the early pregnancy scan is to confirm viability of the pregnancy, ascertain single or multiple pregnancy, estimate the gestational age, and to detect major structural abnormalities [NHS18]. Here, the gestational age is determined

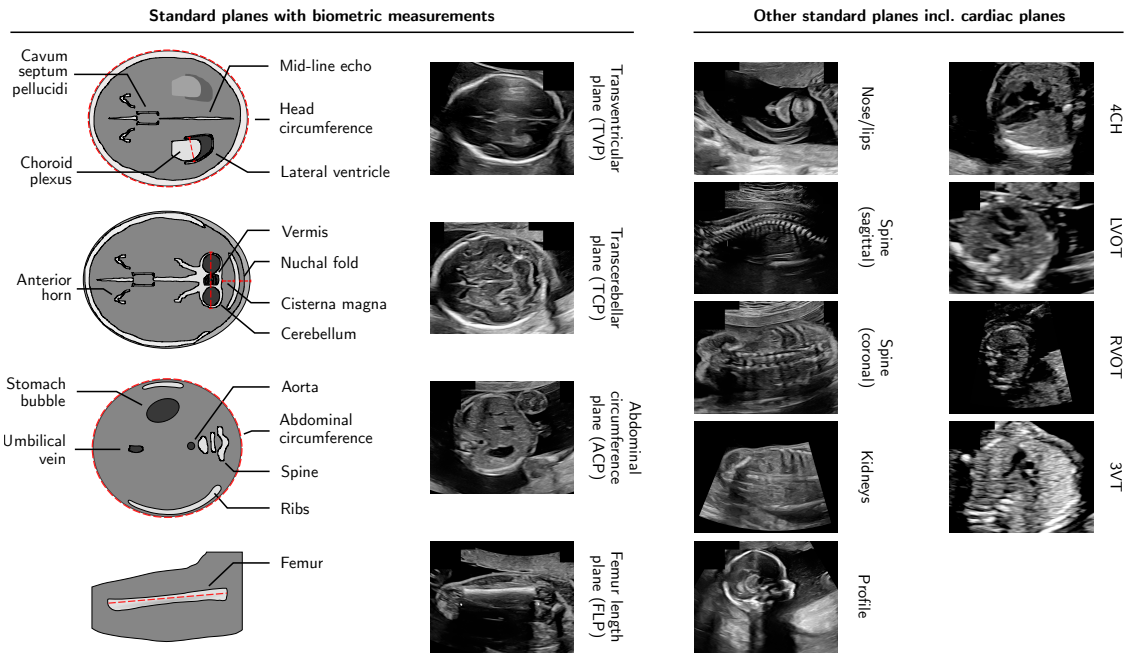

Figure 1.1: FASP standard planes [NHS18]. Cardiac view abbreviations: 4CH: four chamber view; LVOT: left ventricular outflow tract; RVOT: right ventricular outflow tract; 3VT: three vessel and trachea view.

based on the crow-rump length (CRL). Moreover, conditional on the mother's consent, the likelihood of trisomy 21 and/or 18/13 can be determined based on the nuchal translucency, a measurement at the neck of the fetus.

The anomaly ultrasound scan, hereafter referred to simply as *anomaly scan* or *second trimester scan*, has the main purpose of detecting anatomical or growth-related abnormalities of the fetus. Specifically, the scan aims to identify abnormalities related to conditions that: a) indicate that the baby may die shortly after birth, b) may benefit from treatment before birth, c) require planned delivery in an appropriate hospital/centre, and/or d) require optimised treatment after the baby is born.

To detect abnormalities, at least six standard anatomical views are acquired during the anomaly scan:

Transventricular plane (TVP) For the measurement of the *ventricular atrium* (VA) and the *head circumference* (HC) and therefore also referred to as

head circumference plane (HCP). The view is determined primarily by the appearance of the cavum septi pellucidi (CSP) and *lateral ventricle* (LV).

Transcerebellar plane (TCP) For the measurement of the *trans-cerebellar diameter* (TCD), the *cisterna magna* and the *nuchal fold*. The anatomical term for the view is the *suboccipitobregmatic* (SOB) plane.

Nose/Lips Coronal view of the face to detect cleft lip.

Spine *Sagittal* or *coronal* view of the spine including skin covering.

Abdominal circumference plane (ACP) View of the fetal abdomen that includes the *umbilical vein* (UV), *stomach bubble* (SB), aorta, spine and the upper and lower ribs. It is used to measure the *abdominal circumference* (AC).

Femur length plane (FLP) View of the femur to measure the *femur length* (FL).

Besides these six required planes, a coronal or axial view of the fetal kidneys is usually acquired. If the renal pelvis appears large, a measurement of the anterior-posterior renal pelvis diameter is included. The HC, AC and FL are measured and combined with the expected date of delivery (EDD) to assess growth velocity [NHS18].

In addition to the basic anomaly scan, a *fetal cardiac protocol* of seven views is defined [NHS18]. In practice, sonographers often routinely acquire four of these cardiac views: *four chamber view* (4CH), *right ventricular outflow tract* (RVOT), *left ventricular outflow tract* (LVOT), and *three vessel and trachea view* (3VT). Finally, a "*profile*" view of the face is typically acquired. This results in a total of thirteen standard anatomical views (*standard planes*) which are illustrated in Fig. 1.1. During the growth scan, the HC, AC and FL are measured once again in order to determine if the fetus has continued to grow adequately.

Regarding the examination of maternal anatomies, the guidelines recommend an examination of placental position and amniotic fluid as good clinical practice [NHS18].

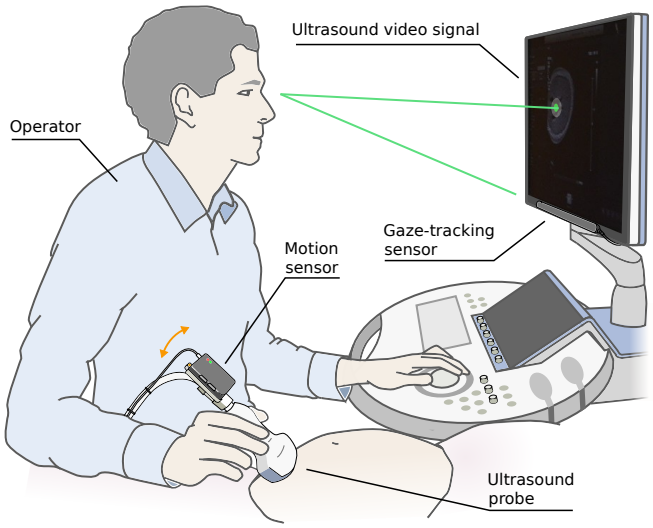

Figure 1.2: PULSE data acquisition setup.

1.3 The PULSE Project

While guidelines for performing obstetric ultrasound scanning exist, correctly and effectively performing a scan requires years of experience. This leads to large inter- and intra-observer [CFL+09] and geographical [GSC01] variations of clinical outcomes. Moreover, many women in developing countries do not receive a single ultrasound examination throughout their pregnancy due to a lack of skilled operators [SBA+15]. The PULSE project aims to address these issues by studying the scanning process and the interaction of sonographers with the ultrasound machine comprehensively. The goals are twofold: First, to develop a better understanding of the expertise that sonographers acquire throughout years of scanning through big data analysis [DDC+20b, DDC+20a, DSD+20, DDNP20, SDD+20, SDC+21] and capture the expertise with machine learning models [SDC+19, DCS+20, WDJ+20, SDC+21]. Second, to employ the machine learning models to develop automated assistive algorithms that enable less experienced operators to effectively perform ultrasound scans [CSCN18a, CSCN18b, DCS+19, ASD+19, AES+20, JCA+20, DDPN20, JDD+20, DCD+20].

In this thesis we present two big data analyses related to biometric measurements and awareness of safety indices (chapter 3, [DDC+20b, DDC+20a]), machine learning

models that capture human visual attention via visual saliency modeling (chapter 4, [DJN20, DCS+20]), and ultrasound image analysis methods that build upon the visual saliency models (chapter 5, [DCS+19, DCD+20]).

1.4 The PULSE Dataset

To enable the comprehensive examination of ultrasound scanning and operator-machine interaction, the project includes the acquisition of novel dataset of unprecedented scale and multi-modality. The data acquisition setup is illustrated in Fig. 1.2. Full-length ultrasound videos of obstetric scans are acquired in a routine clinical setting. The video signal of the ultrasound machine is cloned via a video-splitter module and recorded lossless at 30 Hz via a dedicated acquisition computer. In addition, the gaze points of the operators are recorded at 90 Hz with an eye-tracker (Tobii Eye Tracker 4C, Danderyd, Sweden). The long-term accuracy of the gaze-tracking setup was evaluated in a separate study [CSD+18]. Additionally, the movement of the ultrasound probe is tracked with an *inertial measurement unit* (IMU) and for a subset of scans the sonographer voice is recorded. In this thesis we focus on the video and gaze-tracking data, except for chapter 7 where we provide an outlook for motion-tracking based assistive models.

The ultrasound scans are recorded from May 2018 to March 2020 by sonographers and fetal medicine physicians at the maternity ultrasound unit, Oxford University Hospitals NHS Foundation Trust, Oxfordshire, United Kingdom. All women are offered a 12 weeks dating scan, a 20-week anomaly scan, and a 36-weeks growth scan. Based on risk factors or clinical indication, women were offered additional scans at other gestational ages [Roy13]. In this thesis, these additional scans are excluded when analyses are conducted based on anomaly or growth scans. Ultrasound examinations are carried out or supervised by accredited sonographers or fetal medicine physicians using standard ultrasound equipment. For quality control measures, the stored images and the reliability of measurements are regularly assessed using the INTERGROWTH-21st quality criteria [SIO+13]. All ultrasound scans are performed using a commercial Voluson E8 version BT18 (General Electric

Healthcare, Zipf, Austria) ultrasound machine equipped with standard curvilinear (C2-9-D, C1-5-D), and 3D/4D (RAB6-D) probes. This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051), and written informed consent was given by all participating pregnant women. Sonographers also consented to participate in the study at the outset, but do not have any visual or other signal to know that tracking devices are functioning. This dataset is the foundation for our contributions presented in chapters 3, 5 and 7. Section 4.1 builds upon previously acquired data [CSCN18a] and section 4.2 uses public benchmark datasets. The details of the subsets of data used, data processing and annotation are described in the respective sections.

1.5 Gaze-Tracking for Ultrasound Imaging

Human eye movement during information processing has been extensively studied in experimental psychology and cognitive science [Ray78, Ray98]. Research in scene perception, which viewing ultrasound images and videos can be seen as, has also existed for a long time [Bus35] and experienced steady academic progress [Hen03]. Henderson [Hen03] lists three reasons why gaze is key to understanding scene perception. First, eye movements are actively controlled to search for task-relevant information. Therefore, fixations are directed towards important and informative scene regions. Second, eye movements reveal the allocation of *overt attention* (hereafter simply referred to as attention), which is indicative of *covert attention*, i.e. the allocation of cognitive resources. Third, eye movements provide an "unobtrusive, sensitive, real-time behavioral index of ongoing visual and cognitive processing", enabled by modern eye-tracking technology. These points are complemented by Land and Hayhoe [LH01] who find that gaze precedes action by a fraction of a second, indicating planning and intent instead of merely a response to stimuli.

In contrast to other medical imaging technologies such as MRI and CT, the images acquired by an ultrasound system are displayed in real time and the operator interactively manipulates the probe position and machine settings to acquire the desired images. During this process the operator needs to discern the anatomies

of interest and monitor their appearance. Therefore, tracking the gaze of the operator reveals all three pieces of information that Henderson [Hen03] describes: the visual information that is relevant to fulfilling the task of obstetric ultrasound scanning, the regions of the ultrasound images that demand particular attention from the operator, and the operators cognition when deciding whether to save the current view or performing measurements.

1.6 Visual Saliency Prediction for Ultrasound Imaging

Since chapter 4 and chapter 5 are founded on the concept of visual saliency prediction, a short general introduction to the method is provided here. The related literature is reviewed more comprehensively in Sec. 2.2.

Following the insights presented in the previous section, the question arises: is it possible to *predict* which points in an image or video are likely to be fixated? This would allow to study visual attention on unseen images and videos for which no gaze-tracking data is available. The answer to the question is: yes, the allocation of visual attention can be predicted. Seminal work was performed by Koch and Ullman [KU85] who hypothesized that gaze is attracted by certain low-level features such as color, orientation, direction of movement, etc., and developed an algorithm that fuses these low-level features into a *visual saliency map*. Afterwards, many iterations of this low-level saliency approach were proposed [IKN98, BI13, SF03, HKP07, LMLCBT06, JEDT09].

These early models were then surpassed in performance with a significant and growing margin by deep neural networks [LBH15] trained on large gaze-tracking datasets [JHDZ15a, MSHH11, MS15, WSG+18] due to their ability to learn features in a purely data-driven way, both low-level and high-level. The importance of high-level features is explained by the fact that image components such as people, faces, text and objects [KWGB17, BSI13b] are the most salient.

Beyond the intrinsic value of predicting human gaze in order to understand the features that attract it, there is some early work which showed that the

feature representations are useful for other computer vision applications. In particular, Cornia et al. [CBSC17b] used a trained saliency predictor to support image captioning. In chapter 5 we extend this idea into a full framework for saliency-based image representation learning. The mathematics and implementation details of visual saliency prediction are presented in subsection 5.1.2.

1.7 A Note on the Integrated Thesis Format

This thesis is presented in an *integrated format*, i.e., paper-based format. This means that chapters 3, 4 and 5, which present the research contributions, are based on papers that were peer-reviewed and published in venues with competitive acceptance. The papers are presented as "author accepted manuscript" versions and reproduced in the thesis formatting. Moreover, the acknowledgments are compiled at the beginning of the thesis and the reference markers are standardized, with a biography provided for each paper in addition to a collective full biography at the end. The list of these papers is presented in Sec. 1.8. For each paper, I will provide a brief background section to provide the broader context. Moreover, an overarching literature review (chapter 2), discussion (chapter 6) and outlook (chapter 7) are provided besides this common introduction chapter.

1.8 Publications

- L. Drukker[1], R. Droste[1], P. Chatelain, J. A. Noble, and A. T. Papageorghiou, "Safety Indices of Ultrasound: Adherence to Recommendations and Awareness During Routine Obstetric Ultrasound Scanning," *European Journal of Ultrasound*, vol. 41, no. 2, pp. 138–145, 2020. Sec. 3.1
  Editor's choice.

- L. Drukker[1], R. Droste[1], P. Chatelain, J. A. Noble, and A. T. Papageorghiou, "Expected-value bias in routine third-trimester growth scans," *Ultrasound in Obstetrics & Gynecology*, vol. 55, no. 3, pp. 375–382, 2020. Sec. 3.2

- R. Droste, Y. Cai, H. Sharma, P. Chatelain, A. T. Papageorghiou, and J. A. Noble, "Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction," In: *Medical Image Understanding and*

Analysis (MIUA), pp. 174–186, 2019. Sec. 4.1
Oral presentation, Best Paper Award.

- R. Droste[1], J. Jiao[1], J. A. Noble, "Unified Image and Video Saliency Prediction," In: *European Conference on Computer Vision (ECCV)*, 2020. Sec. 4.2
Spotlight Presentation (top 5% of papers).

- R. Droste, Y. Cai, H. Sharma, P. Chatelain, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention," In: *Info. Proc. in Medical Imaging (IPMI)*, pp. 592–604, 2019. Sec. 5.1

- R. Droste, P. Chatelain, L. Drukker, H. Sharma, A. T. Papageorghiou, and J. A. Noble, "Discovering Salient Anatomical Landmarks by Predicting Human Gaze," In: *IEEE Int. Symp. on Biomedical Imaging (ISBI)*, 2020. Sec. 5.2
Oral presentation, 2nd Runner Up for Best Paper Award.

- R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Automatic Probe Movement Guidance for Freehand Obstetric Ultrasound," In: *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020. Sec. 7.4
Nominated for MICCAI Young Scientist Award.

[1]Equal contribution.

2

Literature Review

Contents

Due to the interdisciplinary nature of this thesis, it builds upon several corpora of literature. Each research item presented in this thesis includes a dedicated summary of related work. In addition, we provide a brief overarching review here. In Sec. 2.1 we first review the literature of gaze-tracking in ultrasound imaging. Next, we briefly summarize prior work on visual saliency prediction in Sec. 2.2. Finally, we present a brief review of computer-aided image analysis for obstetric ultrasound in Sec. 2.3.

2.1 Gaze-Tracking in Ultrasound Imaging

Prior works on the use of gaze-tracking to study and advance ultrasound imaging can be divided into four categories [ZRS20]: (1) to assess the level of expertise of operators [BHK+18, KBV+15], (2) to develop a better understanding of sonographers' visual perception strategies [CBP+15], (3) to design better machine-operator interfaces [HSLF17, LTM+18, ZSR19], and (4) to aid algorithms for computer-aided image analysis [AN16a, AN16b, CSCN18b]. Besides category (4), which is discussed in Sec. 2.3, categories (2) and (3) are most relevant to our work. Category (1), assessing expertise, is currently being further explored by other members of the PULSE study [WDJ+20].

Regarding (2), ultrasound perception strategies, Carrigan et al. [CBP+15] asked sonographers to examine pre-acquired ultrasound images of the female breast and to determine whether the image contains a cancer malignancy while tracking their gaze. The authors found that the difficulty of diagnosis (as determined a priori by an experienced radiologist) correlates with the time-to-first-fixation (TFF), *i.e.*, the time until a lesion was fixated. This is in line with Kosevoi-Tichie et al. [KBV+15] who showed that lower expertise also correlates with the TFF. Carrigan et al. further found that for the malignancy that was overlooked most frequently, sonographers spent more time looking at irrelevant regions.

In regards to (3), better machine-operator interfaces, gaze-based human-computer-

interfaces (HCI) have been developed for ultrasound settings (zoom, gain, depth, Doppler mode) [LTM+18] and for performing measurements [ZSR19]. While these studies proposed *new* HCI systems that rely on the presence of an eye tracker to operate the ultrasound machine, our work in chapter 3 uses eye-tracking to discover limitations of the *clinical* HCI systems that are currently used. Many studies have examined the shortcomings of the outcomes of obstetric ultrasound scanning [ECF+93, BS93, BN00, Mar05, MBE+15], but ours are the first to examine the reasons for these shortcomings through eye-tracking.

2.2 Visual Saliency Prediction

2.2.1 Images

Several comprehensive reviews of image saliency prediction methods are available [BTSI13, BI13, JDT12, BRB+16, Bor18, KBJ+12]. Borji et al. [BSI13a] compared human-model agreement and time complexity for numerous methods. Bylinskii et al. [BRB+16] gave an insightful assessment of current models and their limitations. As mentioned in Sec. 1.6, early visual saliency models relied on maps of hand-crafted, low-level image features [KU85, IKN98, BI13, SF03, HKP07, LMLCBT06, JEDT09]. CNNs, first proposed for saliency prediction by Vig et al. [VDC14a], are seen as a breakthrough for the field in terms of prediction accuracy [KTB15, KWB16, JHDZ15b, PMS+16, KAB15, CBSC16, CBSC17a, CAB+17, WS18, HSBZ15a, JMV16, Jia18] and were reviewed by Borji et al. [Bor18]. Progress has been benchmarked by large, publicly available saliency datasets [JEDT09, MSHH11, JDT12, BI15, JHDZ15b, MS15, WSG+18] and tracked by the ongoing MIT300 Saliency Benchmark [KBJ+12]. A first significant performance gain was achieved by Kümmerer et al. [KTB15, KWB16] with the Deep Gaze I & II models which are pre-training on the ImageNet dataset [DDS+09], a procedure

that is commonplace since. Similar methods include the multi-resolution SALICON model [HSBZ15a] and the FCN DeepFix model [KAB15]. Cornia et al. [CBSC16] presented an LSTM-based architecture which iteratively focusses on different regions to refine the saliency map. State-of-the-art on the MIT300 (prior to our work in Sec. 4.2) was claimed by the heavily engineered EML-NET model [Jia18].

2.2.2 Videos

Saliency prediction in videos naturally differs from images since fixations in videos are highly correlated over time [Cou04] and since temporal features such as sudden changes can attract attention [GC18]. It is of particular interest for applications in robotics [YMG10], surveillance [JXZ14], compression [GZ10, GEV+14], as well as action recognition [VDC12] and salient object detection [WSS18]. Early models with hand-crafted features include Kienzle et al. [KSWF07] for free-viewing and Mathe and Sminchisescu [MS12] for task-related saliency prediction. Chaabouni et al. [CBH16] presented an early CNN-based approach, adding optical flow as an additional input channel to a single-frame CNN. Bak et al. [BKEE18] proposed to include optical flow via a two-stream architecture [SZ14]. Bazzani et al. [BLT17] achieved greater temporal depth with a recurrent mixture density network (RMDN) by aggregating feature vectors with an LSTM. The feature vectors are extracted from short segments by a deep, 3D spatio-temporal CNN [TBF+15]. The LSTM is connected to a mixture density network [Bis94] which outputs the parameters for a Gaussian mixture model generating the saliency map. Jiang et al. [JXW18] observed that fixations in videos mainly fall on moving objects. The authors thus extracted spatio-temporal saliency features with an Object-to-Motion CNN (OM-CNN), a two-stream architecture consisting of a pre-trained YOLO object detector [RDGF15] and FlowNet optical flow network [DFI+15]. The features are aggregated across frames with a two-layer convolutional LSTM [SCW+15]. Wang et al. [WSG+18]

recently released a large (N=1000) video saliency dataset and benchmark, DHF1K. The authors achieved state-of-the-art on this benchmark with a relatively simple architecture consisting of a VGG-16 feature extractor followed by a single-frame saliency prediction attention module and a single-layer convolutional LSTM. The intermediate single-frame saliency predictor allowed the authors to jointly train on large single-image and video saliency datasets. A review of video saliency work concurrent to our own work (Sec. 4.2) is presented in subsection 4.2.2. Overall we note that the summarized methods make saliency predictions based on past and present video frames only. While this allows for real-time applications, we show in Sec. 4.1 that predicting saliency based on the entire video sequence yields higher accuracy and a better model of the underlying cognition for ultrasound imaging. Moreover, we observe that video saliency prediction and image saliency prediction were treated as separate tasks in prior work and therefore evolved into disjoint research areas with different methods and benchmarks. In Sec. 4.2 we show for the first time that these tasks can be approached with a unified model for mutual benefit.

2.3 Obstetric Ultrasound Image Analysis

As described in Sec. 1.3, computer algorithms that aid ultrasound image acquisition and interpretation are key for enabling the broad deployment of the technology. In the following we review prior work on standard plane detection and on segmentation and landmark detection, which is relevant to our work in chapter 5.

2.3.1 Standard Plane Detection

Standard plane detection refers to the task of classifying ultrasound video frames as one of the standard anatomical views introduced in Sec. 1.2 or, if the frame does not precisely match any of them, into a "Background" class. Therefore this method can be used to assist operators in acquiring images of these standard

views. However, the difficulty of distinguishing between standard views and similar-looking non-standard background views makes it a challenging task. For the same reason, the accuracy reported in papers depends on the respective ground truth: stricter definition of standard planes and sampling of Background frames close to standard planes lead to lower accuracy for a given classifier. Chen et al. [CNQ+15] proposed a neural network for the detection of the abdominal circumference plane (ACP), reporting an F1-score 71.2%. The model was pre-trained on the ImageNet dataset [DDS+09] and fine-tuned and evaluated on ultrasound video frames acquired with a simplified sweep-protocol by moving the ultrasound probe from the cervix upwards in one continuous motion. The standard plane frames in the sweeps were labeled retrospectively by a radiologist. The authors later extended this model to detect the ACP, "facial axial standard plane" and the four chamber view in routine freehand scans [CWD+17]. They employed a recurrent neural network which receives sequences of ultrasound video frames, yielding an average F1-score to 68.1%. Baumgartner et al. [BKM+17] pointed out that not all standard views are adequately captured by these works. The authors therefore presented the *SonoNet* model which detects 13 standard planes (see Fig. 1.1) in routine freehand US scans. The model is based on the VGG-Net architecture [SZ15] and achieves an average F1-score of 82.8% in this more practical setting. Afterwards, Schlemper et al. [SOC+18] extended the model with self-gated attention mechanisms (AG-SonoNet), computing a SonoNet F1-score of 93.1% with their testing protocol and an improvement to 93.3% by AG-SonoNet. Apart from 2D ultrasound, localization of standard planes in pre-acquired 3D ultrasound volumes has been explored [YRK+16, LKH+18]. However, this is not directly relevant to our work which focuses on routine 2D acquisition. Moreover, earlier work has addressed the problem of standard plane classification/categorization [YKPN15, MBN+17] but aims to recognize standard plane classes instead of distinguishing standard planes from Background frames and is therefore aimed at retrospective labeling and analysis instead of guidance.

Ahmed et al. [AN16b] developed the first gaze-tracking aided US image classifier, a bag-of-visual-words model was trained with SURF descriptors [BTVG06] around sonographers' fixations. The authors evaluated the model on the task of categorizing ultrasound images into the classes *Head*, *Abdominal* and *Femoral* and found accuracies on clinical data ranging of 76%, 68% and 64% for the HCP, ACP and FLP. The authors also provided the first analysis of sonographer visual search strategies by identifying the most common transition-patterns between these classes. The idea of aiding ultrasound image classification with gaze-tracking data was transferred to the detection of the ACP with deep-learning methodology by Cai et al. [CSCN18a, CSCN18b]. The original SonoEyeNet architecture [CSCN18a] relies on the recorded gaze maps as additional input, while the later Multi-Task SonoEyeNet [CSCN18b] jointly performs saliency prediction and image classification. For the latter, the predicted saliency map is element-wise multiplied with the feature maps before classification, acting as an attention module. The authors showed that saliency-guided prediction is more accurate than purely image-based classification, achieving an F1-score of 93.5% for ACP detection with sweep data compared to 80.7% for the SonoNet [BKM+17] baseline. Moreover, the authors introduced an adversarial loss [GPM+14] for saliency map prediction which increased the F1-score from 93.5% to 96.5% due to an improvement of the recall from 90.5% to 96.2%. Concurrently with this thesis, Cai et al. [CDS+20] extended the idea of saliency-aided detection to the distinction of ultrasound video clips which contain the HCP, ACP or FLP from "background" clips. As for the Multi-Task SonoEyeNet, predicted visual saliency maps are used as attention maps for the detection module. The authors also applied a soft dynamic time warping (sDTW) [CB18] loss which improves saliency prediction for video data. Finally, the authors provided qualitative analyses of the ultrasound saliency predictions to gain insights into the sonographer visual search strategies. In our work in 5.1 and Sec. 5.3 we show for the first time that ultrasound image representations can be learned independently of any manual annotations,

from gaze data alone. This allows us to formulate a large-scale pre-training and fine-tuning framework that significantly boosts performance on the standard plane detection tasks with all 13 planes compared to training without gaze data.

2.3.2 Segmentation and Landmark Detection

Methods for the segmentation of the HCP, ACP and the FLP (as well as the whole fetus at the first trimester) up until 2014 were reviewed in a corresponding grand challenge [RFK+14]. A more recent overview of state-of-the-art deep learning based methods for the segmentation and measurement of the HC was provided by Zeng et al. [ZTW+21]. Most related to the segmentation component of our work in Sec. 5.3 is the work of Wu et al. [WXL+17] who applied fully convolutional networks (FCNs) for the segmentation of the HCP and ACP. The authors reported IoU scores of 96.9% and 96.3% for HC and AC segmentation. In contrast to our work, the method does not segment the internal structures (CSP, UV, SB, spine, aorta, ribs), does not include the femur annotation and applies to the second trimester only. Not covered by Zeng et al. [ZTW+21] is recent work by Sinclair et al. [SBM+18] which claimed the first method that outperforms human annotators at HC measurement. Their method is also based on a FCN and the HC is measured via least-squares fitting of an ellipse to the segmented contour. There is also substantial work on the 3D segmentation of fetal ultrasound volumes, see *e.g.* [YRK+16, YYL+18], but is not directly relevant to our work which focuses on routine 2D freehand ultrasound.

Most prior work related to the landmark detection component of Sec. 5.3 is either based on hand-crafted features [CGGC08], different from the representation learning approach that we take, or pertains to 3D ultrasound [HXAN18, WYD+19]. Ahmed et al. [AN16a] used gaze-tracking to select anatomical structures for the detection of the ACP in ultrasound images based on image intensity and gradient features. Recently, a method was proposed for the weak, heatmap-based localization

of fetal structures [TKS+18] but it only applies to coarse structures such as head, abdomen, *etc.* and not to individual anatomical landmarks. In Sec. 5.3 we show that gaze-tracking data aids the training of ultrasound segmentation and landmark detection and our proposed method can be integrated with any existing encoder-decoder based model.

Related to our work in Sec. 5.2, the concept of salient landmarks is well-established in computer vision and has been applied to medical imaging. Here, *saliency* is often used to refer to low-level features such as local entropy [KB01, GFD06]. Mutually-salient landmarks based on Gabor attributes have been proposed for image registration [OSPD11]. In contrast, our work introduces the concept of *visually salient landmarks*, *i.e.*, anatomical landmarks that are learned based on actual human visual attention as measured with gaze-tracking.

2.4 Summary

In summary, gaze-tracking for ultrasound imaging has been explored in individual prior studies but is nascent compared to purely-image based methods. Regarding sonographer-machine interaction, gaze has been mostly used to evaluate expertise and to develop novel gaze-based machine interfaces but not to study the shortcomings of current interfaces in a clinical setting. Regarding the prediction of gaze with machine learning, CNNs emerged as powerful models for image and video saliency prediction but gaps remain in bridging these two tasks and in methods that are suitable for ultrasound imaging. Finally, ultrasound image analysis research has seen gains in performance through advances in machine learning. Initial works exist that leverage saliency prediction for gaze-based standard plane and landmark detection models. We see promising opportunities to advance this research direction with more general models that are pre-trained with large, unlabeled gaze-tracking data.

3

Analysis of Sonographer Gaze Patterns

Contents

3.1 Expected-Value Bias in Routine Third-Trimester Growth Scans

Authors. Lior Drukker*, Richard Droste*, Pierre Chatelain, J. Alison Noble, Aris T. Papageorghiou *equal contribution

Journal. *Ultrasound in Obstetrics & Gynecology*, vol. 55, no. 3, pp. 375–382, 2020.

Background. A central goal of obstetric ultrasound scanning is to identify fetal growth restriction (FGR). To accomplish this, biometric measurements of the head circumference, abdominal circumference and femur length of the fetus are measured on ultrasound images captured during the second and third trimester. The measurements are then compared against reference distributions for the respective gestational age of the fetus that is estimated based on a first trimester scan. If the measurements falls below the 10th percentile, FGR may be present. However, previous studies have shown that the biometric measurements predict low birth weight with low accuracy [Dud05, MBE+15]. The gaze-tracking data of the PULSE study allowed us to identify a possible cause of these shortcomings: *expected value bias* which arises when sonographers adjust the biometric measurements to match the expected healthy measurement value for the known gestational age.

Statement of Authorship. I was co-lead author, responsible for designing and performing the data extraction, data analyses, visualizations and the majority of statistics of the paper, and contributed to the original draft. Lior Drukker proposed the problem, and was the main responsible for data acquisition, clinical interpretation of results and the original draft. Aris T. Papageorghiou and J. Alison Noble were responsible for supervision and funding acquisition and contributed to conceptualization, methodology and editing the draft. Pierre Chatelain contributed to data acquisition, statistical analyses and editing the draft.

Abstract

Objectives: Operators performing fetal growth scans are usually aware of the actual gestational age. This may lead to an expected value bias when performing biometric measurements.

Methods: We prospectively collected full-length video recordings of routine ultrasound growth scans coupled with operator eye-tracking. The expected value was defined as the actual gestational age at the time of the scan. Expected value bias was defined as occurring when the operator looked at the "measurement box" during the process of caliper adjustment before saving a measurement. We studied the three standard biometric planes on which measurements are taken: Head Circumference (HC), Abdominal Circumference (AC), and Femur Length (FL). We evaluated the occurrence of expected value bias and quantified the impact of biased measurements.

Results: We analyzed 272 third trimester growth scans with a total of 1409 measurements (354 HC, 703 AC, and 352 FL) performed by 16 operators. Expected value bias occurred in 91.4% of the saved standard biometric plane measurements (85.0% for HC, 92.9% for AC, and 94.9% for FL). The operator adjusted the measurement toward / away from the expected value in 48% /20% of the biased standard plane measurements ($p<0.001$). On average, measurements were corrected by 2.3 $\pm$ 5.6, 2.4 $\pm$ 10.4, and 3.2 $\pm$ 10.4 days of gestation towards the actual gestational age for the HC, AC, and FL, respectively. Additionally, we note a statistically significant reduction in measurement variance once the operator was biased ($p=0.026$). Comparing the lowest and highest possible estimated fetal weight (EFW), we note that the discordance, in percentage terms, was 10.1% $\pm$ 6.5% and that in 17% (95% CI 12-21%) of the scans, the fetus could be considered as small-for-gestational-age or appropriate-for-gestational-age if using the smallest or largest possible measurements, respectively. Similarly, in 13% (95%CI 9-16%) the fetus could be considered as large-for-gestational-age or appropriate-for-gestational-age if using the largest or smallest possible measurements, respectively.

Conclusions: During routine third-trimester growth scans, expected value bias frequently occurs and significantly changes measurements of standard biometric planes.

3.1.1 Introduction

In science, the accuracy of measurement is a crucial prerequisite for correct interpretation of results. There are many reasons for inaccurate measurement, and one that is relatively easy to overcome is the observer bias, which is the tendency to see what we expect to see [VGB+18]. The observer bias is also known as "expected value bias", "detection bias", "observer-expectancy effect", "expectancy bias", "observer effect", or "ascertainment bias". This bias may occur if the observer has a preconceived idea of what a measurement ought to be, leading to adjustments of the readings. Hróbjartsson and colleagues undertook a systematic review quantifying the impact of observer bias, by comparing estimates from studies in which outcome assessors were blinded to the intervention with those in which outcome assessors were not blinded [HTE+13]. For clinical trials that used measurement scale outcomes, non-blinded outcome assessment exaggerated effect size by as much as 68% [HTE+13]. In randomized trials, blinding is used to reduce bias; this is usually to prevent knowledge of which intervention or control is being received by a study participant [Sac79, SG02]. Day and Altman highlight that blinding is important in other types of research too, such as evaluation of the performance of a diagnostic test, and reproducibility of measurement techniques [DA00]. Blinding makes it difficult to bias results intentionally or unintentionally and so helps to ensure the credibility of measurements [DA00]. Recently, a review of systematic error and cognitive bias in obstetric ultrasound suggested that the "expectation bias" is pertinent to obstetric ultrasound studies [SO19].

In contrast to trials, measurement blinding is not usually carried out in day to day clinical management. This may be of particular relevance to fetal growth assessment, where screening looks for aberrations from normally expected growth patterns; blinding to the number of gestational weeks to avoid the effect of clinician bias is rarely practiced. In fact, during clinical assessment of fundal height, the guidance

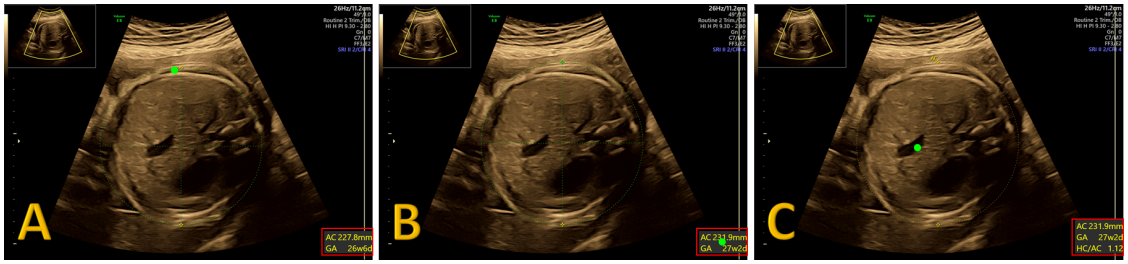

Figure 3.1: Abdominal circumference measurement at 28+0 weeks gestation. A red rectangle was drawn to outline the "measurement box" and the green dot was added to represents the operator eye focus (not visible to the operator during the scan). A. Caliper adjustment in progress. B. Operator eye focus detected at the measurement box; hence this is a biased measurement. C. Accepted Measurement.

suggests that caregivers should hold the tape in a way that the measurement cannot be seen [ESS94]. This is however not usually the case in ultrasound assessment: during a routine growth scan - comprising the three standard biometric plane measurements of Head Circumference (HC), Abdominal Circumference (AC), and Femur Length (FL) [PSI+13] – the ultrasound machine will usually display the reading value (circumference or length, in millimeters or centimeters) as well as an observed gestational age (in weeks + days) corresponding to the measurement (Fig. 3.1). This can lead to an observer (or expected value) bias, which means that the operator may adjust the circumference or length so that the observed gestational age matched the actual gestational age. In turn, this may lead to a biased fetal growth estimation. The use of blinding in this scenario would overcome such bias, and although some studies have blinded the operator to the actual gestational age or to the machine displayed values while performing growth scans [POA+14, ISN+13], measurement blinding is rarely used in routine clinical practice [ITO+12]. In this study we aimed to 1) evaluate the incidence of expected value bias in routine growth scans; 2) assess the impact of expected value bias on standard biometric plane measurements.

3.1.2 Methods

This was a prospective study of routine ultrasound scans performed in women with a singleton pregnancy, undertaken between May 2018 and August 2019 by sonographers and fetal medicine doctors at the Maternity Ultrasound Unit, Oxford University Hospitals National Health Services (NHS) Foundation Trust, Oxfordshire, United Kingdom. There, all women are offered three routine ultrasound scans: first-trimester crown-rump length (CRL) dating [SAB+13] at approximately 12 weeks which includes nuchal translucency measurement for first-trimester aneuploidy screening, a 20-week anomaly scan and a 36-weeks growth scan where estimated fetal weight is computed [HHM91]. Additionally, based on risk factors or clinical indications, women may be offered additional scans at other gestational ages [Roy13]. Ultrasound examinations are carried out or supervised by accredited sonographers or fetal medicine doctors using standard ultrasound equipment. For quality control measures, the stored images and the reliability of measurements are regularly assessed using the INTERGROWTH-21st quality criteria [SIO+13]. This study is part of a project entitled Perception Ultrasound by Learning Sonographic Experience (PULSE) [PUL]. This is an innovative interdisciplinary project that is designed to apply the latest ideas from machine learning and computer vision to build, from real-world video data and other sensory data, computational models that describe how an expert sonographer performs a diagnostic study of a subject from multiple perceptual cues. By understanding closely how experts learn and undertake diagnostic ultrasound, we believe that we will build considerably more powerful assistive interpretation methods than have been possible so far. In PULSE, we capture and record full-length routine ultrasound scan video; record probe movement; and track the point-of-gaze of the sonographer on the monitor of the ultrasound scanner.

All ultrasound scans included in this study were performed using commercial

Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machines equipped with standard curvilinear (C2-9-D, C1-5-D), and 3D/4D (RAB6-D) probes. Synchronized eye-tracking was undertaken using an eye tracker (Tobii Eye-tracking Eye Tracker 4C, Danderyd, Sweden) attached to the ultrasound machine; the validity of eye-tracking has been previously reported [CSD+18].

This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051), and written informed consent was given by all participating pregnant women. Sonographers also consented to participate in the study at the outset, but do not have any visual or other signal to know whether tracking devices are functioning. The PULSE project is funded by the European Research Council (grant ERC-ADG-2015 694581).

Definitions

"Standard biometric planes" are the three standard biometric ultrasound views acquired during growth scans: Head Circumference (HC), Abdominal Circumference (AC), and Femur Length (FL).

"Measurement box" is a box displayed on the ultrasound screen while measuring a standard biometric plane. In the box, the ultrasound machine displays the circumference or length value (in centimeters or millimeters) as well as a gestational age corresponding to the measurement (in weeks + days, see Fig. 3.1).

"Actual gestational age" is the gestational age at the time of growth scan, based on the estimated due date (EDD) that was established at the dating scan.

"Observed gestational age" is the gestational age displayed in the measurement box, which is based on a standard biometric plane measurement.

"Expected value bias": In our settings, like many others, the operator is aware of the *"actual gestational age"*. If before saving a standard biometric plane measurement, the operator looks at the *"measurement box"*, then we considered there to be a risk of bias of the measurement (Fig. 3.1).

Measurement *"before"* and *"after"* bias: we classified the *"before"* measurement as the measurement shown in the *"measurement box"* at the time of first look. The *"after"* measurement, potentially influenced by expected value bias, is the one saved by the operator.

"repeat measurement": After a (either HC, AC, or FL) specific standard biometric plane is saved, any additional same standard biometric plane saved is a *"repeat measurement"*.

Biometry acquisition

The process of acquiring a standard biometry plane is a three-stage process. First, the operator obtains an optimal acquisition of a standard biometric plane and freezes it on the screen. Next, the operator measures the biometric variable by placing calipers on the image; automatic caliper placement is turned on by default in our unit. The operator will often adjust caliper placement to achieve the best visual fit. During caliper placement and adjustment, ultrasound machines display a "measurement box" where parameters are shown and updated in real-time: the measured length or circumference (in centimeters or millimeters or centimeters) and the gestational age derived from the measurement (in weeks + days, see Fig. 3.1). Finally, the operator accepts the standard biometric plane measurement by saving the image with a visible measurement.

Data extraction

Each scan was automatically analyzed on a video frame-by-frame basis with a purpose-built software program implemented in Python (www.python.org, version 3.7.0) using OpenCV (opencv.org, version 3.4) and Tesseract (github.com/tesseract-ocr, version 3.05). For each scan video, the software program first detected episodes where the standard biometric plane was being measured, by the appearance of the measurement box. Next, for each standard biometric measurement, the program

detected uninterrupted fixations of the operator's eye on the measurement box lasting ≥100 ms. If the fixation was interrupted, it was considered as one single episode of eye fixation if this interruption was 400ms or less; or as a separate fixation if it was more than 400 ms [Bry95, SG00]. Additionally, we verified this threshold by randomly looking at more than 50 detected fixations and making sure that the threshold resulted in no false positives. Concurrently, the software program stored the values displayed in the measurement box when the calipers were initially placed, and when the operator "accepted" the measurement. Additionally, the software program stored the values displayed in the measurement box upon each detection of measurement box eye fixation. The measurement box values and parameters were extracted via optical character recognition (OCR).

The Voluson E8 BT18 machine, by design, displays the observed (measured) gestational age as "OOR" (out of range) in the measurement box when no standard curve is available for the measurement. In the current analysis, when this happened, the gestational age values were computed using the appropriate original formula.

Study aims

The study aims were: 1) to evaluate the incidence of expected value bias in routine growth scans; 2) to assess the impact of expected value bias on standard biometric plane measurements.

To evaluate the incidence of expected value bias we evaluated whether the operator looked at the "measurement box" before saving a standard biometric plane. To assess the impact of expected value bias, we measured: I) how often operators adjust the calipers toward or away from the expected value II) evaluated the deviation of measurements expressed as observed and actual gestational age before and after the expected value bias took place III) compared the deviation between the observed and actual gestational age for standard planes that were repeated versus those that were not, and IV) evaluated the impact of expected value

bias on the estimated fetal weight (EFW) by calculating the lowest and highest possible EFW in measurements, using the smallest and largest HC, AC, and FL before and after expected value bias occurred.

Statistical Analysis

We report descriptive statistics. Continuous variables were compared using the Student t-test, Wilcoxon signed-rank test (paired) or Mann–Whitney U-test (unpaired). Comparison of saved measurements and those noted when the operator looked for the first time were investigated using multiple linear regression models. In order to evaluate independent relationships between the number of repeat measurements and the absolute deviation from the actual gestational age (expected value), we conducted a multifactor ANOVA. Analyses were adjusted for the BMI of the pregnant woman and the number of years of scanning experience of the operator. P values less than 0.05 were considered statistically significant. Analyses were carried out in R (www.r-project.org, version 3.5.2), Python (www.python.org, version 3.7.0), Pandas (pandas.pydata.org, version 0.24.0), SciPy (www.scipy.org, version 1.1.0), and Matplotlib (matplotlib.org, version 3.0.0).

3.1.3 Results

During the study period, a total of 272 women attending a routine growth scan were recruited. The demographic characteristics of the participants are displayed in Table 3.1. The mean actual gestational age for women attending a growth scan was 34.6 ± 3.1 weeks. There were 16 operators of which nine were accredited sonographers and seven fetal medicine doctors, with a median of three years (range: four months to 14 years) clinical post-accreditation experience in sonography.

A total of 1409 standard biometric plane measurements were made in the 272 scans: 354 of the HC, 703 of the AC, and 352 of the FL. We observed there to be a

Table 3.1: Characteristics of pregnant women and operators participating in this study. Data are mean ± standard deviation or number (percent).

Pregnant women	n = 272
Maternal age (years)	31.9 ± 5.7
Smoking at booking	21 (7.7%)
In vitro fertilization (IVF)	4 (1.5%)
Body Mass Index at <15 weeks (kg/m2)	25.8 ± 5.3
Dating by crown rump length (CRL)	249 (91.5%)
Nulliparous	123 (45.2%)
Pre-eclampsia	7 (2.6%)
Gestational diabetes mellitus	11 (4.0%)
Actual gestational age at growth scan (weeks)	34.6 ± 3.1
Preterm birth	11 (4.0%)
Vaginal birth	203 (74.6%)
Operators	n = 16
Gender	
Female	14 (87.5%)
Male	2 (12.5%)
Years of clinical experience in scanning	
<2 years	3 (18.8%)
2 - 5 years	7 (43.8%)
5 - 10 years	5 (31.3%)
>10 years	1 (6.3%)
Accreditation	
Sonographer	9 (56.3%)
Fetal medicine doctors	7 (43.8%)

risk of measurement bias in 91.4% of the standard biometric plane measurements, of which 85.0%, 92.9%, and 94.9% were of the HC, AC, and FL measurements, respectively (Table 3.2). Importantly, there was evidence that looking at the measurement box during caliper adjustment was likely due to bias rather than due to other reasons: thus, operators were more likely to adjust measurements towards the actual gestational age than to adjust away from the actual gestational age (47.7% vs. 19.7% overall; 49.5% vs. 16.4% for HC, 51.5% vs. 26.3% for AC, and 38.9% vs. 9.6% for FL) ($p<0.001$) (Table 3.2). The risk of biased measurements applied to all operators, though it varied from 56% to 100% of measurements for the different operators. The correlation between years of scanning experience of an operator and

Table 3.2: Expected value bias according to the standard biometric plane. Data are number, mean ± SD, or percent.

	Saved standard plane measurements	Repeated measurements	Standard plane measurements per growth scan	Biased standard plane measurements
HC	354	82	1.3 ± 0.6	85.0%
AC	703	431	2.6 ± 1.0	92.9%
FL	352	80	1.3 ± 0.7	94.9%
Total	1409	593	5.2 ± 1.7	91.4%
	Adjustment towards the actual gestational age	Adjustment away from the actual gestational age	Mean adjustment (days of gestation)	p-value
HC	49.5%	16.4%	2.3 ± 5.6	<0.001
AC	51.5%	26.3%	2.4 ± 10.4	<0.001
FL	38.9%	9.6%	3.2 ± 10.4	<0.001
Total	47.7%	19.7%	2.6 ± 9.5	<0.001

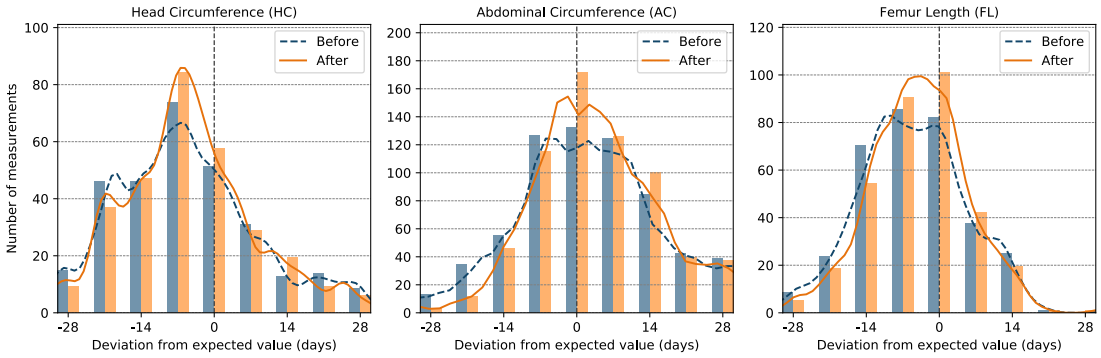

Figure 3.2: Deviation between the observed and the actual gestational age, expressed in days of gestation. Deviation before bias (blue), i.e., before gazing at the measurement box and after (orange) saving the measurement.

the percent of measurements prone to bias was not statistically significant ($p = 0.34$). The deviation between the observed and actual gestational age, expressed in days of gestation, is presented in Fig. 3.2. We found a statistically significant difference in the mean observed gestational age before and after operators looked at the measurement box. Hence, the HC, AC, and FL were closer to the actual gestational age by 2.3 ± 5.6, 2.4 ± 10.4, and 3.2 ± 10.4 days of gestation ($p<0.001$), respectively. Additionally, we note that values were closer to the mean after the operator was

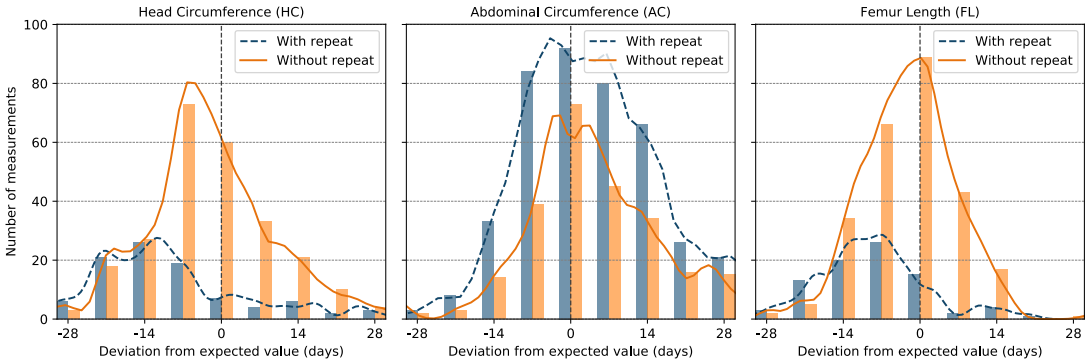

Figure 3.3: Deviation between the observed and the actual gestational age, expressed in days of gestation. Measurements that were repeated afterward are in blue and measurements not repeated in orange.

biased (reduction of the variance, Levene's Test, p=0.0255). These correlations remained statistically significant after multivariable analysis using the BMI of the pregnant woman, and operator identity as confounding variables. Additionally, when there was evidence of bias, we compared the measurement at the time the operator first looks at the measurement box, and that eventually saved. We note that there is a correlation between the initial measurement and the one saved: the further the initial measurement was from the expected value, the larger the adjustment of calipers toward the expected value ($p<0.001$ for AC, HC, and FL). This correlation remained significant after adjusting for operator experience and maternal BMI.

We also compared the deviation between observed and actual gestational age for those measurements that were repeated versus those that were not. A total of 82, 431, and 80 measurements were repeated for the HC, AC, and FL planes, respectively. Operators were more likely to repeat a measurement when the measurement was "far" from the expected value.

The observed gestational age was significantly closer to the actual gestational age for measurements that were not repeated. This means that operators were more likely to acquire another image of the same standard biometric plane and measure again when the initial measurement was far from the expected value (HC:

15.1 ± 8.4 vs. 10.2 ± 10.9 days, p<0.001; AC: 12.4 ± 14.3 vs. 11.5 ± 12.3 days, p=0.036; FL 13.3 ± 11.1 vs. 7.7 ± 9.7 days, p<0.001) (Fig. 3.3). This correlation remained statistically significant after multivariable analysis using the BMI of the pregnant woman, and operator identity as confounding variables.

Finally, to calculate the impact of this possible bias, we calculated the lowest and highest EFW, using the smallest and largest biased HC, AC, and FL measurements. The discordance, expressed in percentage terms, was 10.1% ± 6.5%. The z-score difference between the highest and lowest possible EFW was 0.83 ± 0.58. This means that 46 fetuses (17%, 95%CI 12-21%) could be considered as small-for-gestational-age if using the smallest possible measurements and appropriate-for-gestational-age if using the largest possible measurements. Similarly, in 34 scans (13%, 95%CI 9-16%) the fetus could be considered as large-for-gestational-age or appropriate-for-gestational-age if using the largest or smallest possible measurements, respectively.

3.1.4 Discussion

This study has demonstrated that fetal measurements undertaken during growth scans are often biased by knowledge of the gestational age and the expected measurement for gestation. Operators tend to "correct" caliper placement at the time of the scan toward the expected measurement for the actual gestational age. The amount of "correction" correlates with the amount of deviation from the expected value. Additionally, we noted that operators were more likely to retake an image and repeat a measurement when the first measurement was "far" from the expected value. We did not find a correlation between the tendency to undertake such correction and the number of years of clinical experience or type of accreditation.

It is difficult to compare our findings with previous reports, as observer bias / expected value bias is not well studied in obstetric ultrasound. Nevertheless, unbiased and accurate measurement is a fundamental tenet of science. Such

bias is not limited only to obstetric ultrasound, but can be encountered in many other medical fields and is known to significantly modify clinical measurements as well as experimental results [MSBH18]. For example, knowing what the blood pressure ought to be, might lead to an arbitrary adjustment of a non-automatic reading [BLO01].

The magnitude of the effect is difficult to ascertain and requires a study comparing blinded and non-blinded fetal biometric measurements. Nevertheless, we found that the impact of bias on estimated fetal weight may be as high as 10% and that in 17% of scans the fetus could be considered as small-for-gestational-age or appropriate-for-gestational-age depending on whether the smallest or the largest possible bias measurement is used. The corresponding figure for fetuses that could be considered as large-for-gestational-age or appropriate-for-gestational-age was 13%. This can obviously lead to erroneous detection of growth restriction and thus to unnecessary intervention, maternal anxiety, and iatrogenic perinatal morbidity; or may lead to inadvertently overlooking small for gestational age fetuses and classifying them as normal, putting them at risk for adverse perinatal outcome [MET98]. Hence, when making obstetric decisions, the possibility of bias in the estimation of fetal weight should also be taken into account. Moreover, in clinical practice, it is known that the detection rates for growth restriction during screening remain limited, and one could hypothesize that such an effect may contribute to this.

The finding also has obvious and important implications on research that is based on routine clinical data acquisition, for example when studying normal fetal growth. Bias of measurements means that any underlying formula relating gestation to the fetal measurement that is already programmed in the ultrasound system will have an important effect when aggregating data. It is for this reason that blinding operators to the measurement value is such a crucial step when creating normal ranges [POA+14, ISN+13, ITO+12]. In addition, this study is part of the PULSE project which is designed to apply the latest ideas from artificial intelligence,

machine learning and computer vision to build computational models that describe how expert sonographers scan. Our findings emphasize the importance of minimizing bias when training computer models to perform a task. This is because artificial intelligence is trained by humans who may introduce their own biases to the learning process, resulting in biased models. With current practice, an algorithm training to measure standard biometric planes might end up having a built-in bias when automatically calculating fetal biometry. This bias can potentially even be amplified by the algorithm [CDP+19, Cou18].

In our study all growth scans were routine, and most were of appropriate-for-gestational-age grown fetuses. It is possible that for small- and large-for-gestational-age fetuses this bias may be more pronounced as greater measurement "correction" towards the expected would be anticipated. This may be compounded by the well documented larger errors that exist in estimated fetal weight estimation for small and large for gestational age fetuses [Dud05].

The accuracy and reliability of fetal biometry measurements are determined by the accuracy of standardized biometric plane acquisition [SBD+06] and caliper placement. To study the effect of bias during caliper placement, we tracked the eye movements of the operator. Risk of bias would be present when the operator looks at the measurement box while adjusting caliper placement or saving the image. However, a biased measurement does not necessarily mean that the measurement is incorrect. Extreme values are likely to represent a low-quality acquisition rather than a growth concern. Therefore, operators may commonly look at the displayed measurement to ensure that their measurement meets their expectation before adjusting the calipers. Likewise, adjusting the measurement away from the actual gestational age does not necessarily represent an unbiased measurement. For example, if the operator is aware of gestational diabetes, the operator may unconsciously perceive that the fetus is big, and hence measure it to be large-for-gestational-age. Nevertheless, our findings suggest that, on

average, operators adjust the measurement towards the expected measurement for gestational age. Similarly, performing a repeat standard biometric plane acquisition and measurement may represent good practice [PSI+13]. Nonetheless, operators may choose to acquire an additional standard biometric plane measurement due to an unsatisfactory self-scoring quality assurance [PSI+13], or because of a measurement value that does not match closely enough the expected one. We did notice that measurements which were not repeated afterward were closer to the actual gestational age, which is the expected value.

Our study has some limitations. It was conducted in a single maternity unit which may not represent practice at other centers; nevertheless, we used 16 operators and the same finding was seen in all, making external validity more likely. In addition, operators were aware that scans and eye movements were being recorded. However, the operators were not informed of the aim of the current analysis meaning it is unlikely that they acted differently while participating in this study. Another limitation is that we could only estimate the impact of bias. To accurately examine the impact of bias a study where operators are randomly assigned to blinding of measurements would need to be conducted. However, the principle shown in this paper suggests that expected value bias is both common and clinically significant. We recently reported that operators rarely look at the safety indices while they scan [DDC+20]. This suggests that eye-tracking of the operator is precise in detecting the point of gaze. The finding that operators look at measurements but not bioeffects is in accordance with our assumption. Finally, we used the actual gestational age as the reference value, however, in our settings, the gestational age is based on a measurement performed at the dating scan which may also be biased [NDO+14].

In conclusion, observer bias towards expected values of fetal measurements is prevalent in routine third-trimester growth scans. Further research should evaluate the added value of eliminating this bias to the overall accuracy of growth scans. To overcome it, ultrasound manufacturers should consider including settings that allow

operators to be blinded before saving or ending ultrasound examinations.

Bibliography

[BLO01] G. Beevers, G. Y. Lip, and E. O'Brien. ABC of hypertension. Blood pressure measurement. Part I-sphygmomanometry: Factors common to all techniques. *BMJ*, 322(7292):981–985, April 2001.

[Bry95] Marc Brysbaert. Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, 124(4):434–452, 1995.

[CDP+19] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*, 28(3):231–237, March 2019.

[Cou18] Rachel Courtland. Bias detectives: The researchers striving to make algorithms fair. *Nature*, 558(7710):357–360, June 2018.

[CSD+18] P. Chatelain, H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies. *IEEE Trans. Cybern.*, pages 1–11, 2018.

[DA00] S. J. Day and D. G. Altman. Statistics notes: Blinding in clinical trials and other studies. *BMJ*, 321(7259):504, August 2000.

[DDC+20] Lior Drukker, Richard Droste, Pierre Chatelain, J. Alison Noble, and Aris T. Papageorghiou. Safety Indices of Ultrasound: Adherence to Recommendations and Awareness During Routine Obstetric Ultrasound Scanning. *Ultraschall Med*, 41(2):138–145, April 2020.

[Dud05] N. J. Dudley. A systematic review of the ultrasound estimation of fetal weight. *Ultrasound Obstet Gynecol*, 25(1):80–89, January 2005.

[ESS94] Janet L. Engstrom, Claudia P. Sittler, and Karen E. Swift. Fundal height measurement: Part 5—The effect of clinician bias on fundal height measurements. *Journal of Nurse-Midwifery*, 39(3):130–141, May 1994.

[HHM91] F P Hadlock, R B Harrist, and J Martinez-Poyer. In utero analysis of fetal growth: A sonographic weight standard. *Radiology*, 181(1):129–133, October 1991.

[HTE+13] Asbjørn Hróbjartsson, Ann Sofia Skou Thomsen, Frida Emanuelsson, Britta Tendal, Jørgen Hilden, Isabelle Boutron, Philippe Ravaud, and Stig Brorson. Observer bias in randomized clinical trials with measurement scale outcomes: A systematic review of trials with both blinded and nonblinded assessors. *CMAJ*, 185(4):E201–E211, March 2013.

[ISN+13] Christos Ioannou, Ippokratis Sarris, Raffaele Napolitano, Eric Ohuma, M. Kassim Javaid, and Aris T. Papageorghiou. A longitudinal study of normal fetal femur volume. *Prenat Diagn*, 33(11):1088–1094, November 2013.

[ITO+12] C. Ioannou, K. Talbot, E. Ohuma, I. Sarris, J. Villar, A. Conde-Agudelo, and A. T. Papageorghiou. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. *BJOG*, 119(12):1425–1439, November 2012.

[MET98] M. Mongelli, S. Ek, and R. Tambyrajia. Screening for fetal growth restriction: A mathematical model of the effect of time interval and ultrasound error. *Obstet Gynecol*, 92(6):908–912, December 1998.

[MSBH18] Kamal Mahtani, Elizabeth A. Spencer, Jon Brassey, and Carl Heneghan. Catalogue of bias: Observer bias. *BMJ Evidence-Based Medicine*, 23(1):23–24, February 2018.

[NDO+14] R. Napolitano, J. Dhami, E. O. Ohuma, C. Ioannou, A. Conde-Agudelo, S. H. Kennedy, J. Villar, and A. T. Papageorghiou. Pregnancy dating by fetal crown-rump length: A systematic review of charts. *BJOG*, 121(5):556–565, April 2014.

[POA+14] Aris T. Papageorghiou, Eric O. Ohuma, Douglas G. Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A. Jaffer, Enrico Bertino, Michael G. Gravett, Manorama Purwar, J. Alison Noble, Ruyan Pang, Cesar G. Victora, Fernando C. Barros, Maria Carvalho, Laurent J. Salomon, Zulfiqar A. Bhutta, Stephen H. Kennedy, and José Villar. International standards for fetal growth based on serial ultrasound measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*, 384(9946):869–879, September 2014.

[PSI+13] A. T. Papageorghiou, I. Sarris, C. Ioannou, T. Todros, M. Carvalho, G. Pilu, L. J. Salomon, and International Fetal and Newborn Growth Consortium for the 21st Century. Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project. *BJOG*, 120 Suppl 2:27–32, v, September 2013.

[PUL] Perception Ultrasound by Learning Sonographic Experience, European Research Council (ERC) Advanced Grant. https://cordis.europa.eu/project/id/694581.

[Roy13] Royal College of Obstetricians and Gynaecologists. The Investigation and Management of the Small–for–Gestational–Age Fetus (Green-top Guideline No. 31). https://www.rcog.org.uk/globalassets/documents/guidelines/gtg_31.pdf, 2013.

[SAB+13] L. J. Salomon, Z. Alfirevic, C. M. Bilardo, G. E. Chalouhi, T. Ghi, K. O. Kagan, T. K. Lau, A. T. Papageorghiou, N. J. Raine-Fenning, J. Stirnemann, S. Suresh, A. Tabor, I. E. Timor-Tritsch, A. Toi, and G. Yeo. ISUOG practice guidelines: Performance of first-trimester fetal ultrasound scan. *Ultrasound Obstet Gynecol*, 41(1):102–113, January 2013.

[Sac79] David L. Sackett. Bias in analytic research. *Journal of Chronic Diseases*, 32(1):51–63, January 1979.

[SBD+06] L. J. Salomon, J. P. Bernard, M. Duyme, B. Doris, N. Mas, and Y. Ville. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound Obstet Gynecol*, 27(1):34–40, January 2006.

[SG00] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78, November 2000.

[SG02] Kenneth F. Schulz and David A. Grimes. Blinding in randomised trials: Hiding who got what. *Lancet*, 359(9307):696–700, February 2002.

[SIO+13] I. Sarris, C. Ioannou, E. O. Ohuma, D. G. Altman, L. Hoch, C. Cosgrove, S. Fathima, L. J. Salomon, and A. T. Papageorghiou. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics & Gynaecology*, 120(s2):33–37, 2013.

[SO19] A. Sotiriadis and A. O. Odibo. Systematic error and cognitive bias in obstetric ultrasound. *Ultrasound in Obstetrics & Gynecology*, 53(4):431–435, 2019.

[VGB+18] Jose Villar, Francesca Giuliani, Fernando Barros, Paola Roggero, Irma Alejandra Coronado Zarco, Maria Albertina S. Rego, Roseline Ochieng, Maria Lorella Gianni, Suman Rao, Ann Lambert, Irina Ryumina, Carl Britto, Deepak Chawla, Leila Cheikh Ismail, Syed Rehan Ali, Jane Hirst, Jagjit Singh Teji, Karim Abawi, Jacqueline Asibey, Josephine Agyeman-Duah, Kenny McCormick, Enrico Bertino, Aris T. Papageorghiou, Josep Figueras-Aloy, Zulfiqar Bhutta, and Stephen Kennedy. Monitoring the Postnatal Growth of Preterm Infants: A Paradigm Change. *Pediatrics*, 141(2), February 2018.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Expected-Value Bias in Routine Third-Trimester Growth Scans
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Lior Drukker, Richard Droste, Pierre Chatelain, J. Alison Noble and Aris T. Papageorghiou. "Expected-Value Bias in Routine Third-Trimester Growth Scans". Ultrasound in Obstetrics & Gynecology, vol. 55, no. 3, pp. 375–382, 2020.

Student Confirmation

Student Name:	Richard Droste		
Contribution to the Paper	Co-lead author, responsible for designing and performing the data extraction, data analyses, visualizations and the majority of statistics of the paper. Contributed to the original draft.		
Signature	Richard Droste	Date	31.05.2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. J. Alison Noble			
I confirm that the description above is accurate.			
Signature	J Alison Noble	Date	01.06.2021

This completed form should be included in the thesis, at the end of the relevant chapter.

3.2 Safety Indices of Ultrasound: Adherence to Recommendations and Awareness During Routine Obstetric Ultrasound Scanning

Authors. Lior Drukker*, Richard Droste*, Pierre Chatelain, J. Alison Noble, Aris T. Papageorghiou *equal contribution

Journal. *European Journal of Ultrasound*, vol. 41, no. 2, pp. 138–145, 2020.

Background. While obstetric ultrasound imaging is generally considered safe, it does emit thermal energy into the fetus and it is each operator's responsibility to adhere to the recommended limits [Ame16, Saf09]. However, the adherence to the recommended limits in clinical scanning has previously been studied only for single saved ultrasound images [BSFT14, SA09] and short video segments [NBO+15]. In contrast, the PULSE dataset consists of hundreds of full-length ultrasound videos, which enabled this first comprehensive examination of the adherence to the guidelines. In addition, gaze-tracking enabled the first study of whether the operators were paying attention to the safety indices during scanning.

Statement of Authorship. I was co-lead author, responsible for designing and performing the data extraction, data analyses, visualizations and the majority of statistics of the paper, and contributed to the original draft. Lior Drukker proposed the problem, and was the main responsible for data acquisition, clinical interpretation of results and the original draft. Aris T. Papageorghiou and J. Alison Noble contributed to supervision, funding acquisition, conceptualization, methodology and editing the draft. Pierre Chatelain contributed to the data acquisition, statistical analyses and editing the draft.

Recognition. The paper was selected as Editor's Choice.

Abstract

Purpose: To analyze bioeffect safety indices and assess how often operators look at these indices during routine obstetric ultrasound.

Materials and Methods: Automated analysis of prospectively collected data including video recordings of full-length ultrasound scans coupled with operator eye-tracking was performed. Using optical recognition, we extracted the Mechanical Index (MI), Thermal Index in soft tissue (TIs), and Thermal Index in bone (TIb) values and ultrasound mode. This allowed us to report the bioeffect safety indices during routine obstetric scans and assess adherence to professional organization recommendations. Eye-tracking analysis allowed us to assess how often operators look at the displayed bioeffect safety indices.

Results: A total of 637 ultrasound scans performed by 17 operators were included, of which 178, 216, and 243 scans were first, second, and third-trimester scans, respectively. During live scanning, the mean and range for TIb were 0.14 (0.1 to 3.0), TIs was 0.2 (0.1 to 1.2), and MI was 0.9 (0.1 to 1.3). The mean and standard deviation of TIb were 0.15 ± 0.03, 0.23 ± 0.09, 0.32 ± 0.24 in the first, second, and third trimesters. For B-mode, the highest TIb was 0.8 in all trimesters. The highest TIb was recorded for pulsed wave Doppler mode in all the trimesters. Recommended exposure times were kept in all scans. Analysis of eye-tracking suggested that operators looked at bioeffect safety indices in only 27 (4.2%) of the scans.

Conclusion: In this study, recommended bioeffect indices were adhered to in all routine scans. However, eye-tracking showed that operators rarely assessed safety indices during scanning.

3.2.1 Introduction

Animal studies suggest that prenatal ultrasound may produce biological effects on the exposed fetus [AGD+06, Abr12]. However, no consistent causal relationship between the proper use of diagnostic ultrasound and human biological effects (bioeffects) has been established [HCSA+16, TVM+09] apart from a weak association between ultrasound screening during pregnancy and non-right-handedness in later life [Sal11].

The interaction between ultrasound and tissue generates thermal and mechanical effects. The thermal effect is tissue heating due to the transformation of acoustic energy into heat. The mechanical effect is, in particular, a cavitation effect of microscopic, stabilized gas bubbles in the tissues due to direct tissue reaction to alternating positive and negative pressure. As gas bubbles do not seem to be present in fetuses, the risk of mechanical effects is believed to be minimal [HA17].

The Thermal Index (TI) was designed to indicate the risk of tissue heating, while the Mechanical Index (MI) indicates the risk of inducing cavitation. In obstetrics, TI is reported in two variants: Thermal Index in soft tissue (TIs), assumes that sound is traveling only in soft tissue and is monitored in early pregnancy when bone ossification is low; Thermal Index in bone (TIb) assumes that sound is at or near the bone. The presence of bone within the ultrasound beam increases the likelihood of temperature rise due to direct absorption in the bone itself and conduction of heat from bone to adjacent tissue [NFAC09, KJM+20]. Therefore, after ten weeks of gestation, it is recommended that TIb is used [tH12]. The real-time display of the TI and MI is colloquially known as the Output Display Standard (ODS) which was designed to provide the operator with quantitative safety-related information [BTHZ+00]. As a part of training, ultrasound operators learn about potential bioeffects of ultrasound, and how to monitor these indices while scanning. Nevertheless, knowledge of bioeffects and their output indices are lacking among ultrasound operators [Mar05, SA08].

The aims of this study were to assess values of safety indices in routine obstetric ultrasound practice; the adherence to professional guidelines [Ame16, Saf09]; and to evaluate how frequently on-screen displays are assessed by operators. This was achieved by automated analysis of recordings of full-length ultrasound scans with concurrent operator eye movement tracking.

3.2.2 Methods

This was a prospective study of routine ultrasound scans performed in all trimesters between May 2018 and March 2019 by sonographers and fetal medicine physicians at the maternity ultrasound unit, Oxford University Hospitals NHS Foundation Trust, Oxfordshire, United Kingdom. Here, all women are offered three routine ultrasound scans: first-trimester dating at approximately 12 weeks which includes nuchal translucency measurement for first-trimester aneuploidy screening, a 20-week anomaly scan, and a 36-weeks growth scan. Additionally, based on risk factors or clinical indication, women may be offered additional scans at other gestational ages [Roy13]. Ultrasound examinations are carried out or supervised by accredited sonographers or fetal medicine physicians using standard ultrasound equipment. For quality control measures, the stored images and the reliability of measurements are regularly assessed using the INTERGROWTH-21st quality criteria [SIO+13]. In the United Kingdom, color Doppler and pulsed-wave Doppler are not routinely employed as a part of the first-trimester screening for trisomies [Abr13]. Nevertheless, some operators are familiar with such advanced sonographic screening strategies [ATG+17] and may, therefore, choose to use Doppler and pulsed-wave Doppler as part of the first-trimester screening [SLA+11, RCH+16, vCH+16].

This study is part of a project entitled Perception Ultrasound by Learning Sonographic Experience (PULSE) [PUL]. This is an innovative interdisciplinary project that is designed to apply the latest ideas from machine learning and computer

vision to build, from real-world video data and other sensory data, computational models that describe how an expert sonographer performs a diagnostic study of a subject from multiple perceptual cues. By understanding closely how experts learn and undertake diagnostic ultrasound, we believe that we will build considerably more powerful assistive video navigation and interpretation methods than have been possible so far. In PULSE we capture and record full-length routine ultrasound scan video; record probe movement; and track the point-of-gaze of the sonographer on the monitor of the ultrasound scanner. All ultrasound scans included in this study were performed using a commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machines equipped with standard curvilinear (C2-9-D, C1-5-D), and 3D/4D (RAB6-D) probes. Synchronized eye-tracking was undertaken using an eye tracker (Tobii Eye-tracking Eye Tracker 4C, Danderyd, Sweden) attached to the ultrasound machine; the validity of gaze-tracking was previously validated [CSD+18]. This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051), and written informed consent was given by all participating pregnant women. Sonographers also consented to participate in the study at the outset, but do not have any visual or other signal to know that tracking devices are functioning.

The PULSE project is funded by the European Research Council (grant ERC-ADG-2015 694581).

Definitions and data extraction

The output display standard (ODS) is located at the upper right side of the screen, and it is where the ultrasound machine displays the TIb, TIs and MI values. The screen area around the TIb, TIs and MI values was defined as the ODS box (Fig. 3.4).

Each scan was automatically analyzed on a video frame-by-frame basis with a purpose-built software program implemented in Python (www.python.org, version 3.7.0) using OpenCV (opencv.org, version 3.4) and Tesseract (github.com/tesseract-

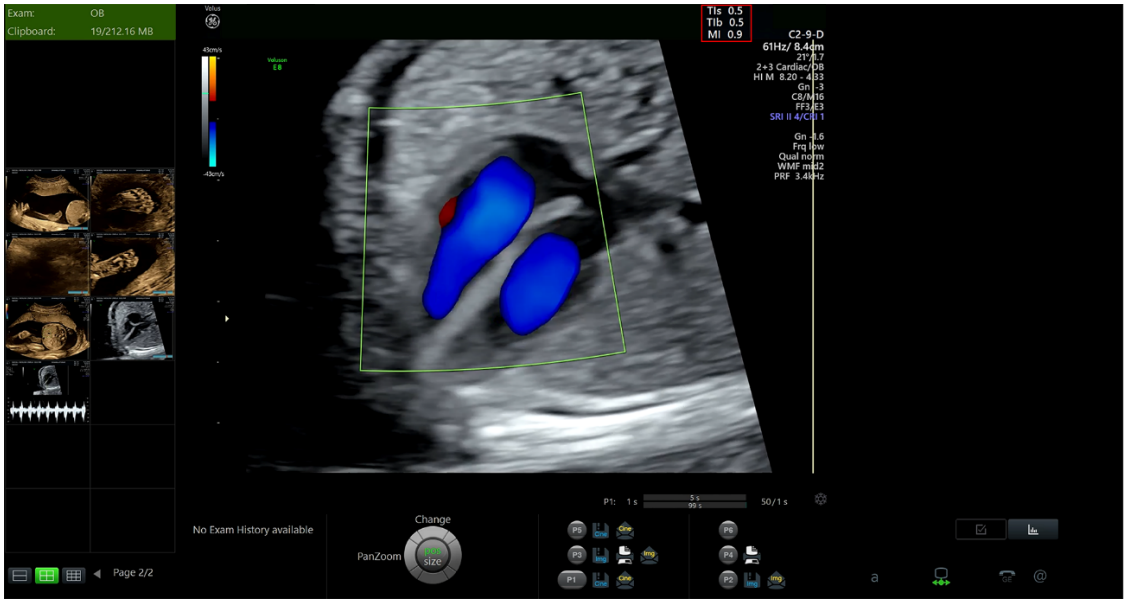

Figure 3.4: Output Display Standard (ODS) of the ultrasound safety bioeffect indices. Frame recorded during a routine scan. The ultrasound safety indices (TIs, TIb, MI), also known as the Output Display Standard (ODS), are displayed in real-time in the top right of the ultrasound image (red rectangle drawn around this for clarity. Patient identifiers as well as timestamp removed for anonymization).

ocr, version 3.05). For each scan video, the software program detected the exam mode by unique features apparent in the different exam modes (i.e., color palette). The "measurement box" values and parameters were extracted via optical character recognition (OCR).

The GE Voluson E8 BT18 machine, by design, displays the TI as <0.1 when energy emission is minimal. In the current analysis, for statistical purposes and because we were looking to ensure that safety is evaluated in a stringent approach when this happened values were recorded as 0.1.

For aim (3), we evaluated operator eye movements. Uninterrupted operator eye fixations on the "measurement box" of 100ms or longer were detected automatically. Eye fixations that were 300ms or less apart were classified as one fixation [Bry95].

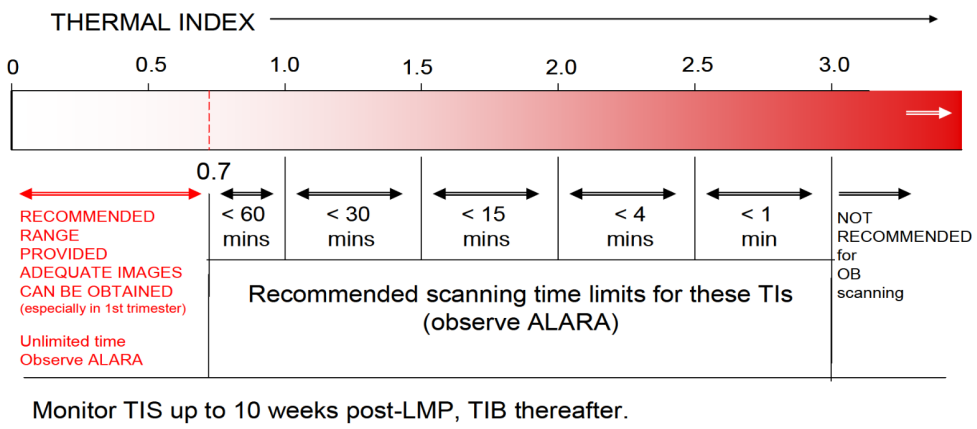

Figure 3.5: Recommended maximum scanning times for displayed Thermal Index (TI) values according to the American Institute of Ultrasound in Medicine (AIUM) and the British Medical Ultrasound Society (BMUS). Adapted with permission of the British Medical Ultrasound Society (BMUS).

Outcomes

Our goals were 1) to evaluate the bioeffect safety indices in routine scans; 2) to evaluate the adherence to the AIUM and BMUS TI safety guidelines (Fig. 3.5) [Ame16, Saf09]; and 3) to determine how often operators look at the displayed bioeffect safety indices.

Statistics

We report descriptive statistics. Analyses were carried out in Python (www.python.org, version 3.7.0), Pandas (pandas.pydata.org, version 0.24.0), SciPy (www.scipy.org, version 1.1.0), and Matplotlib (matplotlib.org, version 3.0.0).

3.2.3 Results

During the study period, a total of 637 women attending a routine obstetric scan agreed to participate. There were 178, 216, and 243 ultrasound scans performed at the first, second and third trimesters, respectively. The scans were performed by 17 operators, of which ten were sonographers and seven fetal medicine

Maternal	
Maternal age	31.7 ± 5.7
Body Mass Index at <15 weeks (kg/m2)	25.5 ± 5.5
Gestational age at scan	
First-trimester dating and NT scan	14.2 ± 7.7
Second-trimester anomaly scan	20.3 ± 0.5
Third-trimester growth scan	34.3 ± 3.3
Operator	
Gender	
Female	15 (88.2%)
Male	2 (11.8%)
Years of experience	
<2 years	2 (11.8%)
2-5 years	7 (41.2%)
5-10 years	5 (29.4%)
>10 years	2 (11.8%)
Accreditation	
Sonographer	10 (58.8%)
Fetal medicine doctor	7 (41.2%)

Table 3.3: Characteristics of 637 pregnant women and 17 operators participating in this study. NT, Nuchal Translucency. Data are mean ± Standard Deviation or number (percent).

	First trimester		Second Trimester		Third trimester	
	Mean ± SD	Max.	Mean ± SD	Max.	Mean ± SD	Max.
TIb	0.15 ± 0.03	2.1	0.23 ± 0.09	2.6	0.32 ± 0.24	3.0
TIs	0.15 ± 0.01	1.2	0.20 ± 0.03	1.2	0.24 ± 0.05	1.2
MI	0.95 ± 0.01	1.3	0.95 ± 0.01	1.3	0.91 ± 0.02	1.3

Table 3.4: Mean and maximal bioeffect safety indices measurements according to the trimester: Thermal Index for bone (TIb), Thermal Index for soft tissue (TIs), Mechanical Index (MI).

physicians. The demographic characteristics of the pregnant women and operators are given in Table 3.3.

The mean and standard deviation (SD) duration of live-scanning were 9.2 ± 6.5, 21.6 ± 11.1, and 7.2 ± 3.9 minutes for the first-trimester dating/NT, second-trimester anomaly, and growth scans, respectively.

During live scanning, the mean (range) first-trimester dating/NT scan TIb was

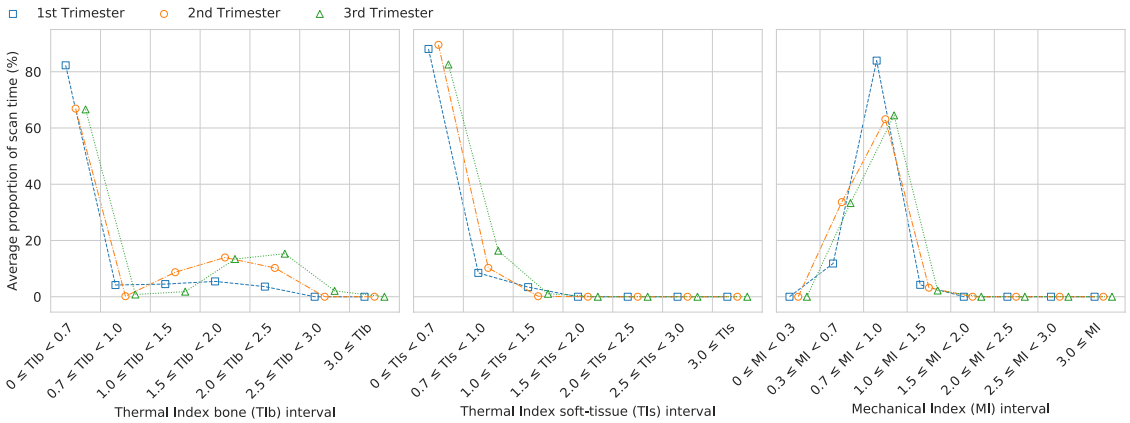

Figure 3.6: Scanning time (mean proportion) for the three ultrasound bioeffect safety indices (TIb, TIs, MI) values, in the different trimesters.

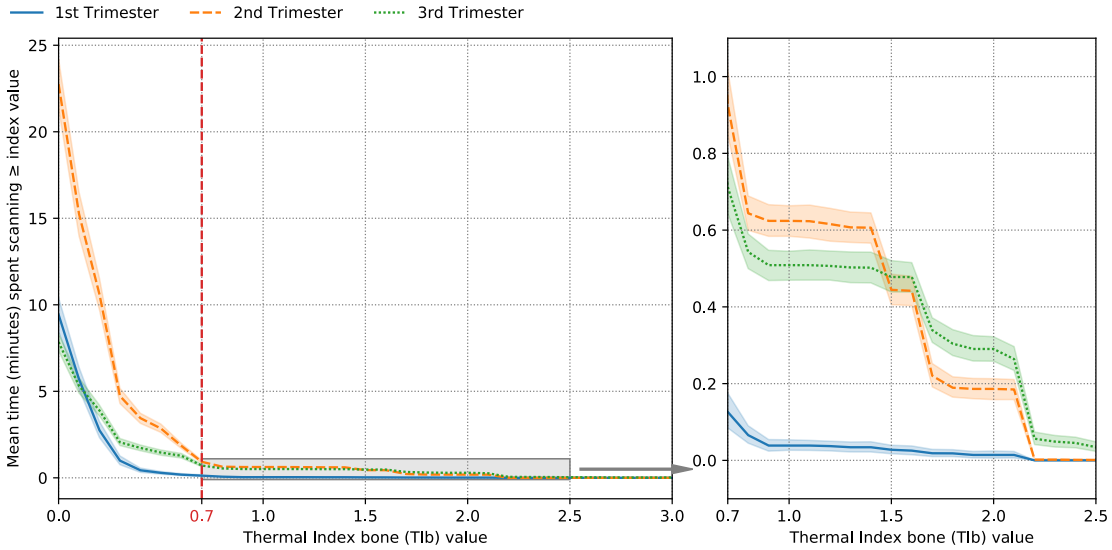

Figure 3.7: Cumulative scan time (mean and 95% CI in minutes) at or above a Thermal Index in Bone (TIb) value, in the different trimesters.

0.15 (0.1 to 2.1), TIs was 0.15 (0.1 to 1.2), and MI was 0.95 (0.1 to 1.3); the second-trimester anomaly scan TIb was 0.23 (0.1 to 2.6), TIs was 0.20 (0.1 to 1.2) and MI was 0.95 (0.1 to 1.3); and the third-trimester growth scan TIb was 0.32 (0.1 to 3.0), TIs was 0.24 (0.1 to 1.2) and MI was 0.91 (0.1 to 1.3). The bioeffect safety indices mean values and proportions for the different trimesters are shown in Table 3.4 and Fig. 3.6.

For B-mode, the highest TIb was 0.8 in all trimesters. The highest TIb was

	First trimester		Second Trimester		Third trimester	
	Mean ± SD	Max.	Mean ± SD	Max.	Mean ± SD	Max.
B-Mode TIb	0.14 ± 0.01	0.8	0.15 ± 0.01	0.8	0.14 ± 0.01	0.8
Color/Power Doppler TIb	0.62 ± 0.01	0.8	0.53 ± 0.01	0.8	0.52 ± 0.01	0.8
Pulsed Wave Doppler TIb	1.65 ± 0.26	2.1	1.67 ± 0.1	2.6	1.89 ± 0.18	3.0

Table 3.5: Mean and range of TIb values according to the ultrasound mode at the different trimesters.

AIUM and BMUS recommendations		Actual exposure time in 637 women					
	Max. exposure	TIb 1. trimester		TIb 2. trimester		TIb 3. trimester	
TIb[1]	time (min)	Mean ± SD	Max.	Mean ± SD	Max.	Mean ± SD	Max.
TIb ≤0.7	Unlimited	9.16 ± 6.37	35.2	21.0 ± 10.9	71.1	7.22 ± 3.93	28.3
0.7 <TIb ≤ 1.0	<60	0.03 ± 0.10	0.84	0.02 ± 0.11	1.11	0.03 ± 0.17	1.56
1.0 <TIb ≤ 1.5	<30	0.01 ± 0.05	0.30	0.17 ± 0.25	1.18	0.03 ± 0.10	0.76
1.5 <TIb ≤ 2.0	<15	0.01 ± 0.03	0.24	0.24 ± 0.21	1.07	0.22 ± 0.25	1.58
2.0 <TIb ≤ 2.5	<4	0.01 ± 0.06	0.40	0.17 ± 0.19	1.10	0.23 ± 0.24	1.31
2.5 <TIb ≤ 3.0	<1	0.00 ± 0.00	0.00	0.00 ± 0.00	0.01	0.03 ± 0.10	0.57
TIb >3.0	Not recommended	0.00 ± 0.00	0.00	0.00 ± 0.00	0.00	0.00 ± 0.00	0.00

Table 3.6: Thermal Index bone (TIb) recommended and actual exposure times in 637 full length routine scans. AIUM, American Institute of Ultrasound in Medicine; BMUS, British Medical Ultrasound Society. Figures are minutes.
[1]TIb should be used after 10 weeks of gestation (AIUM, BMUS guidelines).

recorded for pulsed wave Doppler mode in all the trimesters. Table 3.5 presents the TIb values according to the ultrasound mode (B-Mode, Color/Power Doppler, and Pulsed Wave Doppler) for the different trimesters. The cumulative scanning time at or above a TIb value is presented in Fig. 3.7. There were 41 (23.0%) first-trimester scans where the TIb was > 1.0. During the scans with TIb > 1.0, the average duration of TIb > 1.0 was 9.8 ± 7.6 seconds. The adherence to the AIUM and BMUS guidelines [Ame16, Saf09] in all ultrasound modes combined according to the different trimesters are noted in Table 3.6. In all scans, regardless of trimester, the recommended exposure times were adhered to.

Eye-tracking was successfully undertaken in all cases. This showed that the displayed bioeffect safety indices were looked at in 27 routine scans (4.2%), by

four of the 17 operators. In all 27 scans, we detected that the displayed bioeffect safety indices were checked once.

3.2.4 Discussion

In this paper, we report on the bioeffect safety indices (TIs, TIb, MI) for full-length routine obstetric scans as computed by automated analysis of video. We present results for all of the safety bioeffect indices (TIs, TIb, MI). However, it should be remembered that TIb is the most important indicator of possible bioeffect in pregnancies > 10 weeks of gestation. The recommended exposure times of TIb were kept in accordance with the current guidelines [Ame16, Saf09]. Additionally, we found that operators infrequently visually checked the bioeffect indicators on the ultrasound machine display. It is difficult to compare our results to previous publication as in previous studies appraising operator adherence to safety recommendations in routine practice relied on saved still images [BSFT14, SA09], and short cine-loops [NBO+15]. One study used tape-recorder entire scans. However, that study monitored a selected population of high-risk women in the second half of pregnancy only [DL00].

To simplify the approach to safe use of diagnostic ultrasound the ALARA (As Low As Reasonably Achievable) principle has been proposed [KJM+20, Tom06]. ALARA encourages scans to be restricted to medical indications, by trained professionals, using the lowest intensity power and the shortest duration of scanning as compatible with an accurate diagnosis. The International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) safety statement provides no absolute cutoffs for TI or MI, but states that "Exposure time and acoustic output should be kept to the lowest levels consistent with obtaining diagnostic information..." [AKM+03]. Hence, it is the responsibility of the operator to control the output energy safely. In addition to this general safety principle, the American Institute of Ultrasound in

Medicine (AIUM) and the British Medical Ultrasound Society (BMUS) guidelines provide detailed recommendations for the maximum exposure duration depending on the TI value [Ame16, Saf09]. We were able to show that in routine settings, the AIUM and BMUS recommended thresholds are kept with good adherence to the recommended exposure times, and not even once exceeded. In first-trimester Doppler examinations, ISUOG recommends that the TI should not exceed 1.0 [Ame16]. However, in our settings, pulsed wave Doppler is used only by some certified fetal medicine physicians in the first-trimester, and therefore, the ISUOG recommended TI was not always kept. Nevertheless, the optimal TI value remains elusive in all trimesters, as it is not apparent that there is any particular threshold for thermal induced damage [NFAC09]. Hence, users should adjust machine settings to obtain diagnostic images at the lowest possible acoustic output, recognizing that higher acoustic output does not necessarily improve image quality [SMEK13].

The "Output Display Standard" (ODS) should theoretically provide the necessary safety bioeffect indices information to the operator. However, there are concerns over its practicality [Mar05]. Using eye-tracking, we found that operators infrequently look at the safety bioeffect indices, including after freezing when the moving fetus no longer mandates visual concentration. Several reasons may explain this. It has been previously suggested that operators may not receive enough training on the safety of ultrasound; or how to adjust the output level while keeping the same quality of the image [HAM11]. However, one could also hypothesize that operators do not look at the indices as they feel that safety is granted and it is unnecessary to spend time monitoring the safety bioeffect indices; [Mar05] there is some evidence from our study to suggest this is not unreasonable. Our results do suggest that bioeffect monitoring should be preliminarily done while using Doppler and this could be considered in future ultrasound safety guidelines. This is especially important while using pulsed-wave Doppler since in routine settings [KJM+20], the TIb was as high as 2.1, 2.6, and 3.0 in the first, second, and third-trimester, respectively.

In addition to aspects of operator awareness, there have been some reported concerns over the ability of the bioeffect indices to predict the intensity of ultrasound [Abr13, RKK+17]. This is because the bioeffect indices definitions have several weaknesses: TI and MI do not take time into account practical imaging factors like a long fluid path (full bladder, polyhydramnios) or obesity, and the reported outputs are not necessarily equivalent to those calculated in the laboratory [Abr13, RKK+17]. Our study was limited by data collection at one maternity unit, and with one set of equipment; further research may be required to determine if the results generalize to other settings. This would be especially important to consider in settings where transvaginal ultrasonography is commonly used in the first-trimester. Another potential limitation is the fact that the operators were aware of the eye-tracking component of the study which potentially could alter the operator behavior. This is in contrast to our findings in a recent study where we report that operators look at the biometric measurement values in over 90% of scans [DDC+20]. However, it is unlikely that operators behaved differently since eye-tracking is a passive measurement, operators do not have any visible indication to its function, and operators were not aware of the aims of the current analysis. Women included in this study had a mean BMI of 25.5 ± 5.5. In many regions of the world, the mean BMI is higher and the average times spent scanning in each modality may be higher. Lastly, the bioeffect safety indices values depend on the settings employed which depend on the mode selected, and these in-turn represent the default embedded during the machine setup. Nevertheless, this has probably not majorly altered our findings, as operators commonly optimize the machine settings in real-time. In the current analysis, we do not know whether machine settings adjustment was performed to improve the image quality or to optimize the machine energy output.

In conclusion, we have shown that in routine obstetric settings, safety indices are rarely looked at by operators. Despite this, the safety limits of ultrasound are adhered to. Our findings are reassuring since, despite years of concern, many

operators still fail to demonstrate good knowledge of the bioeffects of ultrasound [Mar05, SA08]. Nevertheless, due to the potential adverse effects, ultrasound should be performed by trained personnel who have received ultrasound safety education.

Bibliography

[Abr12] Jacques S. Abramowicz. Ultrasound and autism: Association, link, or coincidence? *J Ultrasound Med*, 31(8):1261–1269, August 2012.

[Abr13] Jacques S. Abramowicz. Benefits and risks of ultrasound in pregnancy. *Seminars in Perinatology*, 37(5):295–300, October 2013.

[AGD+06] Eugenius S. B. C. Ang, Vicko Gluncic, Alvaro Duque, Mark E. Schafer, and Pasko Rakic. Prenatal exposure to ultrasound waves impacts neuronal migration in mice. *PNAS*, 103(34):12903–12910, August 2006.

[AKM+03] J. S. Abramowicz, G. Kossoff, K. Marsal, G. Ter Haar, and International Society of Ultrasound in Obstetrics and Gynecology Bioeffects and Safety Committee. Executive Board of the International Society of Ultrasound in Obstetrics and Gynecology. Safety Statement, 2000 (reconfirmed 2003). International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). *Ultrasound Obstet Gynecol*, 21(1):100, January 2003.

[Ame16] American Institute of Ultrasound in Medicine. Official Statement: Recommended Maximum Scanning Times for Displayed Thermal Index (TI) Values. https://www.aium.org/officialStatements/65, 2016.

[ATG+17] S. Kate Alldred, Yemisi Takwoingi, Boliang Guo, Mary Pennant, Jonathan J. Deeks, James P. Neilson, and Zarko Alfirevic. First trimester ultrasound tests alone or in combination with first trimester serum tests for Down's syndrome screening. *Cochrane Database Syst Rev*, 3:CD012600, March 2017.

[Bry95] Marc Brysbaert. Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, 124(4):434–452, 1995.

[BSFT14] Bryann Bromley, Jean Spitz, Karin Fuchs, and Loralei L. Thornburg. Do clinical practitioners seeking credentialing for nuchal translucency measurement demonstrate compliance with biosafety recommendations? Experience of the Nuchal Translucency Quality Review Program. *J Ultrasound Med*, 33(7):1209–1214, July 2014.

[BTHZ+00] S. B. Barnett, G. R. Ter Haar, M. C. Ziskin, H. D. Rott, F. A. Duck, and K. Maeda. International recommendations and guidelines for the

safe use of diagnostic ultrasound in medicine. *Ultrasound Med Biol*, 26(3):355–366, March 2000.

[CSD+18] P. Chatelain, H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies. *IEEE Trans. Cybern.*, pages 1–11, 2018.

[DDC+20] L. Drukker, R. Droste, P. Chatelain, J. A. Noble, and A. T. Papageorghiou. Expected-value bias in routine third-trimester growth scans. *Ultrasound in Obstetrics & Gynecology*, 55(3):375–382, 2020.

[DL00] C. Deane and C. Lees. Doppler obstetric ultrasound: A graphical display of temporal changes in safety indices. *Ultrasound Obstet Gynecol*, 15(5):418–423, May 2000.

[HA17] Roxane Holt and Jacques S. Abramowicz. Quality and Safety of Obstetric Practices Using New Modalities- Ultrasound, MR, and CT. *Clin Obstet Gynecol*, 60(3):546–561, September 2017.

[HAM11] Laura E. Houston, Jenifer Allsworth, and George A. Macones. Ultrasound is safe... right?: Resident and maternal-fetal medicine fellow knowledge regarding obstetric ultrasound safety. *J Ultrasound Med*, 30(1):21–27, January 2011.

[HCSA+16] L. Höglund Carlsson, S. Saltvedt, B.-M. Anderlid, J. Westerlund, C. Gillberg, M. Westgren, and E. Fernell. Prenatal ultrasound and childhood autism: Long-term follow-up after a randomized controlled trial of first- vs second-trimester ultrasound. *Ultrasound Obstet Gynecol*, 48(3):285–288, September 2016.

[KJM+20] Christian Kollmann, Klaus-Vitold Jenderka, Carmel M. Moran, Ferdinando Draghi, J. F. Jimenez Diaz, and Ragnar Sande. EFSUMB Clinical Safety Statement for Diagnostic Ultrasound - (2019 revision). *Ultraschall Med*, 41(4):387–389, August 2020.

[Mar05] K. Marsál. The output display standard: Has it missed its target? *Ultrasound Obstet Gynecol*, 25(3):211–214, March 2005.

[NBO+15] Dragos Nemescu, Anca Berescu, Mircea Onofriescu, Dan Bogdan Navolan, and Cristian Rotariu. Safety Indices during Fetal Echocardiography at the Time of First-Trimester Scan Are Machine Dependent. *PLoS One*, 10(5):e0127570, 2015.

[NFAC09] Thomas R. Nelson, J. Brian Fowlkes, Jacques S. Abramowicz, and Charles C. Church. Ultrasound biosafety considerations for the practicing sonographer and sonologist. *J Ultrasound Med*, 28(2):139–150, February 2009.

[PUL] Perception Ultrasound by Learning Sonographic Experience, European Research Council (ERC) Advanced Grant. https://cordis.europa.eu/project/id/694581.

[RCH+16] A. Rempen, R. Chaoui, M. Häusler, K.-O. Kagan, P. Kozlowski, C. von Kaisenberg, and J. Wisser. Quality Requirements for Ultrasound Examination in Early Pregnancy (DEGUM Level I) between 4+0 and 13+6 Weeks of Gestation. *Ultraschall Med*, 37(6):579–583, December 2016.

[RKK+17] K. Retz, S. Kotopoulis, T. Kiserud, K. Matre, G. E. Eide, and R. Sande. Measured acoustic intensities for clinical diagnostic ultrasound transducers and correlation with thermal index. *Ultrasound Obstet Gynecol*, 50(2):236–241, August 2017.

[Roy13] Royal College of Obstetricians and Gynaecologists. The Investigation and Management of the Small–for–Gestational–Age Fetus (Green-top Guideline No. 31). https://www.rcog.org.uk/globalassets/documents/guidelines/gtg_31.pdf, 2013.

[SA08] Eyal Sheiner and Jacques S. Abramowicz. Clinical end users worldwide show poor knowledge regarding safety issues of ultrasound during pregnancy. *J Ultrasound Med*, 27(4):499–501, April 2008.

[SA09] Eyal Sheiner and Jacques S. Abramowicz. Acoustic output as measured by thermal and mechanical indices during fetal nuchal translucency ultrasound examinations. *Fetal Diagn Ther*, 25(1):8–10, 2009.

[Saf09] Safety Group of the British Medical Ultrasound Society. Guidelines for the safe use of diagnostic ultrasound equipment. https://www.bmus.org/static/uploads/resources/BMUS-Safety-Guidelines-2009-revision-FINAL-Nov-2009.pdf, 2009.

[Sal11] K. Å Salvesen. Ultrasound in pregnancy and non-right handedness: Meta-analysis of randomized trials. *Ultrasound Obstet Gynecol*, 38(3):267–271, September 2011.

[SIO+13] I. Sarris, C. Ioannou, E. O. Ohuma, D. G. Altman, L. Hoch, C. Cosgrove, S. Fathima, L. J. Salomon, and A. T. Papageorghiou. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics & Gynaecology*, 120(s2):33–37, 2013.

[SLA+11] K. Salvesen, C. Lees, J. Abramowicz, C. Brezinka, G. Ter Haar, K. Maršál, and Board of International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). ISUOG statement on the safe use of Doppler in the 11 to 13 +6-week fetal ultrasound examination. *Ultrasound Obstet Gynecol*, 37(6):628, June 2011.

[SMEK13] Ragnar K. Sande, Knut Matre, Geir E. Eide, and Torvid Kiserud. The effects of reducing the thermal index for bone from 1.0 to 0.5 and 0.1 on common obstetric pulsed wave Doppler measurements in the second half of pregnancy. *Acta Obstet Gynecol Scand*, 92(7):790–796, July 2013.

[tH12] G. ter Haar. *The Safe Use of Ultrasound in Medical Diagnosis*. British Institute of Radiology, London, 2012.

[Tom06] David A. Toms. The mechanical index, ultrasound practices, and the ALARA principle. *J Ultrasound Med*, 25(4):560–561; author reply 561–562, April 2006.

[TVM+09] M. R. Torloni, N. Vedmedovska, M. Merialdi, A. P. Betrán, T. Allen, R. González, L. D. Platt, and ISUOG-WHO Fetal Growth Study Group. Safety of ultrasonography in pregnancy: WHO systematic review of the literature and meta-analysis. *Ultrasound Obstet Gynecol*, 33(5):599–608, May 2009.

[vCH+16] C. von Kaisenberg, R. Chaoui, M. Häusler, K. O. Kagan, P. Kozlowski, E. Merz, A. Rempen, H. Steiner, S. Tercanli, J. Wisser, and K.-S. Heling. Quality Requirements for the early Fetal Ultrasound Assessment at 11-13+6 Weeks of Gestation (DEGUM Levels II and III). *Ultraschall Med*, 37(3):297–302, June 2016.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Safety Indices of Ultrasound: Adherence to Recommendations and Awareness During Routine Obstetric Ultrasound Scanning
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Lior Drukker, Richard Droste, Pierre Chatelain, J. Alison Noble and Aris T. Papageorghiou. "Safety Indices of Ultrasound: Adherence to Recommendations and Awareness During Routine Obstetric Ultrasound Scanning". European Journal of Ultrasound, vol. 41, no. 2, pp. 138–145, 2020.

Student Confirmation

Student Name:	Richard Droste		
Contribution to the Paper	Co-lead author, responsible for designing and performing the data extraction, data analyses, visualizations and the majority of statistics of the paper. Contributed to the original draft.		
Signature	Richard Droste	Date	31.05.2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. J. Alison Noble			
I confirm that the description above is accurate.			
Signature	J Alison Noble	Date	01.06.2021

This completed form should be included in the thesis, at the end of the relevant chapter.

4

Visual Saliency Modeling for Image and Video Data

Contents

4.1 Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction

Authors. Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Aris T. Papageorghiou, and J. Alison Noble

Conference. *Medical Image Understanding and Analysis (MIUA)*, pp. 174–186, 2019.

Background. Beyond the statistical analysis of gaze patterns presented in the previous chapter, it is possible to predict sonographer gaze on unseen data via visual saliency prediction. In short, a model, typically a convolutional neural network, is trained to recognize the image patterns that attract the visual attention of sonographers. This allows to identify salient landmarks on new ultrasound data and yields learned features that are useful for a variety of image analysis tasks, as we will show in the next chapter. Here, we present the first visual saliency model for ultrasound video data and attain high accuracy through a new, cognition-inspired architecture.

Statement of Authorship. I was the lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft. J. Alison Noble was the main responsible for supervision and funding acquisition and contributed to conceptualization, methodology and editing the draft. Aris T. Papageorghiou contributed to supervision, funding acquisition and editing the draft. Yifan Cai, Harshita Sharma and Pierre Chatelain contributed to conceptualization, methodology and editing the draft.

Recognition. The paper won the MIUA 2019 Best Paper Award.

Abstract

For visual tasks like ultrasound (US) scanning, experts direct their gaze towards regions of task-relevant information. Therefore, learning to predict the gaze of sonographers on US videos captures the spatio-temporal patterns that are important for US scanning. The spatial distribution of gaze points on video frames can be represented through heat maps termed saliency maps. Here, we propose a temporally bidirectional model for video saliency prediction (BDS-Net), drawing inspiration from modern theories of human cognition. The model consists of a convolutional neural network (CNN) encoder followed by a bidirectional gated-recurrent-unit recurrent convolutional network (GRU-RCN) decoder. The temporal bidirectionality mimics human cognition, which simultaneously reacts to past and predicts future sensory inputs. We train the BDS-Net alongside spatial and temporally one-directional comparative models on the task of predicting saliency in videos of US abdominal circumference plane detection. The BDS-Net outperforms the comparative models on four out of five saliency metrics. We present a qualitative analysis on representative examples to explain the model's superior performance.

4.1.1 Introduction

Recently, it has been demonstrated that sonographer gaze-tracking can aid standard plane detection in fetal ultrasound (US) imaging. Cai et al. [CSCN18a] proposed the SonoEyeNet model for abdominal circumference plane (ACP) detection. Recorded gaze-tracking heat maps—hereafter referred to as *saliency maps*—are used as attention on feature maps which are extracted with a fine-tuned SonoNet model [BKM+17]. Next, Cai et al. [CSCN18b] proposed the Multi-task SonoEyeNet. Instead of relying on gaze data as an input, an attention module learns to predict saliency maps so that no gaze-tracking data is required for inference. Recently, Droste et al. [DCS+19] demonstrated that a saliency predictor trained entirely without manual annotations can be transferred to perform standard plane detection in routine clinical videos. These models perform standard plane detection and saliency prediction on single-frames only. However, ultrasound data and human eye movements are inherently spatio-temporal signals.

In this work we aim at improving ultrasound saliency prediction through spatio-temporal modeling, i.e. video saliency prediction. Therefore, we aim to bridge the gap between existing spatio-temporal models which do not leverage gaze information, e.g. for fetal cardiology [HBNZ17, GN17] or US video partitioning [SDC+19], and models like SonoEyeNet that do not utilize temporal information. Chaabouni et al. [CBH16] present an early convolutional neural network (CNN) based approach for video saliency prediction, adding optical flow as an additional input channel to a single-frame CNN. Bak et al. [BKEE18] propose to include optical flow via a two-stream architecture [SZ14]. Bazzani et al. [BLT17] achieve much larger temporal depth with a recurrent mixture density network by aggregating feature vectors with a long short-term memory (LSTM) model. Wang et al. [WSG+18] recently proposed a large video saliency benchmark (DHF1K) and show that existing video saliency predictors do not outperform the best single-frame saliency predictors.

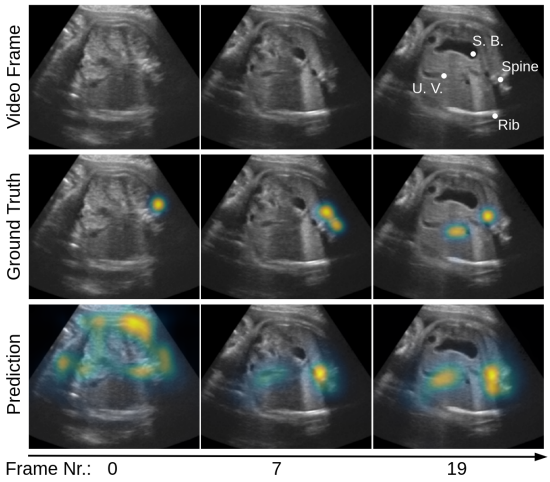

Figure 4.1: BDS-Net predicted saliency maps. The key structures are marked. U. V. denotes umbilical vein and S. B. stomach bubble.

In contrast, Wang et al. achieve state-of-the-art on their benchmark with an architecture consisting of a CNN encoder and a convolutional LSTM decoder.

The above mentioned spatio-temporal models predict the saliency map of each video frame based on aggregated information of the previous frames. However, research in cognitive science suggests that human perception is not just reacting to past and present stimuli. Clark [Cla13] argues that the brain is a 'prediction machine' that strives to minimize the prediction error between expectations versus sensory inputs. We transfer this insight to the problem of predicting sonographer visual saliency on US videos. Since the sonographer's expectations about future visual stimuli are unknown, we use future video frames as a proxy thereof, and ask the question: *To what extent are future video frames predictive for visual saliency?* Song et al. [SWZ+18] recently proposed a bidirectional model for video salient object detection, which is a related application but aims at detecting and segmenting the most salient object in a scene rather than predicting the actual distribution of gaze points. Here, we propose an architecture combining a CNN and a temporally bidirectional recurrent neural network, BDS-Net, that predicts the visual saliency of each frame based on information of the entire video sequence,

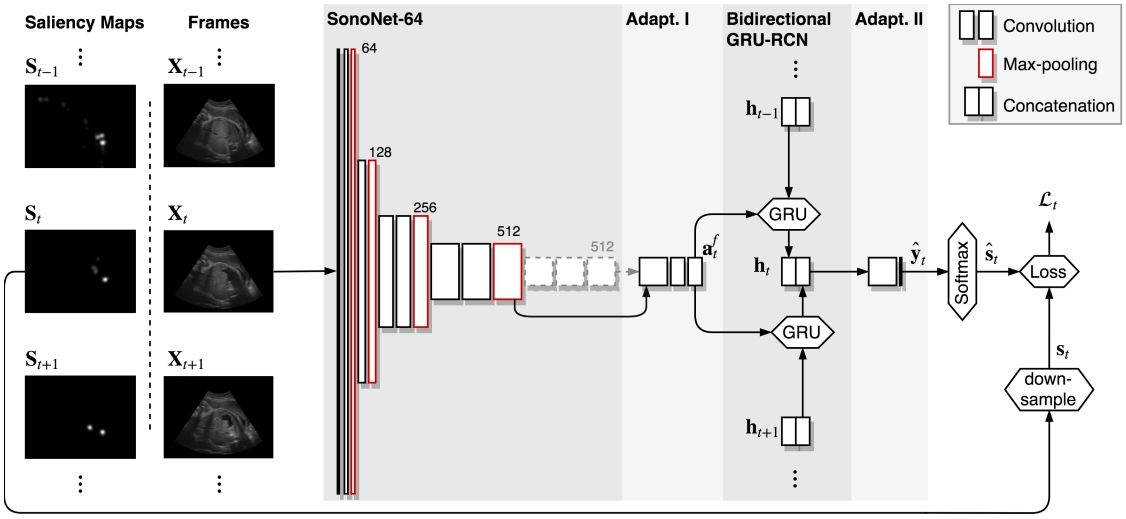

Figure 4.2: Schema of the BDS-Net architecture and training procedure. For better readability, activation and normalization layers are not explicitly shown. The dashed part of the SonoNet-64 is only used for an ablation study.

and compare the performance of the BDS-Net to an equivalent one-directional and a purely spatial model.

Contributions. The contributions of this study are three-fold: (1) To the best of our knowledge, this is the first study to propose a temporally bidirectional model for video saliency prediction, both in medical imaging and in computer vision more generally. Since the model considers the entire video sequence for saliency prediction of each frame, this approach is fundamentally different from previous models that only consider past and present frames. (2) We demonstrate that it is possible to train an effective video saliency predictor with few more than one hundred sequences, despite high inter-sequence variance. We achieve high data-efficiency by employing effective transfer learning and regularization techniques, and by reducing model complexity where possible, e.g. using a gated-recurrent-unit recurrent convolutional network (GRU-RCN) instead of a convolutional LSTM. (3) We demonstrate that the trained US video saliency predictor has learned meaningful aspects of sonographers' cognition in selecting the ACP. Therefore, we expect the model to be beneficial as part of architectures such as the Multi-task SonoEyeNet [CSCN18b].

4.1.2 BDS-Net

The BDS-Net architecture consists of a *truncated SonoNet-64* model as frame-wise encoder, *adaptation I* that extracts the task-relevant features, a *bidirectional GRU-RCN* to aggregate the features temporally, and *adaptation II* to assemble the saliency map, followed by a *softmax* function (Fig. 4.2). In the following, we will use the vector notation $\mathbf{v}_t = [v_0^t, v_1^t, ..., v_n^t]^\top$.

Truncated SonoNet-64.

The SonoNet model was recently proposed for US standard plane detection [BKM+17]. It is derived from the VGG-16 architecture [SZ15], removing the final max-pooling and replacing the fully-connected layers with adaptation layers of 1×1-convolutions followed by global average pooling. Also, batch normalization [IS15] is added to each convolutional layer. The authors present three model variants with different numbers of convolutional kernels and train them on over 27 thousand US standard plane images. For this work, we use the largest variant, SonoNet-64, which was shown to achieve the highest overall precision. To use the model as a feature extractor, we remove the adaptation layers since they are classification-task specific. Further, we truncate the model by discarding the final three 3×3-convolutional layers to obtain lower-level features. We use the remaining 10 convolutional layers as frame-wise encoder of the BDS-Net. Since the SonoNet training data is substantially larger than the data available for this work, we use the model with fixed pre-trained weights.

Bidirectional GRU-RCN.

We propose a bidirectional gated-recurrent-unit recurrent convolutional network (GRU-RCN) as the spatio-temporal decoder of the network. GRU networks [CvG+14] mitigate the exploding/vanishing gradient problem of a regular RNN similarly to LSTM networks [HS97] by updating the hidden state through element-

wise additive and multiplicative gates instead of matrix multiplications. Compared to LSTM networks, however, they yield faster training convergence and higher accuracy on tasks like video captioning, despite reduced complexity and fewer learned parameters [CGCB14]. While the standard GRU operates on 1D feature vectors, the GRU-RCN [BYPC16] is a straightforward extension for stacked 2D feature maps, replacing matrix products with convolutions. This modification vastly reduces the number of parameters compared to a fully-connected GRU and preserves the spatial feature topology. The bidirectional GRU-RCN is constructed from two separate GRU-RCN instances that propagate their hidden states forwards and backwards through time, respectively.

In the forward GRU-RCN, denoted by $\cdot^{\rightarrow}$, the candidate activation $\widetilde{\mathbf{h}}_t^{\rightarrow}$ at time t is computed from the feature activations $\mathbf{a}_t^f$ and the previous hidden state $\mathbf{h}_{t-1}^{\rightarrow}$ as:

$$\mathbf{r}_t^{\rightarrow} = \sigma(\mathbf{W}_r^{\rightarrow} * [\mathbf{h}_{t-1}^{\rightarrow} | \, \mathbf{a}_t^f] + \mathbf{b}_r^{\rightarrow}) \tag{4.1.2.1}$$

$$\widetilde{\mathbf{h}}_t^{\rightarrow} = \tanh\left(\mathbf{W}_h^{\rightarrow} * [\mathbf{r}_t^{\rightarrow} \circ \mathbf{h}_{t-1}^{\rightarrow} \, | \, \mathbf{a}_t^f] + \mathbf{b}_h^{\rightarrow}\right) , \tag{4.1.2.2}$$

where $\mathbf{r}_t^{\rightarrow}$ is the reset gate, $\mathbf{W}_{\cdot}^{\rightarrow}$ and $\mathbf{b}_{\cdot}^{\rightarrow}$ are the respective convolutional filters and biases, $\sigma(\cdot)$ is the logistic sigmoid function, $*$ is the convolution operator, $\circ$ denotes element-wise multiplication and $[\cdot|\cdot]$ denotes concatenation along the feature dimension. The reset gate controls the propagation of the previous hidden state into the current hidden state. Next, the activation $\mathbf{h}_t^{\rightarrow}$ is computed as a linear interpolation between the previous activation and the candidate activation, modulated by the update gate $\mathbf{z}_t^{\rightarrow}$:

$$\mathbf{z}_t^{\rightarrow} = \sigma(\mathbf{W}_z^{\rightarrow} * [\mathbf{h}_{t-1}^{\rightarrow} | \, \mathbf{a}_t^f] + \mathbf{b}_z^{\rightarrow}) \tag{4.1.2.3}$$

$$\mathbf{h}_t^{\rightarrow} = (1 - \mathbf{z}_t^{\rightarrow}) \circ \mathbf{h}_{t-1}^{\rightarrow} + \mathbf{z}_t^{\rightarrow} \circ \widetilde{\mathbf{h}}_t^{\rightarrow} . \tag{4.1.2.4}$$

The backward GRU-RCN activation $\mathbf{h}_t^{\leftarrow}$ is computed equivalently, using the activa-

tion $\mathbf{h}^{\leftarrow}_{t+1}$ of time $t+1$ as the previous activation. Finally, the activations of the forward and backward RCNs are concatenated as $\mathbf{h}_t = [\mathbf{h}^{\rightarrow}_t | \mathbf{h}^{\leftarrow}_t]$.

We normalize the activations and gates throughout the GRU with layer normalization [BKH16]. We found no increase in performance by replacing the layer normalization with instance normalization [UVL16] in the GRU. Batch normalization is not compatible since we set the batch size to one due to the high variability of the sequence lengths. Further, we observed no improvement through dropout on the GRU inputs [ZSV14] or variational recurrent dropout [GG16]. To avoid over-fitting, the kernel size of $\mathbf{W}_r$ and $\mathbf{W}_z$ are set to size 1×1. Only the kernels of $\mathbf{W}_h$ for computing the candidate activation are set to size 3×3.

Adaptations I & II.

The first set of adaptation layers reduces the feature length of the SonoNet output. Since the SonoNet features are learned on several fetal anatomies, only a subset of the SonoNet features is likely to be relevant for the fetal abdomen. A convolutional layer of dimension 1×1×128×512 (2D kernel size × output dimension × input dimension) reduces the feature length, followed by a layer of dimension 3×3×128×128 to adapt the feature maps. Layer normalization [BKH16] is performed after each layer. The final adaptation layer, a single convolutional layer of dimension 1×1×1×128, assembles the final saliency map.

Loss function.

At each time step $t \in \{0, 1, ..., T\}$ and output pixel $i \in \{0, 1, ..., P\}$, we obtain the predicted saliency $\hat{s}^t_i$ from the activations $\hat{y}^t_i$ of the final adaptation layer via the softmax function $\hat{s}^t_i = e^{\hat{y}^t_i} (\sum_{j=0}^{P} e^{\hat{y}^t_j})^{-1}$. Consequently, we implicitly treat the saliency maps as generalized Bernoulli distributions of fixations over the pixels of each frame [JMV16]. We compute the loss as the sum of the Kullback-Leibler

Table 4.1: Data preparation procedure.

Data preparation step	Output
1) Discarding of irrelevant frames	ACP sweeps
For each sweep:	
2) ACP selection by sonographers	ACP search sequences
For each ACP search sequence:	
3) Gaze point aggregation	Gaze maps
4) Gaze map filtering	Saliency maps

Divergences between the predicted distributions[1] and the downscaled ground truth distributions $\mathbf{s}_t$ as $\mathcal{L} = \sum_{t=0}^{T} \sum_{i=0}^{P} s_i^t \cdot (log(s_i^t) - log(\hat{s}_i^t))$

4.1.3 Experiments

Data

The gaze data for this study had previously been recorded based on 33 fetal US videos, which were acquired according to a manual US sweep protocol, moving the probe from the bottom to the top of pregnant women's abdomen. Table 4.1 summarizes the data preparation procedure. *(1) Discarding of irrelevant frames:* From each of the 33 full sweeps an *ACP sweep* was extracted by discarding the frames which do not show the fetal abdomen. *(2) ACP selection by sonographers:* Each ACP sweep was presented to eight sonographers independently with the task of selecting the abdominal circumference plane (ACP). The sonographers were able to scroll through the frames using a keyboard until deciding on one ACP frame. The gaze of the sonographers was recorded at 30 Hz using an eye tracker (The EyeTribe) placed beneath the screen. In addition, the current sweep frame was registered for each gaze point. This yielded $33 \times 8 = 264$ sequences of gaze point-sweep frame pairs, which we will refer to as *ACP search sequences*. The ACP

[1] In our implementation, for numerical stability, we compute $log(\hat{s}_i^t)$ with a log-softmax function instead of computing the softmax and logarithm sequentially.

search sequences represent the way the sonographers moved through the frames to find the ACP. Therefore, they are a potentially useful approximation of freehand US video sequences where the sonographers find the ACP by moving the probe. The sequences were manually inspected and recordings with low-quality gaze data or miscalibration were discarded, leaving 116 sequences for further processing. *(3) Gaze point aggregation:* Since we want our model to learn the search strategy of sonographers from the first glance until the final ACP plane decision, we want to train the model on entire ACP search sequences. At 30 Hz, however, the sequences are too long (at least several hundred frames) and contain high redundancy among consecutive frames. Therefore, the gaze points were aggregated over intervals of 1000 ms at every 8^{th} gaze sampling time, reducing the sampling rate from 30 Hz to 3.75 Hz. Gaze points outside the US image fan were discarded. A gaze map was computed for each aggregated set of gaze points by setting each pixel value to its corresponding number of gaze points. The frames were re-sampled at the same sampling times. The resulting sequences of gaze map-sweep frame pairs are of length 13 to 147 (avg. 33.6). *(4) Gaze maps filtering:* The saliency maps were computed by smoothing the gaze maps with a Gaussian kernel. The resulting final sequences of saliency map-sweep frame pairs are henceforth referred to as *saliency sequences*. The ACP sweeps were divided into 30 sweeps for training and 3 sweeps for validation with five-fold cross-validation.

Implementation Details

Preprocessing. The frames are preprocessed following Baumgartner et al. [BKM+17]. Data augmentation is performed by random rotation with an angle uniformly sampled from $[-25, 25]$ degrees and random horizontal flipping. Scale augmentation is omitted since it results in cropping of parts of the fetal abdomen. Next, the frames are normalized to zero-mean and unit-variance, multiplied by 255 and resized to 288×224 px. For the calculation of the loss, the ground truth saliency maps are

transformed analogously and resized to 18×14 px, which is the output dimension of the network for inputs of 288×224 px.

Training. The model is trained over 140 epochs via stochastic gradient descent (SGD) with Nesterov momentum of 0.9 and initial learning rate 0.004. In accordance with Keskar et al. [KS17], we find that SGD yields better generalization to the validation set compared to ADAM. The learning rate is decayed by a factor of 0.5 each time the validation loss stagnates. The batch size is one to allow for varying sequence lengths. Sequences longer than 60 frames are truncated. The model is regularized via weight decay of $1 \cdot 10^{-6}$, dropout with rate 0.2 before the second and the last adaptation layers, as well as clipping the gradients outside the interval $[-5, 5]$. The z-gate bias is initialized to 1 to stabilize training by learning the spatial features first.

Comparative models. Two comparative models are implemented: A one-directional GRU-RCN model and a purely spatial, single-frame model. The one-directional model is constructed by removing the backward GRU-RCN module from the BDS-Net. All other architectural and training parameters are identical. For the spatial model, the bidirectional GRU-RCN is simply replaced by an additional convolutional layer of dimension 3×3×128×128. Moreover, the layer normalization modules are removed and batch normalization is added to each layer. Training is performed on batches of 16 randomly selected frames and dropout with rate 0.5 is added to all layers. The initial learning rate is increased to 0.01 and no weight-decay is performed. Furthermore, we perform an ablation study to examine the effect of using the full SonoNet-64 (only adaptation layers removed) or the truncated SonoNet-32 models instead of the truncated SonoNet-64. The results are presented in Sec. 4.1.4.

Table 4.2: Cross-validation scores (mean ± standard deviation) of the BDS-Net and the spatial and one-directional models. The best scores are marked in bold. The superscripts * and † denote an improvement with $p < 0.05$ over the spatial and one-directional models, respectively.

Model	NSS ↑	AUC-J ↑	KLD ↓	CC ↑	SIM ↑
Spatial	1.40 ±0.36	0.83 ±0.04	3.01 ±0.37	0.26 ±0.06	**0.23** ±0.04
One-directional	1.49 ±0.34	0.85 ±0.04 *	2.22 ±0.25 *	0.27 ±0.06	0.21 ±0.03
BDS-Net	**1.61** ±0.33 *†	**0.87** ±0.03 *†	**2.16** ±0.27 *†	**0.29** ±0.06 *†	**0.23** ±0.04 †

Evaluation Metrics

We evaluate the models on the metrics of the MIT Saliency Benchmark [BJB+12]. For this, we ported the MATLAB code published by the authors[2] to Python. We consider the fixation point (gaze map) based metrics Normalized Scanpath Saliency (NSS) and Area Under the ROC Curve by Judd (AUC-J), as well as the distribution (saliency map) based metrics Kullback-Leibler divergence (KLD), Linear Correlation Coefficient (CC) and Similarity (SIM). AUC-J, KLD and SIM are more sensitive to false negatives than to false positives, while NSS and CC treat them symmetrically [BJO+19]. To compute the average scores on each validation set, the scores are first averaged across time for each sequence and then across sequences. Thus, shorter and longer sequences weigh equally in the average. The differences between the respective cross-validated model scores are tested for statistical significance with the Wilcoxon test.

4.1.4 Results

Quantitative Results

Table 4.2 shows the validation scores of the BDS-Net and comparative models. The BDS-Net receives the best scores on all metrics except SIM. Moreover, both spatio-temporal models perform better on average than the spatial model on all

[2] https://github.com/cvzoya/saliency

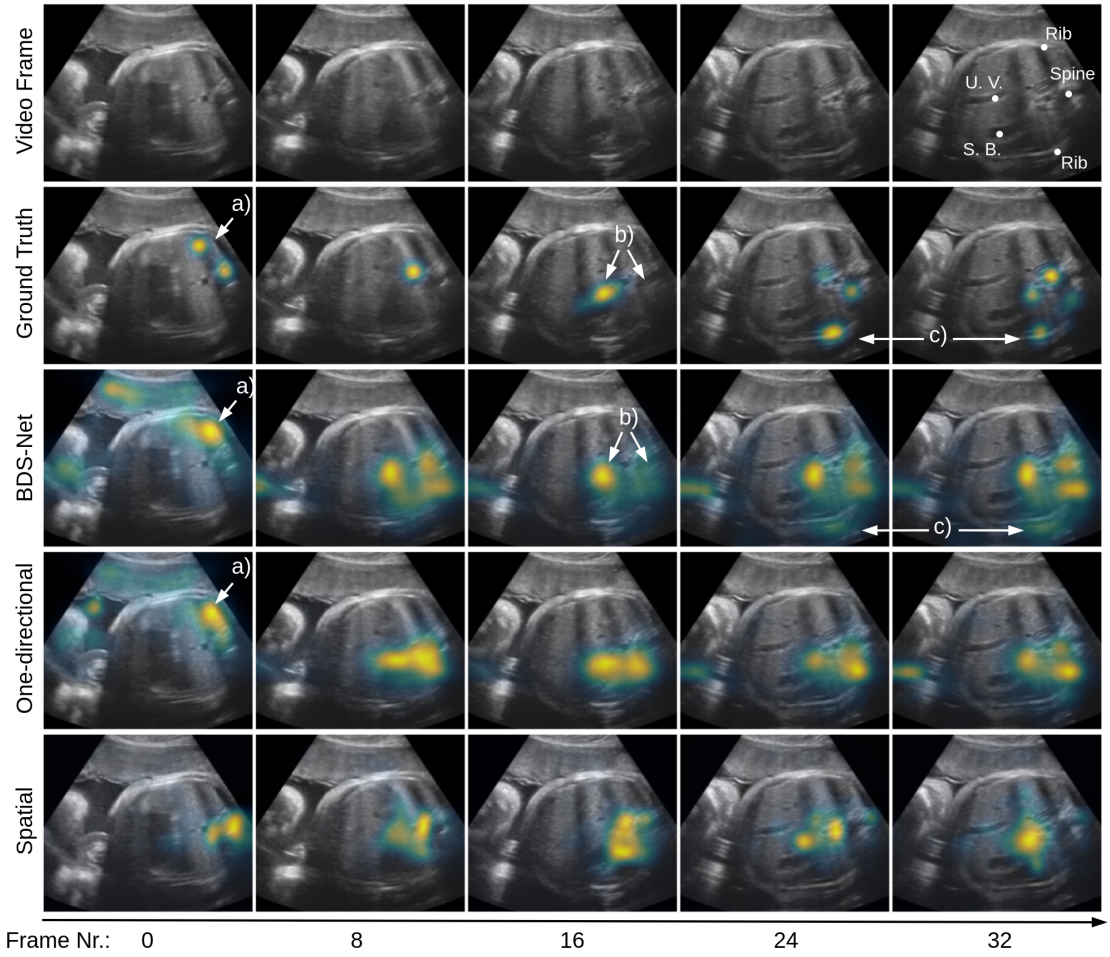

Figure 4.3: Five frames from an exemplary ACP search sequence. The rows show the input frames, the ground truth saliency annotations, and the saliency predictions of the BDS-Net and the spatial and one-directional models, respectively. The relevant anatomical structures are denoted in the last input frame (top right).

metrics except SIM. For SIM, the BDS-Net scores better than the one-directional model but is on par with the spatial model.

Representative Examples

Figures 4.1 and 4.3 show examples of the predictions of the BDS-Net model and the comparative models on validation data. Since training and validation data were divided scan-wise, the frames are entirely unseen by the networks. Moreover, the networks are agnostic as to which sonographer is observing the scan.

Fig. 4.1 shows input frames, ground truth saliency maps and BDS-Net predictions

for three representative frames of one exemplary sequence. At frame zero, the prediction is highly uncertain. Areas throughout the middle and the upper boundary of the abdomen are predicted as fixation candidates. The highest probability is assigned to the area around the upper rib, which is not well visible. The ground truth fixation is between the spine and the upper rib. At frame seven, the BDS-Net assigns high saliency values to the spine and lower values to the umbilical vein. The ground truth fixations are at the spine. At frame nineteen, near the end of the sequence, the network assigns approximately equal probabilities to umbilical vein and spine. The ground truth fixations are indeed on umbilical vein and spine.

Fig. 4.3 shows a more detailed example of BDS-Net predictions for five representative frames of another exemplary sequence. Additionally, the predictions of the spatial and one-directional models are shown. At frame zero, both spatio-temporal models predict uncertain saliency maps with a spread-out peak between spine and upper rib as denoted with *a)* in the figure. The spatial model, which does not have information about the position of the frame in the sequence, predicts saliency at the spine with high certainty. The ground truth fixations are both at spine and the upper rib. Over the next frames, the recurrent models predict temporally smooth saliency maps with slightly varying maxima around the center of the abdomen and the spine. The maxima of the spatial model fall into the same regions but the maps are less temporally smooth. Two key advantages of the BDS-Net predictions are denoted with *b)* and *c)*. At frame sixteen, in the middle of the sequence, the sonographer fixates the center of the abdomen. The spatial model predicts fixations around the spine and the one-directional model predicts fixations at either the spine or the center. Only the bidirectional model, which has information about both the previous and the subsequent frames, correctly predicts the fixation at the center and omits the spine, as denoted by *b)*. On frames 24 and 32, towards the end of the search sequence, the sonographer fixates on the spine, the center of the abdomen and the lower rib, which is not well visible in these frames. Only the BDS-Net

Table 4.3: Scores for the ablation study of the feature extractor on one randomly chosen validation set.

Feature Extractor	NSS ↑	AUC-J ↑	KLD ↓	CC ↑	SIM ↑
Full SonoNet-64	1.50	0.86	2.21	0.29	0.22
Truncated SonoNet-32	2.01	0.90	2.01	0.37	0.25
Truncated SonoNet-64	**2.20**	**0.91**	**1.78**	**0.39**	**0.27**

correctly assigns probability of fixation to the lower rib, indicated by *c)*.

Ablation Study

The quantitative results for the ablation study of the feature extractor are shown in Table 4.3. The ranking of the models is consistent across all five metrics: The full SonoNet-64 with higher-level features performs least favorable, the smaller truncated SonoNet-32 ranks second and the truncated SonoNet-64 performs best.

4.1.5 Discussion

The BDS-Net outperforms both comparative models on the AUC-J and NSS metrics, which are the default metrics of the MIT Saliency Benchmark [BJB+12]. Moreover, the one-directional model outperforms or matches the score of the spatial model on those metrics, despite the fact that training the spatial model is arguably easier for several reasons. First, gradient steps can be computed for each of the 3901 saliency maps per epoch separately. For the recurrent models, in contrast, gradient steps can only be computed for the 104 sequences per epoch. Second, the gradients are noisier for recurrent models in general. We mitigate this problem through gradient clipping, but it is an ad-hoc solution, and it does not resolve vanishing gradients. Finally, batch normalization is applied in the spatial model. For the recurrent models, since we set the batch size to one to account for the varying sequence lengths, we revert to layer normalization, which is known to stabilize training less for most tasks [WH18]. The fact that the recurrent models perform better despite the more difficult training

conditions is a strong indication that spatial information (the current frame) alone is not sufficient to predict US video saliency accurately. This is in accordance with the results of Wang et al. [WSG+18] who have shown for natural videos that a recurrent architecture can outperform sophisticated single-frame saliency predictors.

Moreover, we have shown quantitatively and qualitatively that the bidirectional model performs better than the one-directional model. The added backwards GRU-RCN is the only difference between the two architectures, i.e. no other layers were added or removed and the training procedures are identical. This supports our hypothesis that sonographers are implicitly predicting future frames and focus their visual attention accordingly. Since predicting future frames requires domain expertise, we see this approach as a step towards modeling sonographer experience.

It is difficult to say how much room for improvement remains for the given dataset. Naturally, there is a certain inter-observer variability in the ACP search strategies. The model can only learn to predict the saliency corresponding to some average search strategy across all sonographers. To compute the actual maximum saliency scores, the inter-observer congruence (IOC) would need to be quantified. In our case, however, this is particularly difficult since each gaze sequence corresponds to a unique frame-sequence controlled by the sonographer. Therefore, there is no common reference frame for comparing the gaze sequences.

Nonetheless, despite the missing reference values for the quantitative evaluations, the qualitative analyses have shown that the BDS-Net has learned meaningful spatio-temporal patterns in the sonographers' search strategies. The analyses of Cai et al. [CSCN18b] have shown that similar learned experience can significantly improve US standard plane detection on single frames. We expect that our video saliency predictor can further improve the performance of such models.

4.1.6 Conclusion and Outlook

We have presented a new model for predicting video saliency during ACP plane selection. We have shown that the temporally bidirectional BDS-Net model predicts saliency more accurately than single-frame and one-directional comparative models. The model has learned meaningful spatio-temporal patterns that attract sonographers' attention. Therefore, we expect the model to be beneficial for US standard plane detection tasks. In future work we will transfer the model to a larger dataset, which is currently being acquired. This will allow us to explore the limits of this approach for learning sonographic experience. Furthermore, we plan to integrate the model into architectures for US image analysis tasks such as standard plane detection and video partitioning.

Bibliography

[BJB+12] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT Saliency Benchmark. http://saliency.mit.edu/, 2012.

[BJO+19] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757, 2019.

[BKEE18] Cagdas Bak, Aysun Kocak, Erkut Erdem, and Aykut Erdem. Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction. *IEEE Trans. Multimed.*, 20(7):1688–1698, 2018.

[BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *NIPS - Deep Learning Symposium*, 2016.

[BKM+17] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. *IEEE Trans. Med. Imag.*, 36(11):2204–2215, 2017.

[BLT17] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani. Recurrent Mixture Density Network for Spatiotemporal Visual Attention. *ICLR*, 2017.

[BYPC16] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving Deeper into Convolutional Networks for Learning Video Representations. *ICLR*, 2016.

[CBH16] Souad Chaabouni, Jenny Benois-pineau, and Ofer Hadar. Deep Learning for Saliency Prediction in Natural Video. *arXiv:1604.08010*, 2016.

[CGCB14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *NIPS*, 2014.

[Cla13] Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03):181–204, 2013.

[CSCN18a] Y. Cai, H. Sharma, P. Chatelain, and J. A. Noble. SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking. In *IEEE International Symposium on Biomedical Imaging*, 2018.

[CSCN18b] Yifan Cai, Harshita Sharma, Pierre Chatelain, and J. Alison Noble. Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[CvG+14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *EMNLP*, 2014.

[DCS+19] Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention. In *Information Processing in Medical Imaging*, volume 11492, pages 592–604. 2019.

[GG16] Yarin Gal and Zoubin Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *NIPS*, 2016.

[GN17] Yuan Gao and J Alison Noble. Detection and Characterization of the Fetal Heartbeat in Free-hand Ultrasound Sweeps with Weakly-supervised Two-streams Convolutional Networks. *MICCAI*, 2017.

[HBNZ17] Weilin Huang, Christopher P. Bridge, J. Alison Noble, and Andrew Zisserman. Temporal HeartNet: Towards Human-Level Automatic Analysis of Fetal Cardiac Screening Video. *MICCAI*, 2017.

[HS97] Sepp Hochreiter and Jurgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[IS15] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ICML*, 2015.

[JMV16] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-End Saliency Mapping via Probability Distribution Prediction. *CVPR*, 2016.

[KS17] Nitish Shirish Keskar and Richard Socher. Improving Generalization Performance by Switching from Adam to SGD. *arXiv:1712.07628*, 2017.

[SDC+19] H Sharma, R Droste, P Chatelain, L Drukker, A Papageorghiou, and J Alison Noble. Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans. In *IEEE ISBI*, 2019.

[SWZ+18] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection. *ECCV*, 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. *NIPS*, 2014.

[SZ15] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arxiv:1607.08022*, 2016.

[WH18] Yuxin Wu and Kaiming He. Group Normalization. *ECCV*, 2018.

[WSG+18] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting Video Saliency: A Large-scale Benchmark and a New Model. *CVPR*, 2018.

[ZSV14] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent Neural Network Regularization. *arXiv:1409.2329*, 2014.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Aris T. Papageorghiou, and J. Alison Noble. "Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction". In: Medical Image Understanding and Analysis (MIUA), pp. 174–186, 2019.

Student Confirmation

Student Name:	Richard Droste		
Contribution to the Paper	Lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft.		
Signature	Richard Droste	Date	31.05.2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. J. Alison Noble			
I confirm that the description above is accurate.			
Signature	J Alison Noble	Date	01.06.2021

This completed form should be included in the thesis, at the end of the relevant chapter.

4.2 Unified Image and Video Saliency Modeling

Authors. Richard Droste*, Jianbo Jiao*, J. Alison Noble

*equal contribution

Conference. *European Conference on Computer Vision (ECCV)* 2020.

Background. During ultrasound acquisition, sonographers constantly switch between live scanning (video) and frozen ultrasound frames (images) and our interest is to model visual saliency accurately for both image and video data. Many models have previously been proposed to predict saliency for either of these modalities, but none for both. In this paper, we develop a unified image and video saliency model and find that this results in a simpler model with better saliency prediction accuracy for both modalities. Moreover, our proposed model is the fastest and most lightweight among all existing deep learning based models, which will be important for the deployment during ultrasound scanning (see chapter 7). We evaluate the model not on ultrasound data, but on publicly available datasets of everyday scenes in order to benchmark it against prior work.

Statement of Authorship. I was the co-lead author responsible for conceiving the original idea. Together with Jianbo Jiao I was responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft. J. Alison Noble was the main responsible for supervision and funding acquisition and contributed to conceptualization, methodology and editing the draft.

Recognition. The paper was selected for a Spotlight Presentation (top 5% of all ECCV 2020 papers).

Notes. Table 4.5 was added in this thesis for clarity.

Abstract

Visual saliency modeling for images and videos is treated as two independent tasks in recent computer vision literature. While image saliency modeling is a well-studied problem and progress on benchmarks like SALICON and MIT300 is slowing, video saliency models have shown rapid gains on the recent DHF1K benchmark. Here, we take a step back and ask: Can image and video saliency modeling be approached via a unified model, with mutual benefit? We identify different sources of domain shift between image and video saliency data and between different video saliency datasets as a key challenge for effective joint modelling. To address this we propose four novel domain adaptation techniques— Domain-Adaptive Priors, Domain-Adaptive Fusion, Domain-Adaptive Smoothing and Bypass-RNN— in addition to an improved formulation of learned Gaussian priors. We integrate these techniques into a simple and lightweight encoder-RNN-decoder-style network, UNISAL, and train it jointly with image and video saliency data. We evaluate our method on the video saliency datasets DHF1K, Hollywood-2 and UCF-Sports, and the image saliency datasets SALICON and MIT300. With one set of parameters, UNISAL achieves state-of-the-art performance on all video saliency datasets and is on par with the state-of-the-art for image saliency datasets, despite faster runtime and a 5 to 20-fold smaller model size compared to all competing deep methods. We provide retrospective analyses and ablation studies which confirm the importance of the domain shift modeling. The code is available at https://github.com/rdroste/unisal.

4.2.1 Introduction

When processing static scenes (images) and dynamic scenes (videos), humans direct their visual attention towards important information, which can be measured by recording eye fixations. The task of predicting the fixation distribution is referred to as *(visual) saliency prediction/modeling*, and the predicted distributions as *saliency maps*. Convolutional neural networks (CNNs) have emerged as the most performant technique for saliency modeling due to their capacity to learn complex feature hierarchies from large-scale datasets [Bor18, JHDZ15].

While most prior work focuses on image data, interest in video saliency modeling was recently accelerated through ACLNet, a dynamic saliency model that outperforms static models on the large-scale, diverse DHF1K benchmark [WSG+18]. However, as methods for video saliency modeling progress, it is usually considered a separate task to image saliency prediction [BKEE18, WSX+19, JXL+18, MC19, LMN+19, LWSS19] although both strive to model human visual attention. Current dynamic models use image data only for pre-training [BKEE18, JXL+18, MC19, LMN+19, LWSS19] or auxiliary loss functions [WSG+18]. In addition, many dynamic models are incompatible with image inputs since they require optical flow [BKEE18, LWSS19] or fixed-length video clips for spatio-temporal convolutions [JXL+18, MC19]. In this paper, we ask the question: *Is it possible to model static and dynamic saliency via one unified framework, with mutual benefit?*

First, we present experiments that identify the domain shift between image and video saliency data and between different video saliency datasets as a crucial hurdle for joint modelling. Consequently, we propose suitable domain adaptation techniques for the identified sources of domain shift. To study the benefit of the proposed techniques, we introduce the UNISAL neural network, which is designed to model visual saliency on image and video data coequally while aiming for simplicity and low computational complexity. The network is simultaneously trained on three

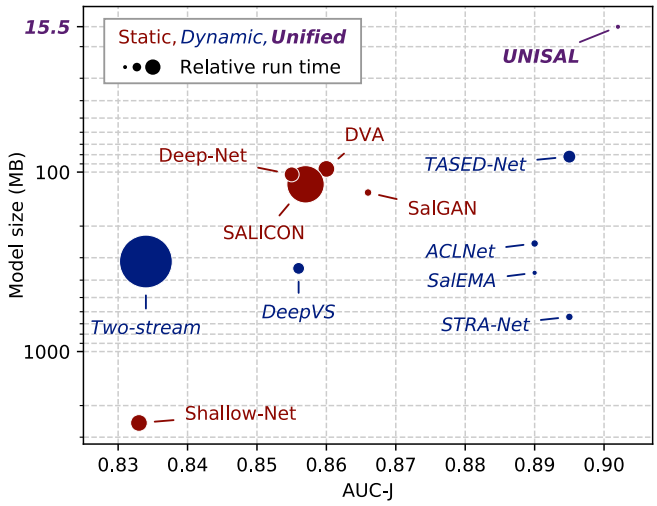

Figure 4.4: Comparison of the proposed model with current state-of-the-art methods on the DHF1K benchmark [WSG+18]. The proposed model is more accurate (as measured by the official ranking metric AUC-J [BJO+19]) despite a model size reduction of 81% or more.

video datasets—DHF1K [WSG+18], Hollywood-2 and UCF-Sports [MS15]—and one image saliency dataset, SALICON [JHDZ15].

We evaluate our method on the four training datasets, among which DHF1K and SALICON have held-out test sets. In addition, we evaluate on the established MIT300 image saliency benchmark [JDT12]. We find that our model significantly outperforms current state-of-the-art methods on all video saliency datasets and achieves competitive performance for the image saliency datasets, with a fraction of the model size and faster runtime than competing models. The performance of UNISAL on the challenging DHF1K benchmark is shown in Figure 4.4. In summary, our contributions are as follows:

- To the best of our knowledge, we make the first attempt to model image and video visual saliency with one unified framework.

- We identify different sources of domain shift as the main challenge for joint image and video saliency modeling and propose four novel domain adaptation techniques to enable strong shared features: Domain-Adaptive Priors, Domain-

Adaptive Fusion, Domain-Adaptive Smoothing, and Bypass-RNN.

- Our method achieves state-of-the-art performance on all video saliency datasets and is on par with the state-of-the-art for all image saliency datasets. At the same time, the model achieves a 5 to 20-fold reduction in model size and faster runtime compared to all existing deep saliency models.

4.2.2 Related Work

Image Saliency Modeling. Most visual saliency modeling literature aims to predict human visual attention mechanisms on static scenes. Early saliency models [IKN98, BI13, SF03, HKP07, LMLCBT06, JEDT09] focus on low-level image features such as intensity/contrast, color, edges, *etc.*, and are are therefore referred to as *bottom-up* methods. Recently, the field has achieved significant performance gains through deep neural networks and their capacity to learn high-level, *top-down* features, starting with Vig *et al.* [VDC14] who propose the first neural network-based approach. Jiang *et al.* [JHDZ15] collect a large-scale saliency dataset, SALICON, to facilitate the exploration of deep learning-based saliency modeling. Zheng *et al.* [ZJCL18] investigate the impact of high-level observer tasks on saliency modeling. Other papers mainly focus on network architecture design with increasing model sizes. For instance, Pan *et al.* [PSGiN+16] evaluate shallow and deep CNNs for saliency prediction, and Kruthiventi *et al.* [KAB15] introduce dilated convolutions and Gaussian priors into the VGG network architecture. Kuemmerer *et al.* [KWB16] propose a simplified VGG-based network while Wang *et al.* [WS18] add skip connections to fuse multiple scales and Cornia *et al.* [CBSC16] add an attentive convolutional LSTM and learned Gaussian priors. Yang *et al.* [YLJL19] expand on the idea of dilated convolutions based on the inception network architecture. While exploration is still ongoing for image saliency modeling, dynamic scenes are arguably at least as relevant to human visual experience, but have received

less attention in the literature to date.

Video Saliency Modeling. Similar to image saliency models, early dynamic models [MPG+09, MV09, RGSZM13, HZ09] predict video saliency based on low-level visual statistics, with additional temporal features (*e.g.*, optical flow). Marat *et al.* [MPG+09] use video frame pairs to compute a static and a dynamic saliency map, which are fused for the final prediction. Marat *et al.* [MPG+09] and Zhong *et al.* [ZLR+13] combine spatial and temporal saliency features and fuse the predictions. By extending the center-surround saliency in static scenes, Mahadevan *et al.* [MV09] use dynamic textures to model video saliency. The performance of these early models is limited by the ability of the low-level features to represent temporal information. Consequently, deep learning based methods have been introduced for dynamic saliency modeling in recent years. Gorji *et al.* [GC18] propose to incorporate attentional push for video saliency prediction, via a multi-stream convolutional long short-term memory network (ConvLSTM). Jiang *et al.* [JXL+18] show that human attention is attracted to moving objects and propose a saliency-structured ConvLSTM to generate video saliency. A recent work [WSX+19] presents a new large-scale video saliency dataset, DHF1K, and propose an attention mechanism with ConvLSTM to achieve better performance than static deep models. The DHF1K dataset, sparked advances [MC19, LWSS19, LMN+19] in video saliency prediction, exploring different strategies to extract temporal features (optical flow, 3D convolutions, different recurrences). However, the above methods either extend prior image saliency models or focus on video data alone with limited applicability to static scenes. Guo *et al.* [GMZ08] present a spatio-temporal model that predicts image and video saliency through the phase spectrum of the Quaternion Fourier Transform but the model lacks the necessary high-level information for accurate saliency prediction. While a recent learning-based approach [LSXW16] extends the image domain to the spatio-temporal domain by using LSTMs, such models

are specialized for video data, rendering them unable to simultaneously model image saliency.

Domain Adaptation. We focus on domain specific learning, a form of domain adaptation which enables a learning system to process data from different domains by separating domain-invariant (shared) and domain-specific (private) parameters [CYS+19]. Domain Separation Networks (DSN) [BTS+16], for instance, are autoencoders with additional private encoders. Instead of an autoencoder, Tsai *et al.* [TC17] introduce an adversarial loss that enforces shared and private encoders networks. Xiao *et al.* [XLOW16] propose Domain Guided Dropout that results in different sub-networks for each domain, and Rozantev *et al.* [RSF19] train entirely separate networks for each domain, coupled through a similarity loss. In contrast to using separate networks, the AdaBN method [LWS+17] adjusts the batch-normalization (BN) parameters of a shared network based on samples from a given target domain. The DSBN method [CYS+19] generalizes this idea by training a separate set of BN parameters for each domain. In general, these existing methods result in a large proportion of domain-specific parameters. In contrast, we propose domain-adaptation techniques that are aimed to bridge the domain gap of saliency datasets with a maximum proportion of shared parameters.

4.2.3 Unified Image and Video Saliency Modeling

Domain-Shift Modeling

In this section we present analyses to examine the domain shift between image and video data and between different video saliency datasets. We use the insights to design corresponding domain adaptation methods. Following Wang *et al.* [WSX+19], we select the video saliency datasets DHF1K [WSX+19], Hollywood-2 and UCF Sports [MS15], and the image saliency dataset SALICON [JHDZ15].

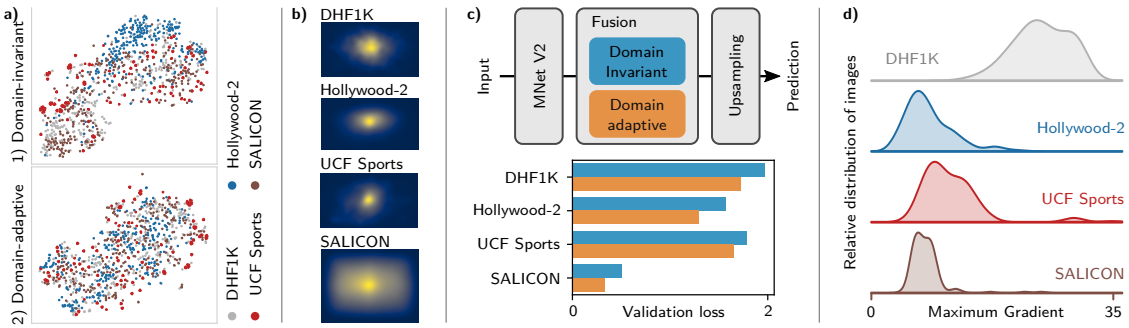

Figure 4.5: Experiments to examine the domain shift between the saliency datasets. **a)** t-SNE visualization of MNet V2 features after domain-invariant and domain-adaptive normalization. **b)** Average ground truth saliency maps. **c)** Comparison of validation losses when training a simple saliency model with domain-invariant and domain-adaptive fusion. **d)** Distributions of ground truth saliency map sharpness.

Domain-Adaptive Batch Normalization. Batch normalization (BN) aims to reduce the internal covariate shift of neural network activations by transforming their distribution to zero mean and unit variance for each training batch. Simultaneously, it computes running estimates of the distribution mean and variance for inference. However, estimating these statistics across different domains results in inaccurate intra-domain statistics, and therefore a performance trade-off. In order to examine the domain shift between the datasets, we conduct a simple experiment: We randomly sample 256 images/frames from each dataset and compute their average pooled MobileNet V2 (MNet V2) features. We then visualize the distribution of the feature vectors via t-SNE [MH08] after normalizing them with the mean and variance of 1) all samples (domain-invariant) or 2) the samples from the respective dataset (domain-adaptive). The results, shown in Figure 4.5 a), reveal a significant domain shift among the different datasets, which is mitigated by the domain-adaptive normalization. Consequently, we employ *Domain-Adaptive Batch Normalization* (DABN), *i.e.*, a different set of BN modules for each dataset. During training and inference, each batch is constructed with data from one dataset and passed through the corresponding BN modules.

Domain-Adaptive Priors. Figure 4.5 b) shows the average ground truth saliency map for each training dataset. Among the video datasets, Hollywood-2 and UCF Sports exhibit the strongest center bias, which is plausible since they are biased towards certain content (movies and sports) while DHF1K is more diverse. SALICON has a much weaker center bias than the video saliency datasets, which can potentially be explained by the longer viewing time of each image/frame (5 s *vs.* 30 ms to 42 ms) that allows secondary stimuli to be fixated. Accordingly, we propose to learn a separate set of Gaussian prior maps for each dataset.

Domain-Adaptive Fusion. We hypothesize that similar image features can have varying visual saliency for images/frames from different training datasets. For example, the Hollywood-2 and UCF Sports datasets are *task-driven*, *i.e.*, the viewer is instructed to identify the main action shown. On the other hand, the DHF1K and SALICON datasets contains *free-viewing* fixations. To test the hypothesis, we design a simple saliency predictor (see Figure 4.5 c): The outputs of the MNet V2 model are fused to a single map by a *Fusion* layer (1×1 convolution) and upsampled through bilinear interpolation. We train the *Fusion* layer until convergence with 1) one set of weights (domain-invariant) or 2) different weights for each dataset (domain-adaptive). We find that the validation loss is lower for all datasets for setting 2), where the network can weigh the importance of the feature maps differently for each dataset. Consequently, we propose to learn a different set of *Fusion* layer weights for each dataset.

Domain-Adaptive Smoothing. The size of the blurring filter which is used to generate the ground truth saliency maps from fixation maps can vary between datasets, especially since the images/frames are resized by different amounts. To examine this effect, we compute the distribution of the ground truth saliency map sharpness for each dataset. Sharpness is computed as the maximum image

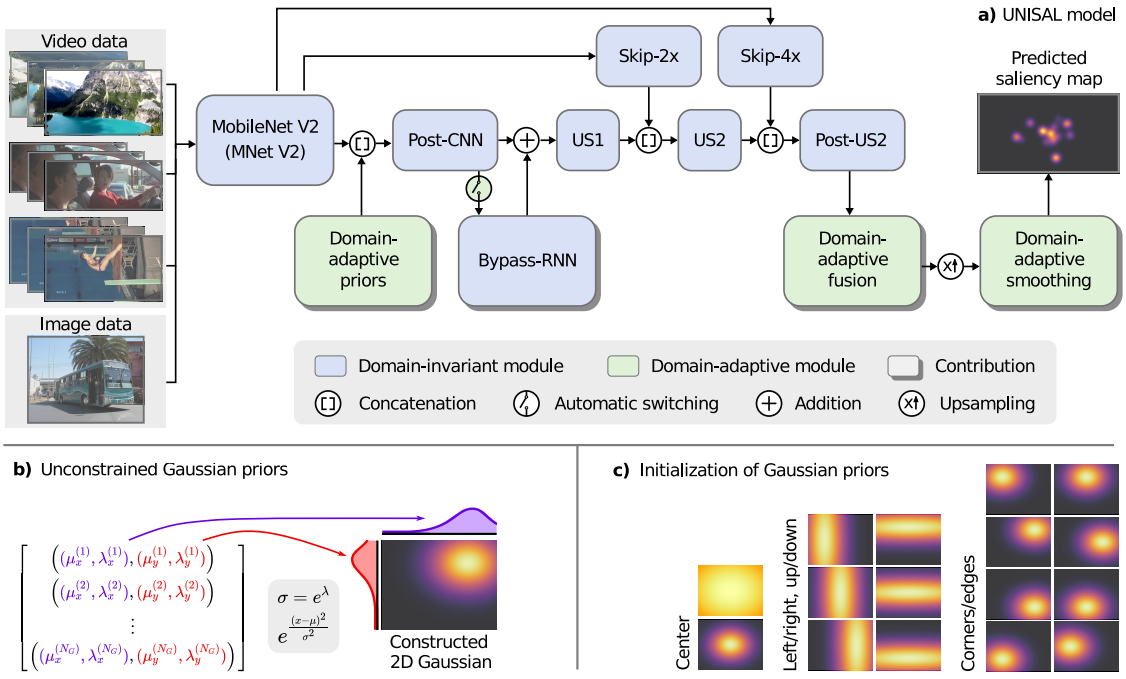

Figure 4.6: a) Overview of the proposed framework. The model consists of a MobileNet V2 (MNet V2) encoder, followed by concatenation with learned Gaussian prior maps, a *Bypass-RNN*, a decoder network with skip connections, and *Fusion* and *Smoothing* layers. The prior maps, fusion, smoothing and batch-normalization modules are domain-adaptive in order to account for domain-shift between the image and video saliency datasets and enable high-quality shared features. **b)** Construction of the prior maps from learned Gaussian meters. **c)** Prior maps initialization.

gradient magnitude after resizing to the model input resolution. The results in Figure 4.5 d) confirm the heterogeneous distributions across datasets, revealing the highest sharpness for DHF1K. Therefore, we propose to blur the network output with a different learned *Smoothing* kernel for each dataset.

UNISAL Network Architecture

We introduce a simple and lightweight neural network architecture termed *UNISAL* that is designed to model image and video saliency coequally and implements the proposed domain-adaptation techniques. The architecture, illustrated in Figure 4.6, follows an encoder-RNN-decoder design tailored for saliency modeling.

Encoder Network. We use MobileNet-V2 (MNet V2) [SHZ+18] as our backbone encoder for three reasons: First, its small memory footprint enables training with sufficiently large sequence length and batch size; second, its small number of floating point operations allows for real time inference; and third, we expect the relatively small number of parameters to mitigate overfitting on smaller datasets like UCF Sports. The main building blocks of MNet V2 are *inverted residuals*, *i.e.*, sequences of pointwise convolutions that decompress and compress the feature space, interleaved with depthwise separable 3×3 convolutions. Overall, for an input resolution of $[r_x, r_y]$, MNet V2 computes feature maps at resolutions of $\frac{1}{2^\alpha}[r_x, r_y]$ with $\alpha \in \{1, 2, 3, 4, 5\}$. The output has 1280 channels and scale $\alpha = 5$. Domain-Adaptive Batch Normalization is not used in MNet V2 since we initialize it with ImageNet-pretrained parameters.

Gaussian Prior Maps. The domain-adaptive Gaussian prior maps are constructed at runtime from learned means and standard deviations. The map with index $i = 1, \ldots, N_G$ is computed as

$$g^{(i)}(x, y) = \gamma \exp\left(-\frac{(x - \mu_x^{(i)})^2}{(\sigma_x^{(i)})^2} - \frac{(y - \mu_y^{(i)})^2}{(\sigma_y^{(i)})^2}\right), \tag{4.2.3.1}$$

where $\gamma = 6$ is a scaling factor since the maps are concatenated with the ReLU6 activations of MNet V2. In this formulation, if the standard deviation $\sigma_{xy}^{(i)}$ is optimized over $\mathbb{R}$, then the resulting variance $(\sigma_{xy}^{(i)})^2$ has the domain $\mathbb{R}_{\geq 0}$, which can lead to division by zero. Prior work which uses non-adaptive prior maps [CBSC16] addresses this by clipping $\sigma_{xy}^{(i)}$ to a predefined interval $[a, b]$ with $a > 0$ and clipping $\mu_{xy}^{(i)}$ to an interval around the center of the map. However, these constraints potentially limit the ability to learn the optimal parameters. Here, we propose *unconstrained Gaussian prior maps* by substituting $\sigma_{xy}^{(i)} = e^{\lambda_{xy}^{(i)}}$ and optimizing $\lambda_{xy}^{(i)}$ and $\mu_{xy}^{(i)}$ over $\mathbb{R}$. Moreover, instead of drawing the initial Gaussian

parameters from a normal distribution, which results in highly correlated maps, we initialize $N_G = 16$ maps as shown in Figure 4.6 c), covering a broad range of priors. Finally, previous work usually introduces prior maps at the second to last layer in order to model the static center bias. Here, we concatenate the prior maps with the encoder output before the RNN and decoder, in order to leverage the prior maps in higher-level features.

Bypass-RNN. Modeling video saliency data requires a strategy to extract temporal features, such as an RNN, optical flow or 3D convolutions. However, none of these techniques are generally suitable to process static inputs, whereas our goal is to process images and videos with one model. Therefore, we introduce a *Bypass-RNN*, *i.e.*, a RNN whose output is added to its input features via a residual connection that is automatically omitted (bypassed) for static batches. during training and inference. Thus, the RNN only models the residual variations in visual saliency that are caused by temporal features.

In the UNISAL model, the *Bypass-RNN* is preceded by a *post-CNN* module, which compresses the concatenated MNet V2 outputs and Gaussian prior maps to 256 channels. For the Bypass-RNN, we use a convolutional GRU (*cGRU*) RNN [VSJR17] due to its relative simplicity, followed by a pointwise convolution. The cGRU has 256 hidden channels, 3×3 kernel size, recurrent dropout [GG16] with probability $p = 0.2$, and MobileNet-style convolutions, *i.e.*, depthwise separable convolutions followed by pointwise convolutions.

Decoder Network and Smoothing. The details of the decoder modules are listed in Table 4.4. First, the Bypass-RNN features are upsampled to scale $\alpha = 4$ by *US1* and concatenated with the output of *Skip-2x*. Next, the concatenated feature maps are upsampled to scale $\alpha = 3$ by *US2* and concatenated with the output of *Skip-4x*. The *Post-US2* features are reduced to a single channel by an

Table 4.4: Network modules and corresponding operations. *ConvDW*(c) denotes a depthwise separable convolution with c channels and kernel size 3×3, followed by batch normalization and ReLU6 activation. *ConvPW*(c_{in}, c_{out}) is a pointwise 1×1 convolution with c_{in} input and c_{out} output channels, followed by batch normalization and, if $c_{\text{in}} \leq c_{\text{out}}$, by ReLU6 activation. DO(p) denotes 2D dropout with probability p. *Up*(c, n) denotes n-fold upsampling with bilinear interpolation of feature maps with c channels.

Module	Operations
Post-CNN	*ConvDW*(1280), *ConvPW*(1280, 256)
Skip-4x	*ConvPW*(64, 128), DO(0.6), *ConvPW*(128, 64)
Skip-2x	*ConvPW*(160, 256), DO(0.6), *ConvPW*(256, 128)
US1	*Up*(256, 2)
US2	*ConvPW*(384, 768), *ConvDW*(768), *ConvPW*(768, 128), *Up*(128, 2)
Post-US2	*ConvPW*(200, 400), *ConvDW*(400), *ConvPW*(400, 64)
Fusion	*ConvPW(64, 1)*

Domain-Adaptive Fusion layer (1×1 convolution) and upsampled to the input resolution via nearest-neighbor interpolation. The upsampling is followed by a *Domain-Adaptive Smoothing* layer with 41×41 convolutional kernels that explicitly models the dataset-dependent blurring of the ground-truth saliency maps. Finally, following Jetley *et al.* [JMV16], we transform the output into a generalized Bernoulli distribution by applying a softmax operation across all output values.

Domain-Aware Optimization

Domain-Adaptive Input Resolution. The images/frames have different aspect ratios for each dataset, specifically 4:3 for SALICON, 16:9 for DHF1K, 1.85:1 (median) for Hollywood-2, and 3:2 (median) for UCF Sports. Our network architecture is fully-convolutional, and therefore agnostic to the exact input resolution. Moreover, each mini-batch is constructed from one dataset due to DABN. Therefore, we use input resolutions of 288×384, 224×384, 224×416 and 256×384 for SALICON, DHF1K, Hollywood-2 and UCF Sports, respectively.

Assimilated Frame Rate. The frame rate of the DHF1K videos is 30 fps compared to 24 fps for Hollywood-2 and UCF Sports. In order to assimilate the

Table 4.5: Overview of the datasets. MIT300 is used only for evaluation.

Dataset	Image/Video	#Videos/images	Resolution	Duration [s]	#Viewers	Task-goal?
SALICON [JHDZ15]	Image	20,000	640×480	N/A	60	No
MIT300 [JDT12]	Image	300	Variable	N/A	39	No
Hollywood-2 [MS15]	Video	1,707	Variable	2-120	19	Yes
UCF Sports [MS15]	Video	150	Variable	2-14	19	Yes
DHF1K [WSG+18]	Video	1,000	640×360	17-42	17	No

frame rates during training, and to train on longer time intervals, we construct clips using every 5^{th} frame for DHF1K and every 4^{th} frame for all others, yielding 6 fps overall. During inference, the predictions are interleaved.

4.2.4 Experiments

In this section, we compare the proposed method with current state-of-the-art image and video saliency models and provide detailed analyses are presented to gain an understanding of the proposed approach.

Experimental Setup

Datasets and Evaluation Metrics. To evaluate our proposed unified image and video saliency modeling framework, we jointly train UNISAL on datasets from both modalities. For fair comparison, we use the same training data as [WSG+18], i.e., the SALICON [JHDZ15] image saliency dataset and the Hollywood-2 [MS15], UCF Sports [MS15], and DHF1K [WSG+18] video saliency datasets. An overview of the datasets if given in Table 4.5. For SALICON, we use the official training/validation/testing split of 10,000/5,000/5,000. For Hollywood-2 and UCF Sports, we use the training and testing splits of 823/884 and 103/47 videos, and the corresponding validation sets are randomly sampled 10% from the training sets, following [WSG+18]. Hollywood-2 videos are divided into individual shots. For DHF1K, we use the official training/validation/testing splits of 600/100/300 videos. We compare against the state-of-the-art methods listed in [WSG+18] and add

newer models with available implementations [MC19, LWSS19, LMN+19, CBSC16, YLJL19]. Moreover, test on the MIT300 benchmark [JDT12], after fine-tuning with the MIT1003 dataset as suggested by the benchmark authors. As in prior work [BI13, WSG+18], we use the evaluation metrics AUC-Judd (AUC-J), Similarity Metric (SIM), shuffled AUC (s-AUC), Linear Correlation Coefficient (CC), and Normalized Scanpath Saliency (NSS) [BJO+19].

Implementation Details. We optimize the network via Stochastic Gradient Descent with momentum of 0.9 and weight decay of 10^{-4}. Gradients are clipped to ± 2. The learning rate is set to 0.04 and exponentially decayed by a factor of 0.8 after each epoch. The batch size is set to 4 for video data and 32 for SALICON. The video clip length is set to 12 frames that are sampled as described in Section 4.2.3. Videos that are too short are discarded for training, which applies to Hollywood-2. For comparability, we use the same loss formulation as Wang *et al.* [WSX+19]. The model is trained for 16 epochs and with early stopping on the DHF1K validation set. To prevent overfitting, the weights of MNet V2 are frozen for the first two epochs and afterwards trained with a learning rate that is reduced by a factor of 10. The pretrained BN statistics of MNet V2 are frozen throughout training. To account for dataset imbalance, the learning rate for SALICON batches is reduced by a factor of 2. Our model is implemented using the PyTorch framework and trained on a NVIDIA GTX 1080 Ti GPU.

Quantitative Evaluation

The results of the quantitative evaluation are shown in Table 4.6 for the video saliency datasets and in Tables 4.7 and 4.8 for the image datasets. For video saliency prediction, in order to analyze the impact of—and generalization across—different datasets, we evaluate six training settings: i) DHF1K, ii) Hollywood-2, iii) UCF Sports, iv) SALICON, v) DHF1K, Hollywood-2, and UCF Sports, vi)

Table 4.6: Quantitative performance on the video saliency datasets. The training settings (i) to (vi) denote training with: (i) DHF1K, (ii) Hollywood-2, (iii) UCF Sports, (iv) SALICON, (v) DHF1K+Hollywood-2+UCF Sports, and (vi) DHF1K+Hollywood-2+UCF Sports+SALICON. Best performance is shown in **bold** while the second best is underlined. The * symbol denotes training under setting (vi), while † indicates that the method is fine-tuned for each dataset.

	Method \ Dataset	DHF1K					Hollywood-2					UCF Sports				
		AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS
Dynamic models	PQFT [GZ09]	0.699	0.139	0.562	0.137	0.749	0.723	0.201	0.621	0.153	0.755	0.825	0.250	0.722	0.338	1.780
	Seo *et al.* [SM09]	0.635	0.142	0.499	0.070	0.334	0.652	0.155	0.530	0.076	0.346	0.831	0.308	0.666	0.336	1.690
	Rudoy *et al.* [RGSZM13]	0.769	0.214	0.501	0.285	1.498	0.783	0.315	0.536	0.302	1.570	0.763	0.271	0.637	0.344	1.619
	Hou *et al.* [HZ09]	0.726	0.167	0.545	0.150	0.847	0.731	0.202	0.580	0.146	0.684	0.819	0.276	0.674	0.292	1.399
	Fang *et al.* [FWLF14]	0.819	0.198	0.537	0.273	1.539	0.859	0.272	0.659	0.358	1.667	0.845	0.307	0.674	0.395	1.787
	OBDL [HKVBS15]	0.638	0.171	0.500	0.117	0.495	0.640	0.170	0.541	0.106	0.462	0.759	0.193	0.634	0.234	1.382
	AWS-D [LGDFVP16]	0.703	0.157	0.513	0.174	0.940	0.694	0.175	0.637	0.146	0.742	0.823	0.228	0.750	0.306	1.631
	OM-CNN [JXL+18]	0.856	0.256	0.583	0.344	1.911	0.887	0.356	0.693	0.446	2.313	0.870	0.321	0.691	0.405	2.089
	Two-stream [BKEE18]	0.834	0.197	0.581	0.325	1.632	0.863	0.276	0.710	0.382	1.748	0.832	0.264	0.685	0.343	1.753
	*ACLNet [WSX+19]	0.890	0.315	0.601	0.434	2.354	0.913	0.542	0.757	0.623	3.086	0.897	0.406	0.744	0.510	2.567
	TASED-Net [MC19]	0.895	0.361	0.712	0.470	2.667	0.918	0.507	0.768	0.646	3.302	0.899	0.469	0.752	0.582	2.920
	STRA-Net [LWSS19]	0.895	0.355	0.663	0.458	2.558	0.923	0.536	0.774	0.662	3.478	0.910	0.479	0.751	0.593	3.018
	†SalEMA [LMN+19]	0.890	**0.465**	0.667	0.449	2.573	0.919	0.487	0.708	0.613	3.186	0.906	0.431	0.740	0.544	2.638
	*SalEMA [LMN+19]	0.895	0.283	**0.739**	0.414	2.285	0.875	0.371	0.663	0.456	2.214	0.899	0.381	0.769	0.521	2.503
Static models	ITTI [IKN98]	0.774	0.162	0.553	0.233	1.207	0.788	0.221	0.607	0.257	1.076	0.847	0.251	0.725	0.356	1.640
	GBVS [HKP07]	0.828	0.186	0.554	0.283	1.474	0.837	0.257	0.633	0.308	1.336	0.859	0.274	0.697	0.396	1.818
	SALICON [HSBZ15]	0.857	0.232	0.590	0.327	1.901	0.856	0.321	0.711	0.425	2.013	0.848	0.304	0.738	0.375	1.838
	Shallow-Net [PSGiN+16]	0.833	0.182	0.529	0.295	1.509	0.851	0.276	0.694	0.423	1.680	0.846	0.276	0.691	0.382	1.789
	Deep-Net [PSGiN+16]	0.855	0.201	0.592	0.331	1.775	0.884	0.300	0.736	0.451	2.066	0.861	0.282	0.719	0.414	1.903
	*Deep-Net [PSGiN+16]	0.874	0.288	0.610	0.374	1.983	0.901	0.482	0.740	0.597	2.834	0.880	0.365	0.729	0.475	2.448
	DVA [WS18]	0.860	0.262	0.595	0.358	2.013	0.886	0.372	0.727	0.482	2.459	0.872	0.339	0.725	0.439	2.311
	*DVA [WS18]	0.883	0.297	0.623	0.397	2.237	0.907	0.497	0.753	0.607	2.942	0.892	0.387	0.740	0.492	2.503
	SalGAN [PFM+17]	0.866	0.262	0.709	0.370	2.043	0.901	0.393	**0.789**	0.535	2.542	0.876	0.332	0.762	0.470	2.238
UNISAL (ours)	Training setting (i)	0.899	0.378	0.686	0.481	2.707	0.920	0.496	0.710	0.612	3.279	0.896	0.443	0.717	0.553	2.689
	Training setting (ii)	0.881	0.313	0.690	0.422	2.352	0.932	0.534	0.762	0.672	3.803	0.892	0.440	0.735	0.566	2.768
	Training setting (iii)	0.869	0.286	0.664	0.375	2.056	0.890	0.392	0.683	0.475	2.350	0.908	0.502	0.764	0.614	3.076
	Training setting (iv)	0.883	0.288	0.715	0.410	2.259	0.912	0.432	0.750	0.565	2.897	0.892	0.428	0.776	0.561	2.740
	Training setting (v)	**0.901**	0.384	0.692	0.488	2.739	**0.934**	**0.544**	0.758	**0.675**	**3.909**	0.917	0.514	**0.786**	0.642	3.260
	Training setting (vi)	**0.901**	0.390	0.691	**0.490**	**2.776**	**0.934**	0.542	0.759	0.673	3.901	**0.918**	**0.523**	0.775	**0.644**	**3.381**

DHF1K, Hollywood-2, UCF Sports and SALICON. For fair comparison, we include state-of-the-art methods that are trained on our best-performing training setting (iv): The ACLNet [WSX+19] video saliency model and the Deep-Net [PSGiN+16] and DVA [WS18] image saliency models. In addition, we provide the performance of SalEMA [LMN+19], which is based on SalGAN [PFM+17], after fine-tuning the model with training setting (vi). Other state-of-the-art video saliency models [JXL+18, MC19, LWSS19] are not suitable for training with image data as discussed in Section 4.2.1. We observe that the proposed UNISAL model significantly outperforms previous static and dynamic methods, across almost all metrics. We obtain the following additional findings: 1) Training with all video saliency datasets

Table 4.7: Performance on the SALICON and MIT300 benchmarks. Best performance is shown in **bold** while the second best is underlined. Training setting (vi) is used for UNISAL. See supplementary material for other settings and updated MIT300 results.

Method \ Dataset	SALICON					MIT300				
	AUC-J	SIM	s-AUC	CC	NSS	AUC-J	SIM	s-AUC	CC	NSS
ITTI [IKN98]	0.667	0.378	0.610	0.205	-	0.75	0.44	0.63	0.37	0.97
GBVS [HKP07]	0.790	0.446	0.630	0.421	-	0.81	0.48	0.63	0.48	1.24
SALICON [HSBZ15]	-	-	-	-	-	**0.87**	0.60	**0.74**	0.74	2.12
Shallow-Net [PSGiN+16]	0.836	0.520	0.670	0.596	-	0.80	0.46	0.64	0.53	-
Deep-Net [PSGiN+16]	-	-	0.724	0.609	1.859	0.83	0.52	0.69	0.58	1.51
SAM-ResNet [CBSC16]	**0.883**	-	0.779	0.842	3.204	**0.87**	**0.68**	0.70	**0.78**	**2.34**
DVA [WS18]	-	-	-	-	-	0.85	0.58	0.71	0.68	1.98
SalGAN [PFM+17]	-	-	0.772	0.781	2.459	0.86	0.63	0.72	0.73	2.04
DINet [YLJL19]	**0.884**	-	**0.782**	0.860	**3.249**	0.86	-	0.71	**0.79**	**2.33**
UNISAL (ours)	0.864	**0.775**	0.739	**0.879**	1.952	**0.872**	**0.674**	**0.743**	**0.784**	**2.322**

Table 4.8: Comparison for dynamic models on the static SALICON benchmark. Best performance is shown in **bold** while the second best is underlined. Training setting (vi) is used for all methods.

Method	AUC-J	SIM	s-AUC	CC	NSS
SalEMA [LMN+19]	0.732	0.470	0.519	0.411	0.760
ACLNet [WSX+19]	0.843	0.688	0.698	0.771	1.618
UNISAL (w/o DA)	0.848	0.690	0.676	0.799	1.654
UNISAL (final)	**0.864**	**0.775**	**0.739**	**0.879**	**1.952**

(setting (v)) *always* improves performance compared to individual video saliency datasets (settings (i) to (iii)). This has not been the case for UCF Sports in a previous cross-dataset evaluation study [WSX+19]. 2) Additionally including image saliency data (setting (vi)) further improves performance for most metrics for DHF1K and UCF Sports. The exception is Hollywood-2, but the performance decrease is less than 1%.

For image saliency prediction, UNISAL performs on par with state-of-the-art image saliency models both on the SALICON and MIT300 benchmark as shown in Table 4.7. In addition, we evaluate state-of-the-art video saliency models on SALICON dataset as shown in Table 4.8. For ACLNet [WSX+19] we use the auxiliary output which is trained on SALICON (using the LSTM output yielded

Figure 4.7: Qualitative performance of the proposed approach on video (top part) and image (bottom part) saliency prediction.

worse performance). For SalEMA [LMN+19], we fine-tuned their best performing model with training setting (vi). A large performance jump can be observed for the domain-adaptive UNISAL model.

Qualitative Evaluation

In Figure 4.7, we show randomly selected saliency predictions for both images and videos. It is visible that the proposed unified model performs well on both modalities. For challenging dynamic scenes with complete occlusion (DHF1K, left), the model correctly memorizes the salient object location, indicating that long-term temporal dependencies are effectively modeled. Moreover, the model correctly predicts shifting observer focus in the presence of multiple salient objects, as evident from the Hollywood-2 and UCF Sports samples. The results on static scenes (bottom part of Figure 4.7) confirm that the proposed unified model indeed generalizes to static scenes.

Ablation Study

We analyze the contribution of each proposed component: 1) Gaussian prior maps; 2) RNN residual connection; 3) skip connections; 4) *Smoothing* layer; 5) domain-adaptive operations (incl. Bypass-RNN); and 6) domain-aware optimization. We

Table 4.9: Ablation study of the proposed approach on the DHF1K and SALICON validation sets. The proposed components are added incrementally to the baseline to quantify their contribution. Training setting (vi) is used for this study.

Config. \ Dataset	DHF1K						SALICON					
	KLD ↓	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑	KLD ↓	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
Baseline	1.877	0.863	0.282	0.659	0.372	2.057	0.551	0.824	0.607	0.633	0.711	1.415
+ Gaussian	1.776	0.879	0.300	0.668	0.411	2.273	0.394	0.848	0.675	0.685	0.801	1.634
+ RNNRes	1.754	0.881	0.302	0.666	0.411	2.274	0.450	0.843	0.648	0.665	0.770	1.531
+ SkipConnect	1.749	0.884	0.308	0.658	0.412	2.301	0.404	0.841	0.673	0.664	0.777	1.600
+ Smoothing	1.770	0.882	0.295	0.677	0.416	2.305	0.369	0.848	0.690	0.676	0.799	1.654
+ DomainAdaptive	1.526	0.907	0.373	0.685	0.482	2.731	0.231	0.867	0.768	0.712	0.877	1.925
Final	1.531	0.907	0.381	0.691	0.487	2.755	0.226	0.867	0.771	0.725	0.880	1.923

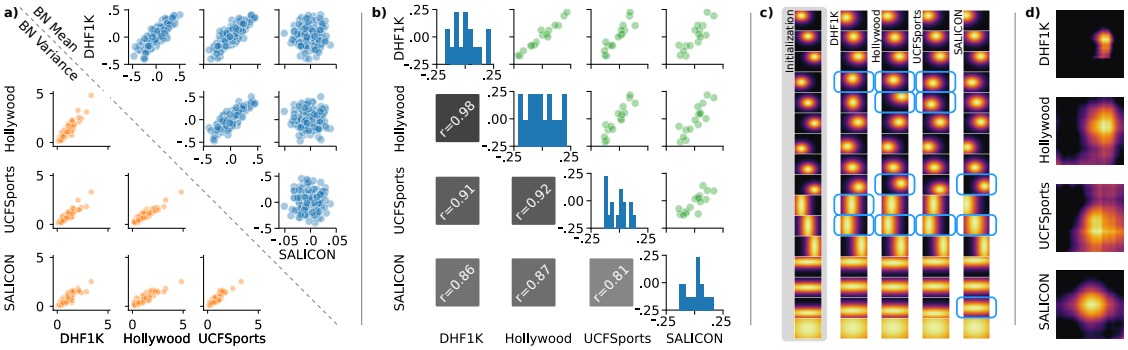

Figure 4.8: Retrospective analysis of the domain-adaptive modules. a) Correlation of the batch normalization statistics between datasets (*US2* module, representative). The upper-right plots correlate the estimated means and the lower-left plots the estimated variances. b) Correlation of the *Fusion* layer weights between datasets. The plots on the diagonal show the distribution of weights of the respective dataset. The lower-left part shows Pearson's correlation coefficients. c) Gaussian prior maps. Significant deviations from the initialization are highlighted. d) *Smoothing* kernel of each dataset.

perform the ablation on the representative DHF1K and SALICON validation sets. The results in Table 4.9 show that each of the proposed components contributes a considerable performance increase. Overall, the domain-adaptive operations contribute the most, both for DHF1K and SALICON. This indicates that mitigating the domain shift between datasets is a crucial component of UNISAL, confirming our initial studies in Section 4.2.3. The Gaussian prior maps yield the second largest gain, indicating the effectiveness of their proposed unconstrained optimization and early position in the model.

Inter-Dataset Domain Shift

Figure 4.8 shows the retrospective analysis of the four domain-adaptive modules. The DABN estimated means in Figure 4.8 a) are correlated among video datasets with Pearson correlation coefficients r between 82% to 83%, but not correlated between SALICON and the video datasets ($r < 3\%$). Similarly, the DABN variances are least correlated between SALICON and the video datasets (90% vs 92%). This confirms the shift of the feature distributions between datasets, especially between SALICON and the video data. The domain-adaptive *Fusion* layer weights shown in Figure 4.8 b) are generally correlated across datasets, with $r > 81\%$. However, as for the DABN, SALICON is the least correlated with the other datasets. Moreover, many of the SALICON *Fusion* weights lie near zero compared to the video datasets, which indicates that only a subset of the video saliency features is relevant for image saliency. The *Domain-Adaptive Fusion* layer models these differences while the remaining network weights are shared. The domain-adaptive Gaussian prior maps shown in Figure 4.8 c) are successfully learned with our proposed unconstrained parametrization, as observed by the deviations from the initialization. Some prior maps are similar across datasets while others vary visibly, indicating that the different domains have different optimal priors. Finally, the learned *Smoothing* kernels shown in Figure 4.8 d) vary significantly across datasets. As expected, the DHF1K dataset, which has the least blurry training targets, results in the most narrow *Smoothing* filter.

Computational Load

With the design of ever more complex network architectures, few studies evaluate the model size, although performance gains can often be traced back to more parameters. We compare the size of UNISAL to the state-of-the-art video saliency predictors in the left column of Table 4.10. Our model is the most light-weight by a

Table 4.10: Model size and runtime comparison of saliency prediction methods (based on the DHF1K benchmark [WSX+19]). Best performance is shown in **bold**.

Method	Model size (MB)	Method	Runtime (s)
Shallow-Net [PSGiN+16]	2,500	Two-stream [BKEE18]	20
STRA-Net [LWSS19]	641	SALICON [HSBZ15]	0.5
SalEMA [LMN+19]	364	Shallow-Net [PSGiN+16]	0.1
Two-stream [BKEE18]	315	DVA [WS18]	0.1
ACLNet [WSX+19]	250	Deep-Net [PSGiN+16]	0.08
SalGAN [PFM+17]	130	TASED-Net [MC19]	0.06
SALICON [HSBZ15]	117	ACLNet [WSX+19]	0.02
Deep-Net [PSGiN+16]	103	SalGAN [PFM+17]	0.02
DVA [WS18]	96	STRA-Net [LWSS19]	0.02
TASED-Net [MC19]	82	SalEMA [LMN+19]	0.01
UNISAL (ours)	**15.5**	UNISAL (ours)	**0.009**

significant margin, with over 5× smaller size than TASED-Net, which is the current state-of-the-art on the DHF1K benchmark (see also Figure 4.4). The same result applies when comparing to the deep image saliency methods from Table 4.7,

Another key issue for real-world applications is the model efficiency. Consequently, we present a GPU runtime comparison (processing time per frame) of video saliency models in the right column of Table 4.10. Our model is the most efficient compared to previous state-of-the-art methods. In addition, we observe a CPU (Intel Xeon W-2123 at 3.60GHz) runtime of 0.43 s (2.3 fps), which is faster than some models' GPU runtime. Considering both the model size and the runtime, the proposed saliency modeling approach achieves state-of-the-art performance in terms of real-world applicability. While the MNet V2 encoder makes a large contribution to low model size and runtime, other contributing factors are: Separable convolutions throughout the cGRU and decoder; cGRU at the low-resolution bottleneck; bilinear upsampling. Without these measures the model size and runtime increase to 59.4 MB and 0.017 s, respectively.

4.2.5 Discussion and Conclusion

In this paper, we have presented a simple yet effective approach to unify static and dynamic saliency modeling. To bridge the domain gap, we found it crucial to account for different sources of inter-dataset domain shift through corresponding novel domain-adaptive modules. We integrated the domain-adaptive modules into the new, lightweight and simple UNISAL architecture which is designed to model both data modalities coequally. We observed state-of-the-art performance on video saliency datasets, and competitive performance on image saliency datasets, with a 5 to 20-fold reduction in model size compared to the *smallest* previous deep model, and faster runtime. We found that the domain-adaptive modules capture the differences between image and video saliency data, resulting in improved performance on each individual dataset through joint training. We presented preliminary and retrospective experiments which explain the merit of the domain-adaptive modules. To our knowledge, this is the first attempt towards unifying image and video saliency modeling in a single framework. We believe that our work can serve as a basis for further research into joint modeling of these modalities.

Bibliography

[BI13] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[BJO+19] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757, 2019.

[BKEE18] Cagdas Bak, Aysun Kocak, Erkut Erdem, and Aykut Erdem. Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction. *IEEE Trans. Multimed.*, 20(7):1688–1698, 2018.

[Bor18] Ali Borji. Saliency Prediction in the Deep Learning Era: An Empirical Investigation. *arXiv:1810.03716*, 2018.

[BTS+16] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In

International Conference on Neural Information Processing Systems, 2016.

[CBSC16] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2016.

[CYS+19] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, 2019.

[FWLF14] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE TIP*, 23(9):3910–3921, 2014.

[GC18] Siavash Gorji and James J Clark. Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *CVPR*, 2018.

[GG16] Yarin Gal and Zoubin Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *NIPS*, 2016.

[GMZ08] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, 2008.

[GZ09] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP*, 19(1):185–198, 2009.

[HKP07] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NeurIPS*, 2007.

[HKVBS15] Sayed Hossein Khatoonabadi, Nuno Vasconcelos, Ivan V Bajic, and Yufeng Shan. How many bits does it take for a stimulus to be salient? In *CVPR*, 2015.

[HSBZ15] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015.

[HZ09] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. In *NeurIPS*, 2009.

[IKN98] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998.

[JDT12] Tilke Judd, Frédo Durand, and Antonio Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations. *Mit-Csail-Tr-2012*, 1:1–7, 2012.

[JEDT09] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to Predict Where Humans Look. In *ICCV*, 2009.

[JHDZ15] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, 2015.

[JMV16] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-End Saliency Mapping via Probability Distribution Prediction. *CVPR*, 2016.

[JXL+18] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. DeepVS: A Deep Learning Based Video Saliency Prediction Approach. In *European Conference on Computer Vision*, 2018.

[KAB15] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations. *IEEE Trans. Image Process.*, 26(9):4446–4456, 2015.

[KWB16] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563*, October 2016.

[LGDFVP16] Victor Leboran, Anton Garcia-Diaz, Xosé R Fdez-Vidal, and Xosé M Pardo. Dynamic whitening saliency. *IEEE TPAMI*, 39(5):893–907, 2016.

[LMLCBT06] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE TPAMI*, 28(5):802–817, 2006.

[LMN+19] Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Kevin McGuinness, Xavier Giro-i Nieto, and Noel E. O'Connor. Simple vs complex temporal recurrences for video saliency prediction. In *BMVC*, 2019.

[LSXW16] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016.

[LWS+17] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting Batch Normalization For Practical Domain Adaptation. In *ICLR Workshops*, 2017.

[LWSS19] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE TIP*, 2019.

[MC19] Kyle Min and Jason J. Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *ICCV*, 2019.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[MPG+09] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International journal of computer vision*, 82(3):231, 2009.

[MS15] Stefan Mathe and Cristian Sminchisescu. Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1408–1424, 2015.

[MV09] Vijay Mahadevan and Nuno Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE TPAMI*, 32(1):171–177, 2009.

[PFM+17] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv:1701.01081*, 2017.

[PSGiN+16] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016.

[RGSZM13] Dmitry Rudoy, Dan B Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, 2013.

[RSF19] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):801–814, April 2019.

[SF03] Yaoru Sun and Robert Fisher. Object-based visual attention for computer vision. *Artificial intelligence*, 146(1):77–123, 2003.

[SHZ+18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[SM09] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009.

[TC17] Jen-Chieh Tsai and Jen-Tzung Chien. Adversarial domain separation and adaptation. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, September 2017.

[VDC14] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014.

[VSJR17] Sepehr Valipour, Mennatullah Siam, Martin Jagersand, and Nilanjan Ray. Recurrent Fully Convolutional Networks for Video Segmentation. In *ArxiV*, 2017.

[WS18] Wenguan Wang and Jianbing Shen. Deep Visual Attention Prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2018.

[WSG+18] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting Video Saliency: A Large-scale Benchmark and a New Model. *CVPR*, 2018.

[WSX+19] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE TPAMI*, 2019.

[XLOW16] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. In *arXiv:1604.07528 [Cs]*, 2016.

[YLJL19] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A dilated inception network for visual saliency prediction. *IEEE TMM*, 2019.

[ZJCL18] Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson WH Lau. Task-driven webpage saliency. In *ECCV*, 2018.

[ZLR+13] Sheng-hua Zhong, Yan Liu, Feifei Ren, Jinghuan Zhang, and Tongwei Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *AAAI*, 2013.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Unified Image and Video Saliency Modeling
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Richard Droste, Jianbo Jiao, and J. Alison Noble. "Unified Image and Video Saliency Modeling". In: European Conference on Computer Vision (ECCV) 2020.

Student Confirmation

Student Name:	Richard Droste		
Contribution to the Paper	Co-lead author responsible for conceiving the original idea. Together with Jianbo Jiao I was responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft.		
Signature	Richard Droste	Date	31.05.2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. J. Alison Noble			
I confirm that the description above is accurate.			
Signature	J Alison Noble	Date	01.06.2021

This completed form should be included in the thesis, at the end of the relevant chapter.

5

Ultrasound Image Analysis with Visual Saliency Models

Contents

5.1 Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention

Authors. Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T. Papageorghiou, J. Alison Noble

Conference. *International Conference on Information Processing in Medical Imaging (IPMI)* 2019.

Background. The method of learning features that are predictive for a task of interest is known as feature representation learning, or simply representation learning. A notable recent advancement in this field is *self-supervised learning*, i.e., learning feature representations by training a neural network to predict surrogate labels which can be obtained automatically, such as transformations applied to the input images [GSK18] or an audio signal [AZ18]. Much in the same way as an audio signal is recorded alongside a video, the PULSE study records gaze-tracking data alongside ultrasound scanning. Consequently, in this study we explore visual saliency prediction of sonographer gaze for ultrasound image representation learning. Following the goal of this thesis to advance ultrasound image analysis, we evaluate the model on ultrasound standard plane detection, which can be used to assist the scanning process by identifying views of interest in real-time [BKM+17].

Statement of Authorship. I was the lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft. J. Alison Noble was the main responsible for supervision and funding acquisition and contributed to conceptualization, methodology and editing the draft. Aris T. Papageorghiou contributed to supervision, funding acquisition and editing the draft. Yifan Cai, Harshita Sharma, Pierre Chatelain and Lior Drukker contributed to conceptualization, methodology and editing the draft. Lior Drukker and Pierre Chatelain were responsible for data acquisition.

Abstract

Image representations are commonly learned from class labels, which are a simplistic approximation of human image understanding. In this paper we demonstrate that transferable representations of images can be learned without manual annotations by modeling human visual attention. The basis of our analyses is a unique gaze-tracking dataset of sonographers performing routine clinical fetal anomaly screenings. Models of sonographer visual attention are learned by training a convolutional neural network (CNN) to predict gaze on ultrasound video frames through visual saliency prediction or gaze-point regression. We evaluate the transferability of the learned representations to the task of ultrasound standard plane detection in two contexts. Firstly, we perform transfer learning by fine-tuning the CNN with a limited number of labeled standard plane images. We find that fine-tuning the saliency predictor is superior to training from random initialization, with an average F1-score improvement of 9.6% overall and 15.3% for the cardiac planes. Secondly, we train a simple softmax regression 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 a fully-supervised baseline model for all but the last layer.

5.1.1 Introduction

When interpreting images, humans direct their attention towards semantically informative regions [WWP14]. This allocation of visual attention is typically quantified via the distribution of gaze points, which can be recorded with gaze-tracking. There has been great interest in developing models of human visual attention that, given an image, predict the likelihood that each pixel is fixated upon, hereafter referred to as *visual saliency map*. Currently, convolutional neural networks (CNNs) are the most effective visual attention models (VAMs) due to their ability to learn complex feature hierarchies through end-to-end training [Bor18]. Here, we explore the following question: *To what extent can models of human visual attention transfer to tasks such as automatic image classification?*

We explore this question using the application of fetal anomaly ultrasound scanning. The scan is performed during mid-pregnancy in order to detect fetal anomalies that require prenatal treatment and to determine the place, time and mode of birth. Previous work related to this application has focused on detecting so-called ultrasound standard imaging planes through fully-supervised training of image classifiers [BKM+17, CSCN18, SOC+18]. Here, in contrast, we aim to learn transferable representations of the scan data without manual supervision by modeling sonographer visual attention. To this end, we acquire the gaze of sonographers in real-time through unobtrusive gaze-tracking alongside anomaly scan recordings.

Sonographer visual attention is modeled by training a CNN to predict gaze on random video frames. We consider this to be *self-supervised representation learning* since it does not require any manual annotations and gaze data is acquired fully automatically. We extract high-resolution image features by introducing dilated convolutions [KAB15, YK16] into a recently proposed image classification architecture [HSS17]. Two methods for training the model for gaze prediction are evaluated: (i) *Visual saliency prediction:* Ground truth visual saliency maps are

generated and used as training targets [Bor18]. (ii) *Gaze-point regression:* The approach of gaze-point regression [NM17] is much less explored in the literature but is simpler since it does not require explicit modeling of foveal vision for ground truth saliency map generation. An existing mathematically differentiable method is based on a fully-connected layer [NM17] which does not scale well to high-resolution feature maps due to the exponentially increasing number of learnable parameters. Here, we propose a method based on the soft-argmax algorithm by Levine et al. [LFDA15] with no additional learnable parameters compared to saliency prediction.

The learned representations are evaluated on the task of standard plane detection in two contexts. (1) *Transfer learning:* We fine-tune the weights of the entire CNN with a limited number of training samples, thereby assessing the transferability of the learned representations in a realistic scenario. (2) *Softmax regression:* We fix the weights of the CNN and train a simple softmax regression on the spatially average-pooled feature activations of each layer. This procedure determines the generality of the representations independently of any transfer learning hyper-parameters.

Related Work. Visual saliency predictors have previously been employed to aid computer vision tasks. Cornia et al. [CBSC17] use a pre-trained saliency predictor as an attention module within an image captioning architecture. However, no representations are shared between the saliency predictor and the task-specific architecture. Cai et al. [CSCN18] show that saliency prediction can aid fetal abdominal standard plane detection. The authors fine-tune an existing standard plane detector [BKM+17] with manually labeled data, using saliency prediction as an auxiliary task and as an attention module. In contrast, we show that transferable representations can be learned without manual annotations via visual attention modeling only. Moreover, we evaluate our framework on full-length freehand clinical fetal anomaly scans instead of short sequences (sweeps) of the fetal abdomen.

Within the field of unsupervised representation learning, our method is most

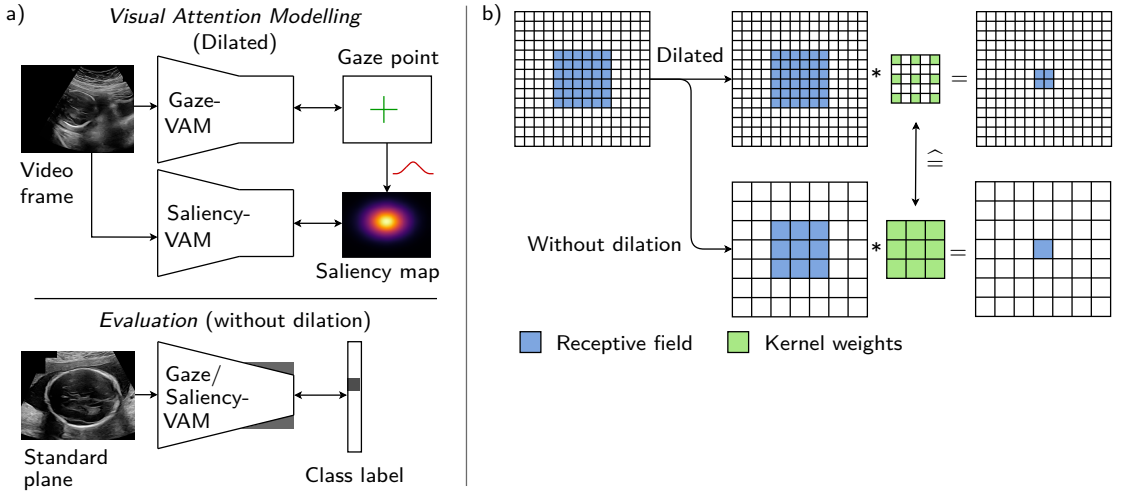

Figure 5.1: a) Illustration of our framework for learning and evaluating visual attention models (VAMs). b) The upper part illustrates a dilated convolution after removed down-sampling operation. When the down-sampling operation is reintroduced for classification as shown in the lower part, the dilation is removed from the kernel without changing the learned kernel weights. The receptive field of the corresponding output neurons is unchanged and the operation is reversible.

closely related to *self-supervised learning*. The general idea is to exploit "free" supervision signals, i.e., supervision signals that can be extracted from the data itself without any manual annotation, which is comparable to our approach of using automatically acquired gaze for supervision. Specifically, representations are learned by either altering the data and inferring the alteration (e.g., spatial and color transformations [DFS+14]) or by predicting certain properties of the data that are withheld (e.g., the relative position of image patches [DGE15] or the order of video frames [MZH16]). However, all existing methods design artificial tasks that yield transferable representations as a "by-product". Human gaze, in contrast, is inherently a strong prior for semantic information [WWP14].

Contributions. Our contributions are three-fold: (1) We propose an original framework for self-supervised image representation learning by modeling human visual attention. The method does not require manual annotations, is generic, and has the potential to be applied in any setting where gaze-tracking and image

data can be acquired simultaneously. To the best of our knowledge, this is the first attempt to study human visual attention modeling in the context of self-supervised representation learning; (2) we propose a method to regress gaze point coordinates via the soft-argmax algorithm, which is significantly simpler and more computationally efficient than the existing method by Ngo et al. [NM17]; (3) finally, we evaluate the attention models on the exemplary task of fetal anomaly ultrasound standard plane detection, both for transfer learning and as fixed feature extractors, thus demonstrating the applicability to a challenging real-world medical imaging task. The framework is illustrated in Fig. 5.1 a).

5.1.2 Representation Learning by Modeling Visual Attention

In this section we describe our method of learning image representations from video and gaze data in general terms. Let $\mathcal{X} \subset \mathbb{R}^{N_c \times H \times W}$ be the set of video frames with width W, height H and N_c channels and let $\mathcal{P} = [0, W] \times [0, H]$ be the set of all valid gaze points. Each frame $\mathbf{X} \in \mathcal{X}$ has a corresponding gaze point set $G = \{\mathbf{p}_i \mid \mathbf{p}_i \in \mathcal{P}\}_{i=1}^{N_G}$ with $N_G \geq 1$. The dataset $\mathcal{D} = \left\{\left(\mathbf{X}^{(t)}, G^{(t)}\right)\right\}_{t=1}^{N_\mathbf{x}}$ consists of $N_\mathbf{x}$ pairs of video frames and gaze point sets.

Let $\boldsymbol{f}_\theta : \mathcal{X} \to \mathbb{R}^{N_f \times \frac{1}{2^d} H \times \frac{1}{2^d} W}$ be a CNN with N_f feature channels, 2^d-fold spatial down-sampling and learnable parameters θ. The final, classification-specific operations (global pooling, fully-connected layers and softmax layers) are removed from the network at this stage. In our experiments we use the SE-ResNeXt [HSS17] model, but any similar feed-forward CNN is suitable. Since such models are designed for image classification, they perform strong down-sampling in order to increase the receptive field of the higher-level neurons and to reduce computational complexity. In contrast, for visual attention modeling, it is desirable to preserve spatial information throughout the network. Consequently,

we remove the last N_D down-sampling operations, i.e., max-pooling or strided convolutions. However, this modification reduces the receptive field of the subsequent neurons. If the down-sampling operations were reintroduced to restore the original architecture and use the representations for classification tasks, the learned weights would be invalid. Therefore, the 3×3 convolutions after the removed down-sampling operations are replaced with *dilated* 3×3 convolutions [YK16] such that each convolutional kernel maintains the same receptive field as in the original architecture, as illustrated in Fig. 5.1 b). Formally, given a matrix $\mathbf{M}$ and the kernel $k : [-r, r]^2 \cap \mathbb{Z}^2 \to \mathbb{R}$ of size $(2r+1)^2$, the l-fold dilated convolution operator $*_l$ is defined as:

$$(\mathbf{M} *_l k)_{i,j} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} \mathbf{M}_{i+ln,j+lm}\, k(n,m) \tag{5.1.2.1}$$

The resulting dilated CNN $\boldsymbol{f}_\theta^{\oplus}$ has the increased output resolution of $(H_D, W_D) \coloneqq (\frac{1}{2^{d-N_D}}H, \frac{1}{2^{d-N_D}}W)$. Next, we want to reduce the high-dimensional feature activations to a single probability map that can be used to model visual attention. Hence, a series of *adaptation layers* consisting of a 7×7 depthwise convolution and several 1x1 convolutions is appended that outputs a single activation map $\mathbf{A} \in \mathbb{R}^{H_D \times W_D}$. A probability map $\hat{\mathbf{S}}$ is then computed by applying a *softmax* across the activations with $\hat{S}_{i,j} = e^{A_{i,j}} / \sum_{i,j} e^{A_{i,j}}$.

We investigate two methods of training the CNN to predict gaze in a differentiable, and therefore end-to-end trainable, manner: *visual saliency prediction* and *gaze-point regression*.

Visual Saliency Prediction. Given an image and a gaze point set $(\mathbf{X}, G) \in \mathcal{D}$, the idea is to generate a *visual saliency map* $\mathbf{S} \in]0,1]^{H \times W}$, where $S_{i,j}$ is the probability that pixel $X_{i,j}$ is fixated upon. The saliency map is then used as the target for the predicted probability map $\hat{\mathbf{S}}$. We generate $\mathbf{S}$ as a sum of Gaussians

around the gaze points in G, normalized such that $\sum_{i,j} S_{i,j} = 1$. The standard deviation of the Gaussians is equivalent to ca. 1° visual angle to account for the radius of visual acuity and the uncertainty of the eye tracker measurements [CSD+18]. Next, the saliency map is downscaled to the size of $\hat{\mathbf{S}}$, yielding the training target $\mathbf{S}^* \in]0,1]^{H_D \times W_D}$. Finally, the training loss is computed via the Kullback-Leibler divergence (KLD) between the predicted and the downscaled true distribution:

$$\mathcal{L}_s(\mathbf{S}^*, \hat{\mathbf{S}}) = D_{\mathrm{KL}}(\mathbf{S}^* \| \, \hat{\mathbf{S}}) = \sum_{i,j} S^*_{i,j} \cdot (log(S^*_{i,j}) - log(\hat{S}_{i,j})) \tag{5.1.2.2}$$

Gaze-Point Regression. We propose a method for reducing $\hat{\mathbf{S}}$ to a single gaze point in order to compare it to the true gaze points. This eliminates the need to model the probability distribution of gaze points via a visual saliency map. First, $\hat{\mathbf{S}}$ is transformed into image coordinates via the soft-argmax algorithm [LFDA15]. With $\boldsymbol{g}(i,j) \coloneqq \left(\frac{j-0.5}{W_D} W, \frac{i-0.5}{H_D} H\right)$ as the function that maps entry (i,j) of $\hat{\mathbf{S}}$ to its corresponding point on the image plane, the predicted gaze point $\hat{\mathbf{p}}$ is computed as the expected value of the probability mass function defined by $\hat{\mathbf{S}}$:

$$\hat{\mathbf{p}} = \sum_{i,j} \hat{S}_{i,j} \, \boldsymbol{g}(i,j) \tag{5.1.2.3}$$

Next, the target gaze point $\mathbf{p}^*$ is obtained from the gaze point set G via the geometric median:

$$\mathbf{p}^* = \underset{\mathbf{p}' \in [0,W] \times [0,H]}{\arg\min} \sum_{p_i \in G} \|\mathbf{p}_i - \mathbf{p}'\|_2 \tag{5.1.2.4}$$

This reduction is justified by the fact that the gaze points on each frame tend to be highly localized due to the short frame period (ca. 33 ms). Finally, the training loss is obtained as $\mathcal{L}_g(\mathbf{p}^*, \hat{\mathbf{p}}) = \|\mathbf{p}^* - \hat{\mathbf{p}}\|_2$, i.e., the Euclidean distance between the predicted and the target gaze point.

Table 5.1: SE-ResNeXt-50 (half-width) [XGD+17] and SonoNet-64 [BKM+17] (variant of VGG-16 [SZ15]) architectures. Convolutional layers are denoted as 'conv <kernel-size>, <output-channels>[, $<C = cardinality>$]', where cardinality is the number of grouped convolutions. SE modules are denoted as 'fc' followed by the dimensions of the corresponding fully-connected layers. Scales in parentheses correspond to the dilated networks for attention modeling. The lower part of the table shows the heads for attention modeling and classification, respectively.

	SE-ResNeXt-50 (half-width, 7.4M parameters)			**SonoNet-64** (14.9M parameters)	
Layer name	Scale	Layers		Scale	Layers
layer 1	224×288 112×144	conv, 7×7, 64, stride 2 max pool, 3, stride 2		224×288	[conv, 3×3, 64] $\times 2$
layer 2	56×72	conv, 1×1, 64 conv, 3×3, 64, $C = 16$ conv, 1×1, 128 fc, $[8, 128]$	$\times 3$	112×144	[conv, 3×3, 128] $\times 2$
layer 3	28×36	conv, 1×1, 128 conv, 3×3, 128, $C = 16$ conv, 1×1, 256 fc, $[16, 256]$	$\times 4$	56×72	[conv, 3×3, 256] $\times 3$
layer 4	14×18 (28×36)	conv, 1×1, 256 conv, 3×3, 256, $C = 16$ conv, 1×1, 512 fc, $[32, 512]$	$\times 6$	28×36	[conv, 3×3, 512] $\times 3$
layer 5	7×9 (28×36)	conv, 1×1, 512 conv, 3×3, 512, $C = 16$ conv, 1×1, 1024 fc, $[64, 1024]$	$\times 3$	14×18	[conv, 3×3, 512] $\times 3$
adaptation (attention)	28×36	conv 7×7, 1024, $C = 1024$ [conv 1×1, 256] $\times 2$ conv 1×1, 1		—	—
adaptation (classification)	7×9	conv 1×1, 256 conv 1×1, N_C avg. pool and softmax		14×18	conv 1×1, 256 conv 1×1, N_C avg. pool and softmax

5.1.3 Experiments

Data. We acquired a novel dataset of clinical fetal ultrasound exams with real-time sonographer gaze-tracking data. The exams are performed on a GE Voluson E8 scanner (General Electric, USA) while the video signal of the machine's monitor is recorded lossless at 30 Hz. Gaze is simultaneously recorded at 90 Hz with a Tobii Eye Tracker 4C (Tobii, Sweden). Ethics approval was obtained for data recording and data are stored according to local data governance rules. For our experiments, we use 135 fetal anomaly scans, which are randomly split into three equally sized subsets for cross-validation.

CNN Architecture. Recent empirical evidence suggests that ImageNet performance is strongly correlated with performance on other vision tasks [KSL18]. Therefore, we base our CNN on SE-ResNeXt [HSS17], a ResNet-style model with aggregated convolutions and channel recalibration (*squeeze-and-excitation*, short *SE*) modules, which won the 2017 ImageNet classification competition. For attention modeling, layers 4 and 5 are dilated as described in subsection 5.1.2. In preliminary experiments we found that halving the number of feature channels (except for layer 0) greatly reduced the computational cost without performance losses on our dataset. The resulting architecture is summarized in column 1 of Table 5.1. Column 2 shows SonoNet-64 [BKM+17], which we use as a reference for standard plane detection since the authors published network weights trained on over 22k standard plane images.

Visual Attention Modeling

Experimental methods. Two visual attention models (VAMs) were trained on the ultrasound video and gaze data as described in subsection 5.1.2, namely a visual saliency predictor (*Saliency-VAM*) and a gaze-point regressor (*Gaze-VAM*). For pre-processing, all video frames that did not correspond to 2D B-mode live scanning (e.g., Doppler, 3D/4D or frozen frames) or without gaze data were discarded. Further, all but every 8^{th} frame were discarded to reduce temporal redundancy, resulting in a total of 403 070 video frames. Next, the frames were cropped down to the region of the actual ultrasound image. Data augmentation was performed by uniformly sampling sub-crops of 70-90% side length that contained the gaze points, random horizontal flipping, and varying gamma and brightness by ±25%. Finally, the frames were down-sampled to a size of 224×288 pixels and normalized to zero-mean and unit-variance.

Both attention models were trained via stochastic gradient descent (SGD) with momentum of 0.9, weight decay of 10^{-4} and mini-batch size of 32. The Saliency-VAM was trained for 8 epochs at a learning rate (LR) of 0.1 while the Gaze-VAM

Table 5.2: Results of visual saliency prediction and gaze-point regression compared to static baselines (mean ± standard deviation). Next to the training loss (KLD), the Saliency-VAM is evaluated on the metrics normalized scanpath saliency (NSS), AUC-Judd, Pearson's correlation coefficient (CC) and histogram intersection (SIM) (for references see [Bor18]). Best values are marked bold.

	Saliency-VAM					Gaze-VAM
	KLD	NSS	AUC [%]	CC [%]	SIM [%]	ℓ_2-norm
Static	3.41 ±0.02	1.39 ±0.05	85.9 ±0.3	14.9 ±0.4	8.5 ±0.1	54.4 ±0.6
Learned	**2.43** ±0.03	**4.03** ±0.05	**96.7** ±0.2	**31.6** ±0.3	**18.5** ±0.2	**27.4** ±0.4

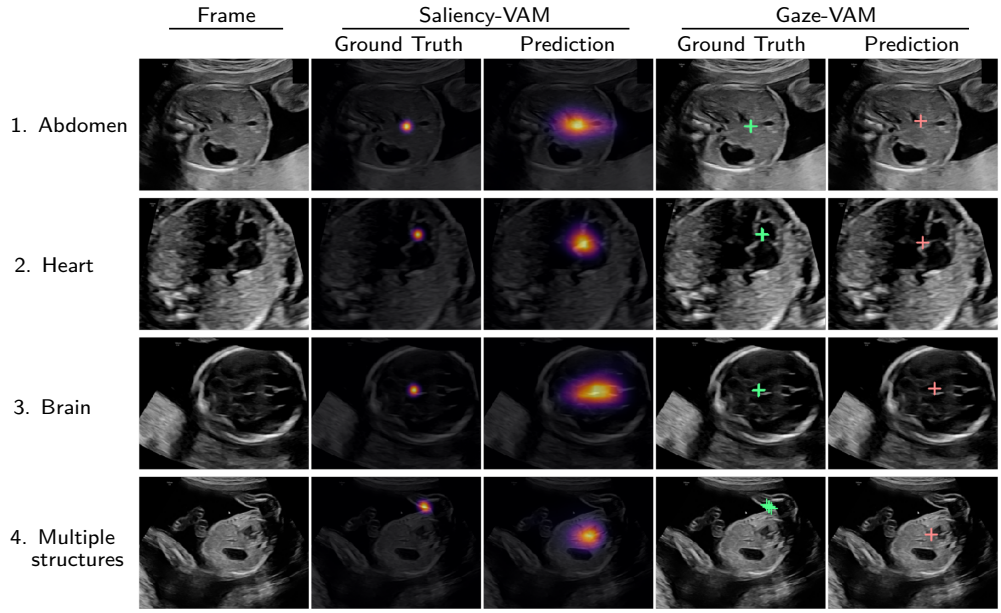

Figure 5.2: Visual saliency and gaze point predictions with corresponding ground truths for representative validation set frames.

converged more slowly and was trained for 10 epochs at a LR of 0.01. In each case, the LR was decayed by a factor of 10 for the final two epochs. All experiments were implemented in the PyTorch framework. Each training run was performed in 9 h to 16 h on a single Nvidia GTX 1080 Ti.

Results. Table 5.2 summarizes the quantitative evaluation of the attention models. The static baseline for the Saliency-VAM is the normalized sum of all ground truth saliency maps. The baseline for the Gaze-VAM is the geometric median of all gaze points. The learned models clearly outperform the static baselines

on every metric. Fig. 5.2 shows visual saliency and gaze point predictions for four representative frames from the validation set. Frames 1.-3. each contain one anatomical structure and show examples of accurate prediction. Frame 4. contains several structures, which creates ambiguity.

Fetal Anomaly Standard Plane Detection

Experimental methods. For comparison with Baumgartner et al. [BKM+17], we consider the same 13 standard plane classes and "background" class, except that our data contains the three vessels and trachea view (3VT) which is similar to their three-vessel view (3VV). From the available 135 anomaly scans, we obtained a total of 1129 standard plane frames with 62 to 148 samples per class (a plane may be acquired twice or may be skipped in a scan). Moreover, we sampled 1127 background frames in the vicinity of the standard planes. The same scan-level three-fold cross-validation split as for attention modeling was applied. For pre-processing, frames were cropped to the ultrasound image region as for attention modeling. The images were then augmented by random horizontal flipping, rotation by $\pm 10°$, varying the aspect-ratio by $\pm 10\%$, sampling a sub-crop of 95-100% side length, and varying gamma and brightness by $\pm 25\%$. As before, the images were down-sampled and normalized.

The trained visual attention models were fine-tuned (FT) on the standard plane detection task, yielding *Saliency-FT* and *Gaze-FT*. Moreover, two baselines were generated: A SE-ResNeXt model trained from random initialization and a fine-tuned SonoNet-64 (*SonoNet-FT*). Each epoch consisted of 1024 randomly sampled images. Analogous to Baumgartner et al., we overcome the class imbalance problem by sampling images from each standard plane class with the same frequency and sampling one background image per standard plane image. Fine-tuning was performed via SGD with momentum of 0.9, weight decay of 5×10^{-4}, mini-batch size of 16 and a cross-entropy loss. The attention models were fine-tuned for 50

epochs with a LR of 0.01, decayed by a factor of 10 at epochs 20 and 35. For the randomly initialized model, the LR was increased by a factor of 4. The SonoNet model was initialized with pre-trained weights published by the authors and fine-tuned for 25 epochs with a LR of 0.01, decayed at epochs 10 and 20. Longer training or higher learning rates led to overfitting for the latter two models due to the relatively small number of training samples. Due to the class imbalance, the overall precision, recall and F1-scores were computed as *macro-averages*, i.e., the average of the scores per standard plane.

Besides fine-tuning, we trained a multinomial logistic regression (softmax regression) on the spatially average-pooled feature activations of each layer of the attention models and two baselines: an SE-ResNeXt model with random weights and the pre-trained SonoNet model. For each regression, the entire respective training set was sampled without augmentation. The classifier was trained with the L-BFGS solver and balanced class weights. The L2 regularization parameter was selected for each regression from a range of 16 logarithmically spaced values from 10^{-5} to 10^{1} based on the validation F1-score.

Results. A quantitative evaluation of the fine-tuned attention models is shown in Table 5.3. The Saliency-FT model improves standard plane detection compared to the model trained from random initialization on every metric and for each standard plane. The largest improvement per anatomy is observed for the right ventricular outflow tract (RVOT) with an average 20.8% increase in F1-score, followed by the left ventricular outflow tract (LVOT) and the four chamber view (4CH). Further, the average F1-score of Saliency-FT exceeds that of SonoNet on 3/14 classes. The Gaze-FT model under-performs compared to training from random initialization. In general, the average precision, recall and F1-score of SonoNet-FT are in good agreement with the literature values. The authors do not provide per-anatomy scores for SonoNet-64.

Table 5.3: Standard plane detection results after fine-tuning (mean ± standard deviation [%]). *Rand. Init.* denotes the SE-ResNeXt model trained from scratch. The best score among the first three models is marked in bold. Scores of the fine-tuned SonoNet that exceed all three models are marked in bold as well. The literature SonoNet scores are given in parenthesis.

	Rand. Init.	Gaze-FT	Saliency-FT	Δ(Saliency, Rand. Init.)	SonoNet-FT (Lit. value [BKM+17])
Precision	70.4 ± 2.3	67.2 ± 3.4	**79.5** ± 1.7	9.1 ± 2.1	**82.3** ± 1.3 (81)
Recall	64.9 ± 1.6	57.3 ± 4.5	**75.1** ± 3.4	10.2 ± 1.9	**87.3** ± 1.1 (86)
F1-score	67.0 ± 1.3	60.7 ± 3.9	**76.6** ± 2.6	9.6 ± 2.1	**84.5** ± 0.9 (83)
F1-scores per class:				↓	
RVOT	37.9 ± 3.8	30.4 ± 4.9	**58.7** ± 2.7	20.8 ± 5.5	**71.2** ± 2.8
LVOT	30.3 ± 4.7	25.8 ± 5.9	**48.6** ± 3.3	18.4 ± 7.3	**69.9** ± 5.3
4CH	43.1 ± 6.7	33.5 ± 8.5	**57.3** ±10.8	14.2 ±11.9	**75.7** ± 9.1
Kidneys	71.4 ± 5.5	68.5 ±12.1	**84.7** ± 6.3	13.3 ± 5.7	81.0 ± 5.0
Profile	77.5 ± 7.2	61.7 ± 7.7	**87.2** ± 7.5	9.7 ± 3.7	**88.1** ± 4.5
Lips	76.7 ± 2.6	74.2 ± 7.3	**85.6** ± 4.5	8.8 ± 6.7	**92.9** ± 0.8
Brain (Cb.)	84.9 ± 7.0	83.2 ± 1.5	**93.7** ± 4.6	8.8 ± 2.3	92.8 ± 1.1
3VT	50.4 ± 1.9	47.1 ± 7.1	**58.3** ± 7.1	7.9 ± 5.6	**77.9** ± 1.6
Brain (Tv.)	86.1 ± 7.7	88.8 ± 2.8	**92.9** ± 5.0	6.8 ± 2.8	92.1 ± 4.5
Spine (cor.)	72.9 ± 3.6	57.2 ± 2.8	**79.0** ± 3.7	6.1 ± 6.2	**90.3** ± 4.9
Abdominal	67.9 ± 5.1	60.8 ± 6.7	**72.9** ± 2.9	5.0 ± 3.7	**85.0** ± 1.4
Spine (sag.)	86.5 ± 3.5	80.2 ± 2.5	**89.1** ± 2.1	2.7 ± 5.2	**91.6** ± 2.5
Femur	85.7 ± 2.0	77.7 ± 0.1	**87.6** ± 1.3	1.9 ± 1.5	**89.5** ± 1.8
Background	85.2 ± 0.9	83.3 ± 0.7	**89.0** ± 0.4	3.8 ± 1.2	**90.3** ± 0.4

RVOT: right ventricular outflow tract; LVOT: left ventricular outflow tract; 4CH: four chamber view; 3VT: three vessel and trachea view; Brain (Cb.): brain cerebellum suboccipitobregmatic plane; Brain (Tv.): brain transventricular plane; Cor.: Coronal plane; Sag.: Sagittal plane.

The results of the regression analysis are shown in Fig. 5.3 a). The scores of both attention models monotonously increase up to layer 4 and stagnate at layer 5, peaking at F1-scores of $58.0 \pm 1.5\%$ for the Gaze-VAM and $64.9 \pm 1.5\%$ for the Saliency-VAM. The scores of the Saliency-VAM and SonoNet are at similar levels up to layer 4, while the Gaze-VAM achieves lower scores. For SonoNet the scores continue to increase at layer 5, reaching an F1-score of $79.9 \pm 0.6\%$. In general, the scores of the random features are comparable to those of the other models at layers 0 and 1 but decline afterwards, peaking at an F1-score of $49.5 \pm 2.7\%$.

The differences between the feature embeddings are illustrated for selected layers in Fig. 5.3 b) via t-SNE [vdMH08], a non-linear dimensionality reduction algorithm that visualizes high-dimensional neighborhoods. Compared to random features, a

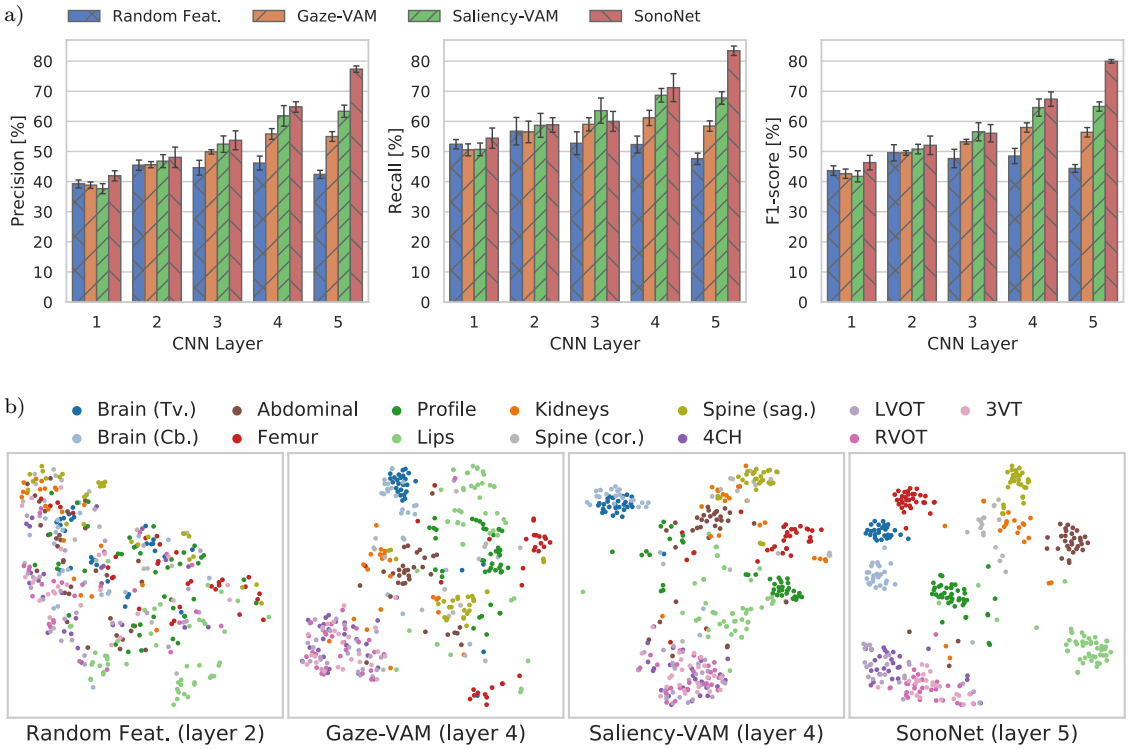

Figure 5.3: a) Results of the regression analysis of the fixed-weight attention models, and baselines. b) t-SNE visualization of the feature embeddings at the respective layers with the highest F1-score (Background class omitted for legibility). Best viewed in color.

separation of the different standard plane classes emerges in the embeddings of the visual attention models. However, a large overlap remains among the two brain views and the cardiac views, respectively. Moreover, the views of coronal spine, kidneys, profile and lips are not well localized. In the embedding of SonoNet, most classes are well-separated with overlap remaining among the cardiac views.

5.1.4 Discussion and Conclusion

The evaluations have shown that the visual attention models have learned meaningful representations of ultrasound image data, which was the main goal of this work. In the transfer learning context, the Saliency-FT model clearly outperforms the model trained from random initialization. The largest benefit is observed for the cardiac views with an average increase in F1-score of 15.3%. In fact, the performance of Saliency-FT is closer to that of SonoNet-FT, although the latter had been pre-

trained with over 22k labeled standard plane images, while the attention models are pre-trained only with sonographer gaze on unlabeled video frames. Since fine-tuning is performed with 753 standard plane images on average, this is a 30-fold reduction in the amount of manually annotated training data. Gaze-FT did not yield an improvement, indicating that visual saliency prediction is better suited to learn transferable representations.

Even without fine-tuning, the high-level features of the attention models are predictive for fetal anomaly standard plane detection, outperforming the baseline with random weights for softmax regression on the feature activations. Up to last network layer, the features of the Saliency-VAM are almost as predictive as those of SonoNet, even though it had received no explicit information about the concept of standard planes during training. This confirms our hypothesis, motivated by Wu et al. [WWP14], that gaze is a strong prior for semantic information. At the last layer, the attention models fall behind SonoNet, indicating the task-specificity of that layer. The qualitative analysis through t-SNE confirms that some standard plane classes are well-separated in the respective feature spaces of the attention models, with overlap remaining for standard planes with similar appearance such as the brain views and the cardiac views, respectively. It should be noted that we did not compare our models to a recently proposed variation of SonoNet [SOC+18] with multi-layer attention-gating due to the added complexity of that architecture.

The results for both visual saliency prediction and gaze-point regression indicate successful learning of sonographer visual attention. This is supported by the fact that the scores on the key metrics of AUC and NSS are higher than the scores reported by Cai et al. [CSCN18] and than typical scores on the public MIT Saliency Benchmark of natural images (`saliency.mit.edu`). However, our scores on the CC, SIM and KLD metrics are worse compared to these sources and in general, the comparability is very limited since the maximum attainable values are dataset-dependent. For gaze-point regression, the proposed method based on the

soft-argmax algorithm was found to be an effective solution.

In conclusion, we have shown that visual attention modeling is a promising method to learn image representations without manual supervision. The trained CNNs generalize well to the task of fetal anomaly standard plane detection, both for transfer learning and as fixed feature extractors. We have evaluated two methods for visual attention modeling, visual saliency prediction and gaze-point regression, and found that the representations learned with the former method generalize better. The representation learning framework presented herein is generic and therefore has the potential to be applied in many settings where gaze and image data can be readily acquired.

Bibliography

[BKM+17] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. *IEEE Trans. Med. Imag.*, 36(11):2204–2215, 2017.

[Bor18] Ali Borji. Saliency Prediction in the Deep Learning Era: An Empirical Investigation. *arXiv:1810.03716*, 2018.

[CBSC17] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Visual saliency for image captioning in new multimedia services. In *ICMEW*, 2017.

[CSCN18] Yifan Cai, Harshita Sharma, Pierre Chatelain, and J. Alison Noble. Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[CSD+18] P. Chatelain, H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies. *IEEE Trans. Cybern.*, pages 1–11, 2018.

[DFS+14] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. In *NIPS*, 2014.

[DGE15] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*, 2015.

[HSS17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *CVPR*, 2017.

[KAB15] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations. *IEEE Trans. Image Process.*, 26(9):4446–4456, 2015.

[KSL18] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? In *arXiv:1805.08974*, 2018.

[LFDA15] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-End Training of Deep Visuomotor Policies. *J. Mach. Learn. Res.*, 17(1):1334–1373, 2015.

[MZH16] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *ECCV*, 2016.

[NM17] Thuyen Ngo and B S Manjunath. Saccade gaze prediction using a recurrent neural network. In *ICIP*, 2017.

[SOC+18] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention-Gated Networks for Improving Ultrasound Scan Plane Detection. In *MIDL*, 2018.

[SZ15] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9(Nov):2579–2605, 2008.

[WWP14] Chia-Chien Wu, Farahnaz Ahmed Wick, and Marc Pomplun. Guidance of visual attention by semantic information in real-world scenes. *Front. Psychol.*, 5:54, 2014.

[XGD+17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*, 2017.

[YK16] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*, 2016.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T. Papageorghiou, J. Alison Noble. "Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention". In: International Conference on Information Processing in Medical Imaging (IPMI) 2019.

Student Confirmation

Student Name:	Richard Droste		
Contribution to the Paper	I was the lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft.		
Signature	Richard Droste	Date	31.05.2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. J. Alison Noble			
I confirm that the description above is accurate.			
Signature	J Alison Noble	Date	01.06.2021

This completed form should be included in the thesis, at the end of the relevant chapter.

5.2 Discovering Salient Anatomical Landmarks by Predicting Human Gaze

Authors. Richard Droste, Pierre Chatelain, Lior Drukker, Harshita Sharma, Aris T. Papageorghiou, J. Alison Noble

Conference. *IEEE International Symposium on Biomedical Imaging (ISBI)* 2020.

Background. In this study we directly build upon the work of the previous section and demonstrate that the feature representations learned through visual saliency prediction are useful not only for image-level classification, but also for the discovery and detection of *salient anatomical landmarks*. In prior work, salient landmarks have been considered only from a purely computational perspective, with saliency defined through low-level image features. In contrast, we measure saliency directly by predicting operator gaze, providing a detector for landmarks that are salient to operators in practice. This method is useful for ultrasound image analysis applications, to advance our understanding of sonographer visual search strategies, and for sonographer training.

Statement of Authorship. I was the lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft. J. Alison Noble was the main responsible for supervision and funding acquisition and contributed to conceptualization, methodology and editing the draft. Aris T. Papageorghiou contributed to supervision, funding acquisition and editing the draft. Pierre Chatelain, Lior Drukker and Harshita Sharma contributed to conceptualization, methodology and editing the draft. Lior Drukker and Pierre Chatelain were responsible for data acquisition.

Recognition. The paper won 2nd Runner Up for the ISBI 2020 Best Paper Award.

Abstract

Anatomical landmarks are a crucial prerequisite for many medical imaging tasks. Usually, the set of landmarks for a given task is predefined by experts. The landmark locations for a given image are then annotated manually or via machine learning methods trained on manual annotations. In this paper, in contrast, we present a method to automatically discover and localize anatomical landmarks in medical images. Specifically, we consider landmarks that attract the visual attention of humans, which we term *visually salient landmarks*. We illustrate the method for fetal neurosonographic images. First, full-length clinical fetal ultrasound scans are recorded with live sonographer gaze-tracking. Next, a convolutional neural network (CNN) is trained to predict the gaze point distribution (saliency map) of the sonographers on scan video frames. The CNN is then used to predict saliency maps of unseen fetal neurosonographic images, and the landmarks are extracted as the local maxima of these saliency maps. Finally, the landmarks are matched across images by clustering the landmark CNN features. We show that the discovered landmarks can be used within affine image registration, with average landmark alignment errors between 4.1% and 10.9% of the fetal head long axis length.

5.2.1 Introduction

An *anatomical landmark* is "a point of correspondence on each object that matches between and within populations" and is assigned "in some scientifically meaningful way" [DM16, p. 3]. For brevity, we will refer to *anatomical landmarks* simply as *landmarks*. The selection and localization of landmarks are essential steps for medical image analysis tasks such as image registration and shape analysis. Usually, the set of landmarks for a given task is selected by experts a priori. The landmark locations for a given image are then either annotated manually or via machine learning models trained on manual annotations. However, when clinicians interpret images in practice based on experience, they may consider only a subset of the predefined landmarks, or use additional, unspecified landmarks. Moreover, it might be desirable to automatically localize landmarks without the need for manual annotations.

Contribution. In this work we overcome these limitations by presenting a method to automatically discover and localize anatomical landmarks. Specifically, the method reveals landmarks that attract the visual attention of clinicians, which we term *visually salient landmarks*. The backbone of the proposed system is a CNN that is trained to predict the gaze-point distributions (saliency maps) of clinicians observing images from the domain of interest. For modalities like ultrasound imaging, gaze-tracking data can be acquired during image acquisition with no additional expert time expenditure. The trained CNN is then used to reveal visually salient landmarks on unseen images and to assign them semantic labels that can be used to match them across images. To the best of our knowledge, this is the first work to present a method to automatically discover landmarks based on *visual* saliency.

Related Work. In previous work, *saliency* is often used to refer to low-level features such as local entropy [KB01, GFD06]. Moreover, mutually-salient landmarks based on Gabor attributes have been proposed for image registration [OSPD11]. Here, in contrast, we use *visual saliency*, i.e., the predicted allocation

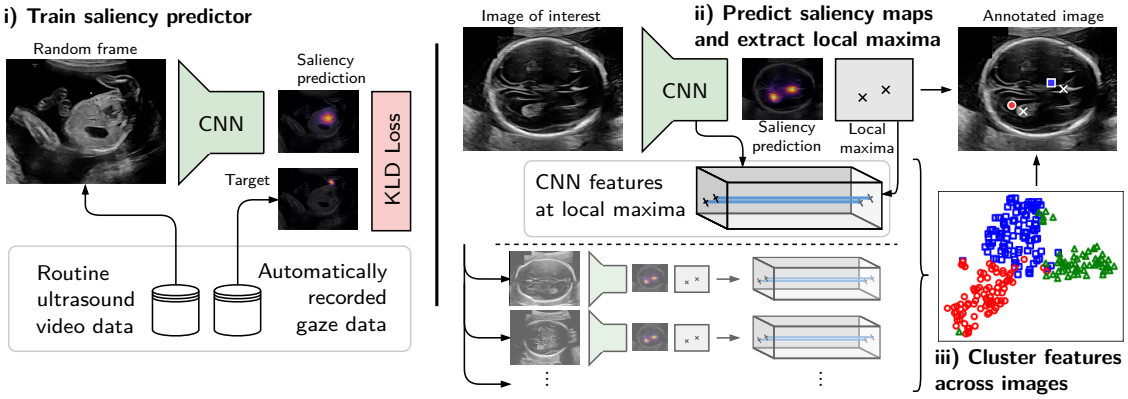

Figure 5.4: Overview of the proposed method for the discovery and localization of visually salient landmarks.

of human visual attention based on gaze-tracking data, to discover *anatomical landmarks*. We apply the method to neurosonographic standard views in fetal anomaly ultrasound scans. The landmarks for these standard views are defined by a set of international practice guidelines [SAB+11]. A landmark detector has previously been developed but is trained on manual annotations of a pre-defined set of landmarks [YKPN17]. Moreover, gaze data has been used to support the detection of standard views in fetal ultrasound scans [CSCN18, DCS+19], but these works do not consider the problem of identifying landmarks.

5.2.2 Methods

Data

The data were acquired as part of the PULSE (Perception Ultrasound by Learning Sonographic Experience) project, a prospective study of routine fetal ultrasound scans performed in all trimesters by sonographers and fetal medicine doctors at the maternity ultrasound unit, Oxford University Hospitals NHS Foundation Trust, Oxfordshire, United Kingdom. The exams were performed on a GE Voluson E8 scanner (General Electric, USA) while the video signal of the machine monitor was recorded lossless at 30 Hz. Operator gaze was simultaneously recorded at 90 Hz with a Tobii Eye Tracker 4C (Tobii, Sweden). This study was approved by the UK

Research Ethics Committee (Reference 18/WS/0051), and written informed consent was given by all participating pregnant women and operators. In this paper, we use ultrasound video and corresponding gaze data of 212 second trimester scans acquired between May 2018 and February 2019.

We selected 90 scans to train the saliency predictor and used the remaining 122 scans to evaluate the landmark discovery method. We considered the fetal neurosonographic standard views, i.e., the transventricular (TV) and the transcerebellar (TC) plane (first row in Fig. 5.5). On the TV plane the operators measure the head circumference (HC) and the lateral ventricle (LV). On the TC plane they measure the transcerebellar diameter (TCD), the nuchal fold and the cisterna magna. The views are defined by the visibility of these structures as well as the appearance of the cavum septi pellucidi (CSP). From the 122 ultrasound scans, we automatically extracted 143 TV and 124 TC plane images by performing optical character recognition on the machine's graphical interface.

Visually Salient Landmark Discovery

Visually salient anatomical landmarks are discovered in three steps (see Fig. 5.4): i) training a CNN to predict the sonographer gaze point distributions (saliency maps) on random video frames of the routine fetal ultrasound scan data described above; ii) predicting the visual saliency maps of the neurosonographic images and extracting the landmark locations as the local maxima of the saliency maps; and iii) clustering the CNN feature vectors which correspond to the landmark locations.

i) To train the saliency predictor, we use the CNN architecture and training procedure detailed in previous work [DCS+19] (model *Saliency-VAM*). The precise architecture and training procedure are not repeated here as they are not essential for the proposed landmark discovery method. The CNN takes ultrasound images of dimension 288×244 as input and performs three two-fold down-sampling operations, which results in output saliency maps of dimensions $W_s \times H_s = 36 \times 28$.

ii) Let $s_i : [1, W_s] \times [1, H_s] \cap \mathbb{Z}^2 \rightarrow [0, 1]$ be the function which, for an image with index $i = 1, \ldots, N_i$, maps each saliency map location to its predicted saliency value (i.e., the probability that the location is gazed at). The local maxima of this predicted saliency map are found with the scikit-image (https://scikit-image.org/) *peak_local_max* algorithm. The algorithm first applies a maximum filter

$$s_i^{\max}(x, y) := \max_{(x', y') \in [-d, d]^2 \cap \mathbb{Z}^2} s_i(x + x', y + y') \ , \tag{5.2.2.1}$$

where d is the minimum distance of any two local maxima (empirically $d = 2$). The local maxima are then extracted as the points where the s equals $s^{\max}$ and s is above a threshold t to suppress spurious maxima (empirically $t = 0.1$):

$$\mathcal{M}_i := \{(x, y) | s_i(x, y) = s_i^{\max}(x, y) \wedge s_i(x, y) \geq t\} \tag{5.2.2.2}$$

The landmark locations are obtained by fitting a 2D Gaussian peak to a 3×3 neighborhood around the saliency map maxima.

iii) Once the landmark locations are extracted, their correspondence across images is still unknown. Recent work has shown that saliency predictors implicitly learn *global* semantic features which are useful for image classification [DCS+19]. Here, we hypothesize that saliency predictors can also be used to extract *local* semantic features which allow automatic landmark classification. Let $f_i : [1, W_s] \times [1, H_s] \cap \mathbb{Z}^2 \rightarrow \mathbb{R}^{N_f}$ be the function which, for image i, maps each location of the saliency map to the corresponding feature activations of the last CNN layer, where N_f is the number of channels. Then the set of all landmark feature vectors $\mathcal{F}$ across N_i images is obtained as

$$\mathcal{F} := \bigcup_{i=1}^{N_i} \{f_i(x, y) | (x, y) \in \mathcal{M}_i\} \ . \tag{5.2.2.3}$$

Finally, the feature vectors are classified via k-means clustering of $\mathcal{F}$. The number of clusters is automatically selected by maximizing the *Silhouette Coefficient* $\frac{1}{N_i}\Sigma_{i=1}^{N_i}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, where $a(i)$ is the mean intra-cluster distance and $b(i)$ the mean nearest-cluster distance of sample i [Rou87].

Application to Image Registration

In order to examine a simple practical use of the visually salient landmarks, we consider the task of aligning the standard view images. For each plane, we use two landmarks to construct an affine transformation of optional horizontal flipping, translation, rotation and isotropic scaling.

Consider the TV plane (the generalization to the TC plane is straightforward). For image index i, let $C^i = \left(c_x^i, c_y^i\right) \in \mathbb{R}^2$ be the coordinates of the salient landmark corresponding to the CSP, and let $D^i = \left(d_x^i, d_y^i\right) \in \mathbb{R}^2$ be the coordinates of the landmark corresponding to the LV (or the cerebellum for the TC plane). Let j and k be the indices of the source and target images to be aligned. For a point $p = (p_x, p_y)$ on the images with width W_i, optional flipping of the x-coordinate is performed with the function $f : \mathbb{R} \to \mathbb{R}$ with

$$f(p_x) = \begin{cases} W_i - p_x & \text{if } \operatorname{sgn}(c_x^t - d_x^t) \neq \operatorname{sgn}(c_x^j - d_x^j) \\ p_x & \text{otherwise,} \end{cases} \tag{5.2.2.4}$$

which makes use of the fact that the horizontal ordering of the landmarks determines the orientation of the fetal head (see Fig. 5.5). Let $C^{j,f} = \left(f(c_x^j), c_y^j\right)$ and $D^{j,f} = \left(f(d_x^j), d_y^j\right)$ be the source image landmarks after optional horizontal flipping. Next, the images are aligned with the translation vector $\mathbf{t} = (t_x, t_y) = \overrightarrow{C^{j,f}C^k}$, the isotropic scaling factor $\rho = \|\overrightarrow{C^k D^k}\| / \|\overrightarrow{C^{j,f} D^{j,f}}\|$ and the rotation angle $\theta = \angle(\overrightarrow{C^{j,f} D^{j,f}}, \overrightarrow{C^k D^k})$, where the latter two operations are performed with center C^k. The resulting affine transformation $\mathcal{T}^{j,k} : \mathbb{R}^2 \to \mathbb{R}^2$ of a point $P^j = (p_x^j, p_y^j)$ on the source image to

the estimated point $\hat{P}^k = (\hat{p}_x^k, \hat{p}_y^k)$ on the target image is

$$\begin{bmatrix} \hat{p}_x^k \\ \hat{p}_y^k \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & \beta & (1-\alpha)\, c_x^k - \beta\, c_y^k \\ -\beta & \alpha & \beta\, c_x^k + (1-\alpha)\, c_y^k \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f(p_x^j) + t_x \\ p_y^j + t_y \\ 1 \end{bmatrix}, \tag{5.2.2.5}$$

where $\alpha = \rho \cdot \cos(\theta)$ and $\beta = \rho \cdot \sin(\theta)$.

We evaluate the alignment method for all unique image pairs of each plane. First, we manually annotated the CSP, LV, TCD and HC as shown in the first two rows of Fig. 5.6. Each transformation is then evaluated based on the distances of the CSP, LV and TCD centers. In addition, the alignment of the fetal skull is assessed via the distance of the ellipse centers. All distances are reported as percent of the respective HC long axis length. Three baselines are implemented: First, no alignment ("None"); second, manually aligning the head orientation via horizontal flipping ("Left-Right" (LR)); and third, manually aligning the head orientation plus subsequent intensity-based registration ("LR + Intensity"). For the latter, we compute similarity transformations via the *SimpleElastix* library [MBSK16], using the normalized cross-correlation metric with default settings and a maximum of 256 iterations per scale.

5.2.3 Results

Salient Landmark Discovery. Fig. 5.5 shows exemplary results of the salient landmark discovery method. All shown predicted saliency maps have two peaks: one at the CSP and one at the LV (TV images) or at the cerebellum (TC images). The cluster labels correctly match the landmarks across images.

Application to Image Registration. After assigning the anatomical structures to the corresponding cluster labels, 88.0% of the discovered landmarks were near the correct annotated structure (within a radius of 10% of the HC long axis). Conversely,

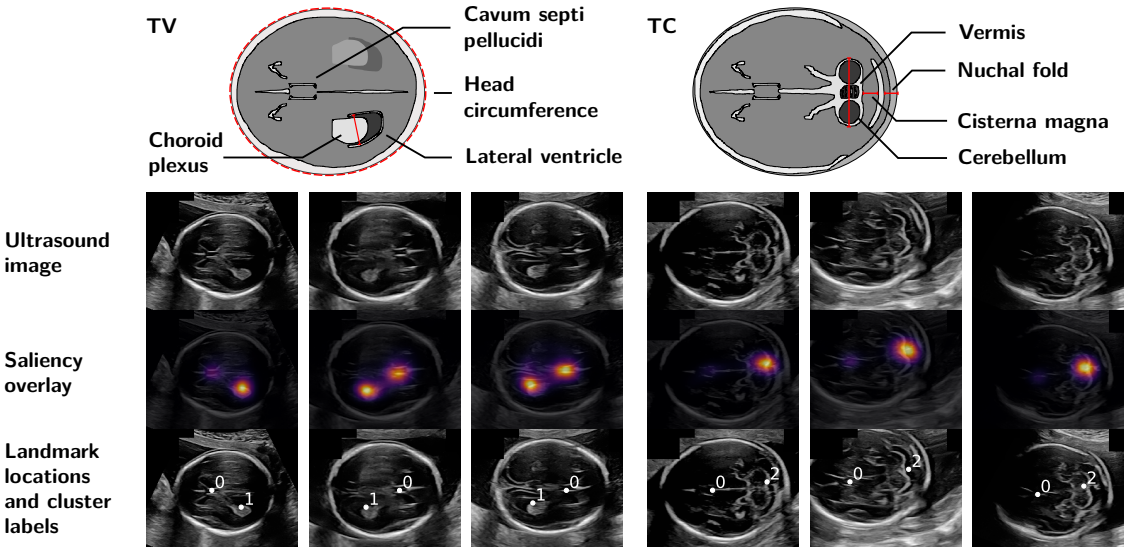

Figure 5.5: Exemplary results of the visually salient landmark discovery method. The top row illustrates the anatomy of the respective standard view, with biometric measurements highlighted in red [NHS18]. The first row of the image grid shows exemplary neurosonographic images. The second row shows an overlay of the predicted saliency map. The third row shows the discovered landmarks with cluster labels.

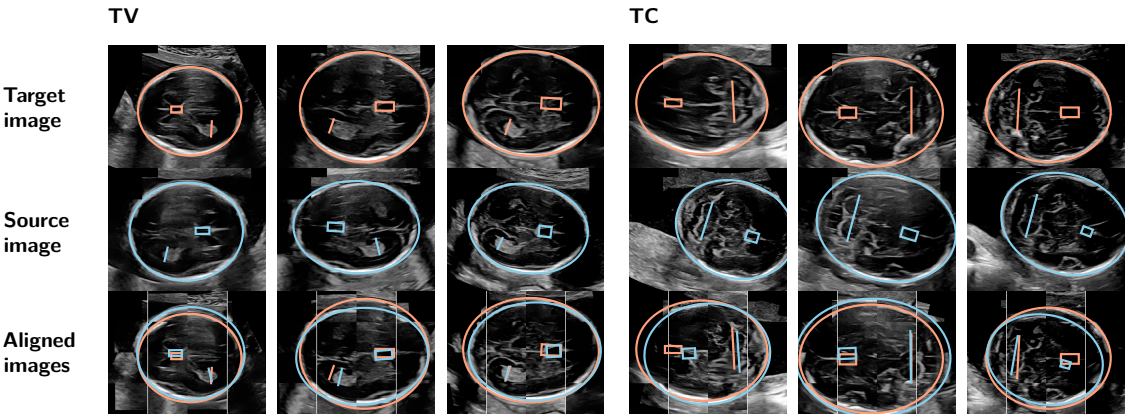

Figure 5.6: Exemplary results of the image registration via visually salient landmarks. The first and second row show target and source images with overlaid annotations of the CSP (box), LV (TV line) and TCD (TC line). The third row shows the transformed images overlaid with the transformed annotations.

77.1% of the annotated structures were near a corresponding discovered landmark. Alignment was performed for 89 (62%) TV images and 67 (54%) TC images which had all annotated structures correctly identified. Fig. 5.6 shows exemplary results and Table 5.4 shows the corresponding quantitative evaluation. The alignment

Table 5.4: Quantitative results of the image registration with visually salient landmarks and baselines. The errors for the CSP, LV, cerebellum ("Cereb.") and HC center are given in percent of the respective HC long axis length.

Plane	Alignment	CSP	LV/Cereb.	HC Center
TV	None	39.3 ± 0.3	21.9 ± 0.2	15.1 ± 0.1
	Left-Right (LR)	16.9 ± 0.1	8.9 ± 0.1	15.2 ± 0.1
	LR + Intensity	15.5 ± 0.1	8.2 ± 0.1	13.9 ± 0.1
	Salient LM	**9.8 ± 0.1**	**4.1 ± 0.0**	**7.1 ± 0.0**
TC	None	58.1 ± 0.4	24.8 ± 0.2	28.5 ± 0.2
	Left-Right (LR)	28.4 ± 0.2	12.0 ± 0.1	24.8 ± 0.1
	LR + Intensity	27.2 ± 0.2	11.6 ± 0.1	24.4 ± 0.2
	Salient LM	**10.9 ± 0.1**	**5.7 ± 0.1**	**6.7 ± 0.0**

errors are consistently lower for salient landmarks compared to the baselines.

5.2.4 Discussion and Conclusion

The results of subsection 5.2.3 show that the proposed method successfully discovers visually salient landmarks based on predicted human gaze. While the guidelines define a large set of standard plane criteria via the illustration shown in Fig. 5.5, the landmark discovery method reveals which structures the operators pay attention to in practice. Specifically, the landmarks correspond to key anatomical structures in the brain, i.e., the LV, cerebellum and CSP. The CSP itself is not part of any measurement, but it helps the sonographer assess the horizontal orientation of the fetal head and is part of both views [SAB+11]. In general, the only prerequisite for applying the landmark discovery method is a set of images from the domain of interest with recorded gaze data in order to train the saliency predictor.

For image registration, the results show that our approach can achieve good alignment without explicit supervision. The landmarks are successfully matched based on the local features of the saliency prediction CNN. The intensity-based registration performs significantly worse and only slightly above the trivial "No align" and "Flip" baselines since intensity-based alignment of ultrasound images is

inherently difficult due to noise, shadowing, artifacts and the visibility of maternal anatomies [CMG17]. The landmark discovery based on visual saliency prediction effectively ignores the irrelevant structures as a human would. A limitation is that landmark-based alignment is only possible if all necessary landmarks are detected. Moreover, the quality of alignment may be limited by the affine transform, as visible for the TC plane in Fig. 5.6, and a non-rigid transformation might yield an improvement.

In conclusion, we have presented a new method to discover visually salient anatomical landmarks by predicting human gaze. We have applied the method to fetal neurosonographic images and shown the merit for image alignment compared to intensity-based registration. Avenues for future work include a comparison of the registration performance to keypoint descriptors (e.g. SIFT), and the application of the proposed visually salient landmarks in other areas of radiology, in biological imaging and in cognitive science.

Bibliography

[CMG17] Chengqian Che, Tejas Sudharshan Mathai, and John Galeotti. Ultrasound registration: A review. *Methods*, 115:128–143, February 2017.

[CSCN18] Yifan Cai, Harshita Sharma, Pierre Chatelain, and J. Alison Noble. Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[DCS+19] Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention. In *Information Processing in Medical Imaging*, volume 11492, pages 592–604. 2019.

[DM16] Ian L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis, with Applications in R*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, September 2016.

[GFD06] Guorong Wu, Feihu Qi, and Dinggang Shen. Learning-based deformable

registration of MR brain images. *IEEE Transactions on Medical Imaging*, 25(9):1145–1157, September 2006.

[KB01] Timor Kadir and Michael Brady. Saliency, Scale and Image Description. *Int. J. Comput. Vision*, 45(2):83–105, November 2001.

[MBSK16] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 574–582, Las Vegas, NV, USA, June 2016. IEEE.

[NHS18] NHS. Fetal Anomaly Screening Programme Handbook, 2018.

[OSPD11] Yangming Ou, Aristeidis Sotiras, Nikos Paragios, and Christos Davatzikos. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. *Medical Image Analysis*, 15(4):622–639, August 2011.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[SAB+11] L. J. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. L. Johnsen, K. Kalache, K. Y. Leung, G. Malinger, H. Munoz, F. Prefumo, A. Toi, and W. Lee. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound Obstet. Gynecol.*, 37(1):116–126, 2011.

[YKPN17] Mohammad Yaqub, Brenda Kelly, Aris T. Papageorghiou, and J. Alison Noble. A Deep Learning Solution for Automatic Fetal Neurosonographic Diagnostic Plane Verification Using Clinical Standard Constraints. *Ultrasound in Medicine & Biology*, 2017.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Discovering Salient Anatomical Landmarks by Predicting Human Gaze
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Richard Droste, Pierre Chatelain, Lior Drukker, Harshita Sharma, Aris T. Papageorghiou, J. Alison Noble. "Discovering Salient Anatomical Landmarks by Predicting Human Gaze". In: IEEE International Symposium on Biomedical Imaging (ISBI) 2020.

Student Confirmation

Student Name:	Richard Droste		
Contribution to the Paper	I was the lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft.		
Signature		Date	31.05.2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. J. Alison Noble			
I confirm that the description above is accurate.			
Signature		Date	01.06.2021

This completed form should be included in the thesis, at the end of the relevant chapter.

5.3 General Saliency Representation Learning for Ultrasound Imaging

5.3.1 Introduction

In this section we present a generalization and extension of our work on "Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention" (Sec. 5.1). Since the original work considers the downstream task of standard plane detection, it is based on a CNN that is developed for image classification [HSS17] and hence does not preserve spatial information. In order to retain sufficient spatial information during pre-training with saliency prediction, dilated convolutions were introduced during the pre-training phase and removed during the downstream classification stage.

Yet, many important ultrasound image analysis tasks such as segmenting anatomical structures or localizing/detecting key anatomical landmarks are crucial for applications like quality assurance [WCL+17] and biometric measurements [JPK+18]. Therefore, we generalize the original framework such that it is suitable for these applications. At the same time, the generalized framework remains applicable to classification tasks and is shown to outperform the previous method on the original standard plane detection downstream task (see Sec. 5.3.3). In addition, the new network is trained and evaluated on data from both the second and third trimester of pregnancy in contrast to second trimester data only in the prior work, making the new network a versatile feature extractor for later gestations. We term our method "*SalT-Net*", which is short for Saliency Transfer Network.

The generalized framework with improved performance is achieved through three novelties. (1) First, the design of a CNN with optional dilated layers is replaced with a U-Net style architecture, where a CNN encoder is followed by a decoder with skip connections. As a result, the decoder can simply be removed for

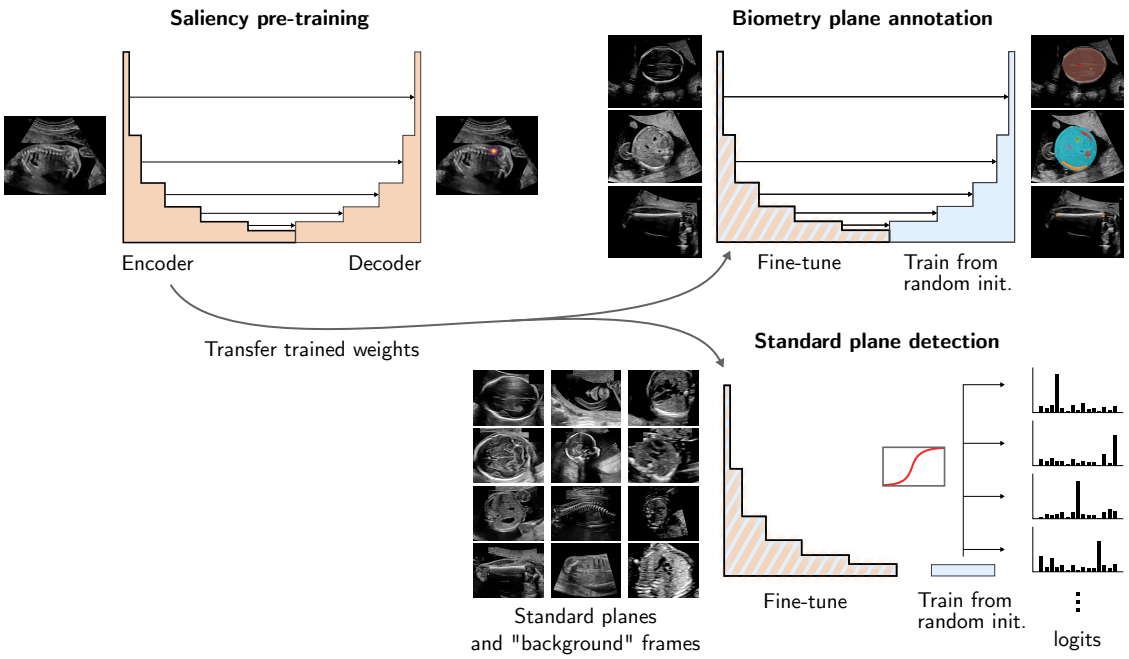

Figure 5.7: Framework of generalized saliency transfer learning (*SalT-Net*). The two stages of training are (1) pre-training for visual saliency prediction and (2) transfer to the downstream tasks of biometry plane annotation and standard plane detection.

classification downstream tasks or can be re-trained for segmentation, localization and detection downstream tasks. (2) Inspired by recent advances in transfer learning [KBZ+20], we replace the original CNN with a simplified architecture that has no batch-normalization layers. Thus, the new architecture is independent of batch feature-statistics which may differ between pre-training and downstream tasks. Also, it allows for arbitrarily small batch sizes which is useful for training with very few samples. (3) Realizing that the pre-training dataset size correlates with the downstream performance [KBZ+20], with no known upper bound and even if label noise is introduced, we increase the number of ultrasound scans for pre-training ca. nine-fold and train with all scans of the second and third trimester.

5.3.2 Method

Overview. The method is illustrated in Fig. 5.7 and consists of two stages. First, general ultrasound image representations (i.e., features) are learned by training

a U-net style network, *SalT-Net*, to predict the operator gaze point distribution on video frames that are sampled from *key segments* (as defined below) of a large number of routine obstetric ultrasound scans. No manual annotations are required for this stage and the gaze points are acquired automatically alongside the routine scanning. Second, the learned representations are transferred to downstream tasks through fine-tuning of the network. Here, we consider two downstream tasks: biometry plane annotation and standard plane detection.

SalT-Net architecture

Recently, Kolesnikov et al. [KBZ+20] achieved significant improvements of the transfer learning performance of convolutional neural networks (CNNs) for classification, especially for downstream tasks with few samples. The key insight was that the downstream performance increases with pre-training dataset size and model size. the authors demonstrated this relationship based on a standard ResNet-50 model [HZRS16] with two simple modifications: The Batch Normalization (BN) [IS15] layers of ResNet-50 are replaced with Group Normalization (GN) [WH18], making the network independent of the mini-batch size during stochastic gradient descent and allowing arbitrarily small batch sizes for few-sample training. To recover the performance of BN for larger batch sizes, the authors found that it is necessary to combine GN with Weight Standardization (WS) [QWL+20]. Here, we extend this work for image classification into a generalized architecture that is suitable for pre-training with visual saliency prediction and for downstream tasks including both classification and segmentation/detection/localization. Specifically, we add a decoder network which consists of the same ResNet-style blocks with GN and WS, with nearest-neighbor up-sampling after each block. The upsampled features after each decoder block are concatenated with features of the ResNet encoder, as introduced with the U-Net architecture [RFB15]. This encoder-decoder design allows the SalT-Net to be used for tasks that require the preservation

of spatial information and, by simply removing the decoder, tasks that depend on semantic information only.

Saliency Pre-Training.

For saliency pre-training we mostly follow the procedure proposed in [DCS+19] (Sec. 5.1), with the key difference that we created a significantly larger version of the initial dataset and sample only from key segments (defined below). First, the number of anomaly scans used is increased from 135 to 626 by extending the original acquisition interval of May to October 2018 until March 2020. Next, 374 third trimester growth scans are added to the dataset. This also serves the purpose of obtaining more general features that are useful for second and third trimester downstream tasks alike. Finally, we use five-fold instead of three-fold cross validation, increasing the respective training fold size by 20%. Overall this results in a new dataset of exactly 1000 scans and an increase in the number of scans per training fold by a factor of 8.89. In order to keep the resulting number of training video frames computationally manageable and, the video frame sampling procedure is adjusted. In our original work, every 8^{th} video frame is sampled from the entire set of live (i.e., non-frozen) B-mode frames. Applying this same method to our new dataset would result in over two million video frames per training fold. Our aim is to lower this number by reducing the amount of redundant and irrelevant data. We attempt this by only including *key segments*, i.e., the time-windows before key events, where a key event is the operator either saving or freezing the live B-mode image. In addition, the sampling interval is set to 10 frames for anomaly scans and 5 frames for growth scans in order to balance the smaller number of growth scans. We set a key segment length of 20 s which results in 822 522 frames in total.

The training procedure is performed analogously to [DCS+19]. The weights are optimized by minimizing the Kullback-Leibler Divergence (KLD) between the SalT-Net output and the target saliency maps. The network output consists of a

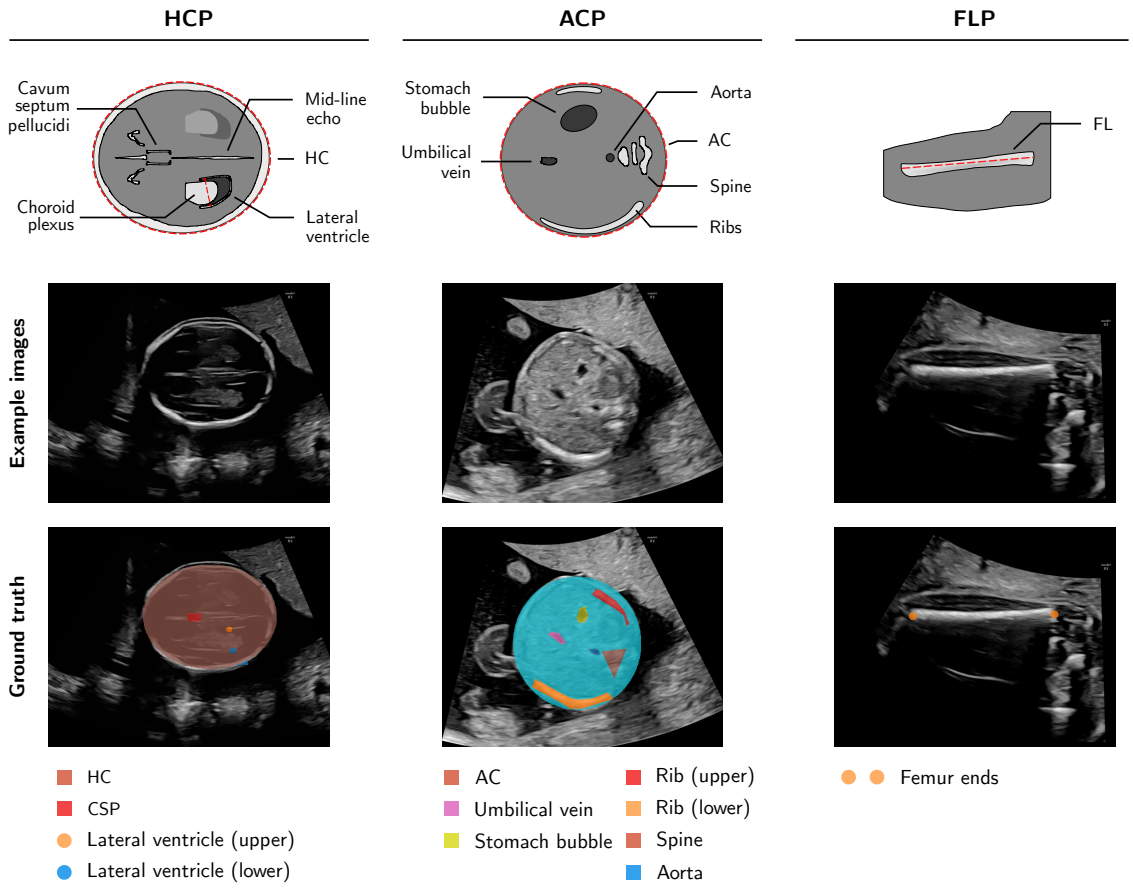

Figure 5.8: Illustration of the biometry plane annotation task.

single channel that is normalized via a softmax function across all output pixels. The target saliency maps are normalized 2D Gaussian blobs around the recorded gaze points and with a standard deviation of approx. 1° visual angle. The training data are augmented through random horizontal flipping, cropping and random variations of contrast, gamma and brightness. The data are then standardized to zero-mean and unit-standard deviation and downscaled to 224×288 pixels. The weights are optimized via stochastic gradient descent (SGD) with momentum of 0.9. The same processing and optimizer are used for the downstream evaluation tasks.

Automatic Biometry Plane Annotation

We choose *automatic biometry plane annotation* as a varied and difficult downstream task that has wide applications and covers second and third trimester annotation.

The task is illustrated in Fig. 5.8 and consists of the head circumference plane (HCP), the abdominal circumference plane (ACP) and the femur length plane (FLP). The annotation of the HCP includes both measurements: the head circumference (HC) and the ventricular atrium (VA) of the lateral ventricle (LV). Moreover, we include the cavum septi pellucidi (CSP) because it its appearance determines the correctness of the HCP view, which makes it the most fixated HCP structure [DDNP20]. For the ACP, all structures that are relevant according to the FASP guidelines are annotated: abdominal circumference (AC), umbilical vein (UV), stomach bubble (SB), upper and lower rib (separately), spine and aorta. The FLP is annotated with the two measurement caliper points of the femur length (FL) measurement.

The challenge arises to train a network that predicts x,y-coordinates for the points (VA and FL) and binary masks for the remaining anatomies. To address this, we generate a target map with a Gaussian blob for each point which allows us to train all classes simultaneously with a standard segmentation loss. Specifically, each annotated point is taken as the center of a 2D Gaussian with diagonal covariance matrix and standard deviation empirically set to 4.82 pixels. The target maps of the two FL points are merged since it is not always possible to distinguish between the proximal and distal ends of the femur. All other target maps, including the HC and AC measurements, are represented with binary masks. The number of network output channels is set to the respective number of resulting target maps (HCP: 4, ACP: 7, FLP: 1). We formulate segmentation as *multi-label* dense classification, meaning that each pixel can be classified as any subset of the relevant classes. Therefore, the output is passed through a sigmoid function and then per-class intersection over union losses [RW16] that are summed. The pre-trained SalT-Net encoder weights are loaded, a randomly initialized decoder with the desired number of output channels is added, and the network is fine-tuned for a fixed number of epochs (see 5.3.2 for the hyperparameters). Training the entire network from random initialization is used as a baseline.

For testing, the points of the VA and FL are retrieved as the maxima of the respective output maps. The two points for the FL are extracted from a single output map as the global maximum and the largest local maximum with a minimum distance of 20 pixels to the global maximum. The output maps of the remaining classes are thresholded at a value of 0.5.

We selected the five sonographers with the largest number of available scans and an experienced fetal medicine specialist annotated 20 standard biometry planes per sonographer, per trimester and per plane, resulting in $20 \times 5 \times 2 = 200$ samples per plane. The annotator discarded 5, 1, 3 samples of the HCP, ACP and FLP plane respectively as unsuitable since they did not contain all required anatomies, leaving 195, 199 and 197 samples. With five-fold cross-validation this left an average of just under 160 samples per training fold. We also explored the impact of the training dataset size on the network's performance and by reducing the number of training samples per trimester to 1, 4 and 16. To further increase the statistical power of our experiments, we re-ran each cross-validation fold with four different random seeds including four different subsets for the reduced training set sizes. The cross-validation and random seed variation resulted in $5 \times 4 = 20$ training runs per training set size and plane, $20 \times 4 = 80$ training runs per plane, and $80 \times 3 = 160$ training runs in total.

Standard Plane Detection

The standard plane detection (SPD) downstream task is conducted with the same dataset and training procedure as [DCS+19]. The only difference is that we use the same five-fold cross-validation split as is used for pre-training and automatic biometry plane annotation, instead of the original three-fold cross validation. To maintain comparability to prior work, the SPD training fold size is kept constant at 90 scans. For fine-tuning, the same data augmentations, class-balanced sampling scheme and cross-entropy loss are used. To adapt SalT-Net for classification, the

Table 5.5: Training hyperparameters of SalT-Net tasks and baseline. Abbreviations as follows. *LR*: Learning rate. *LR Decay*: epochs at which the learning rate (LR) is decayed by a factor of 10, either in absolute epochs or relative to the total number of epochs. *WD*: Weight decay. *BS*: Batch size. *SalT-Enc.* and *SalT-Dec.* SalT-Net encoder and decoder.

Task	Model	Plane	LR	Epochs	LR Decay	WD	BS
Pre-training	SalT-Net	*N/A*	0.002	7	(3, 5)	10^{-4}	24
Biometry plane annotation	SalT-Net	HCP	0.04	100	(50%, 90%)	10^{-4}	8
		ACP	0.1	150	(50%, 90%)	10^{-4}	8
		FLP	0.04	60	(50%, 90%)	10^{-4}	8
	Baseline	HCP	0.04	150	(50%, 90%)	10^{-3}	8
		ACP	0.1	150	(50%, 90%)	10^{-3}	8
		FLP	0.04	100	(50%, 90%)	10^{-3}	8
Standard plane detection	SalT-Enc. SalT-Dec.	*N/A*	0.01 0.004	60	(30, 54)	10^{-4}	16

decoder network is removed and a classification head equivalent to that used in the original work [DCS+19] is added and trained from random initialization. As before, we use training from random initialization as a baseline. As additional benchmark references we include fine-tuning SonoNet [BKM+17] (SonoNet-FT) and fine-tuning ResNet-50 from ImageNet-21k (14.2M images) pre-trained weights, the current state-of-the-art in general transfer learning for image classification.

Hyperparameters

The hyperparameters were tuned via grid-search for pre-training, each of the downstream tasks and the baselines. The results are shown in Table 5.5. We found five notable observations. (1) The larger number of epochs for the ACP compared to HCP and FLP was caused mostly by the slow convergence of the segmentation of the aorta due to its small size. (2) The lower learning rate for HCP and FLP is obtained since the loss component of the Gaussian target maps otherwise caused training instability. (3) Training from scratch demanded more training epochs to converge for HCP and FLP, which lead to larger weight decay to mitigate overfitting. (4) For SPD with SalT-Net, the learning rate for the decoder is reduced in order to

stabilize fine-tuning. We found that it was necessary to additionally introduce a linear learning rate warm-up of 20 epochs for training stability. (5) The optimal LR and mini-batch size for SalT-Net SPD are the same as in [DCS+19].

For the biometry plane annotation with reduced training sets, the number of epochs is increased such that the total number of batches per epochs is 25% of that for the full training set.

Evaluation Metrics

Segmented anatomies are evaluated via the intersection over union (IoU) with a classification threshold of 0.5. Point-based annotations are evaluated by extracting the predicted points as the maxima of the corresponding prediction maps and computing the delta of the estimated biometric measurement value relative and the true measurement value. Statistically significance of the differences between the SalT-Net based architecture and baselines is determined via two-sided Student's t-tests with a significance level of 0.05.

5.3.3 Results

Automatic Biometry Plane Annotation

The quantitative results for the SalT-Net downstream task of automatic biometry plane annotation are shown in Fig. 5.9. For the 1 and 4 samples per trimester settings of the ACP, the mean IoU of the saliency pre-trained SalT-Net is higher than that of the baseline trained from random initialization for all anatomies except the aorta. Statistically significant improvements are observed for the AC, UV, ribs, spine and the average across all anatomies. For the 16 samples per trimester setting, statistically significant improvements are observed for AC, UV and SB. Regarding the annotation of the HCP, the HC annotation is significantly improved for the 1 and 16 sample per trimester settings and the CSP and the LV annotations for the 4

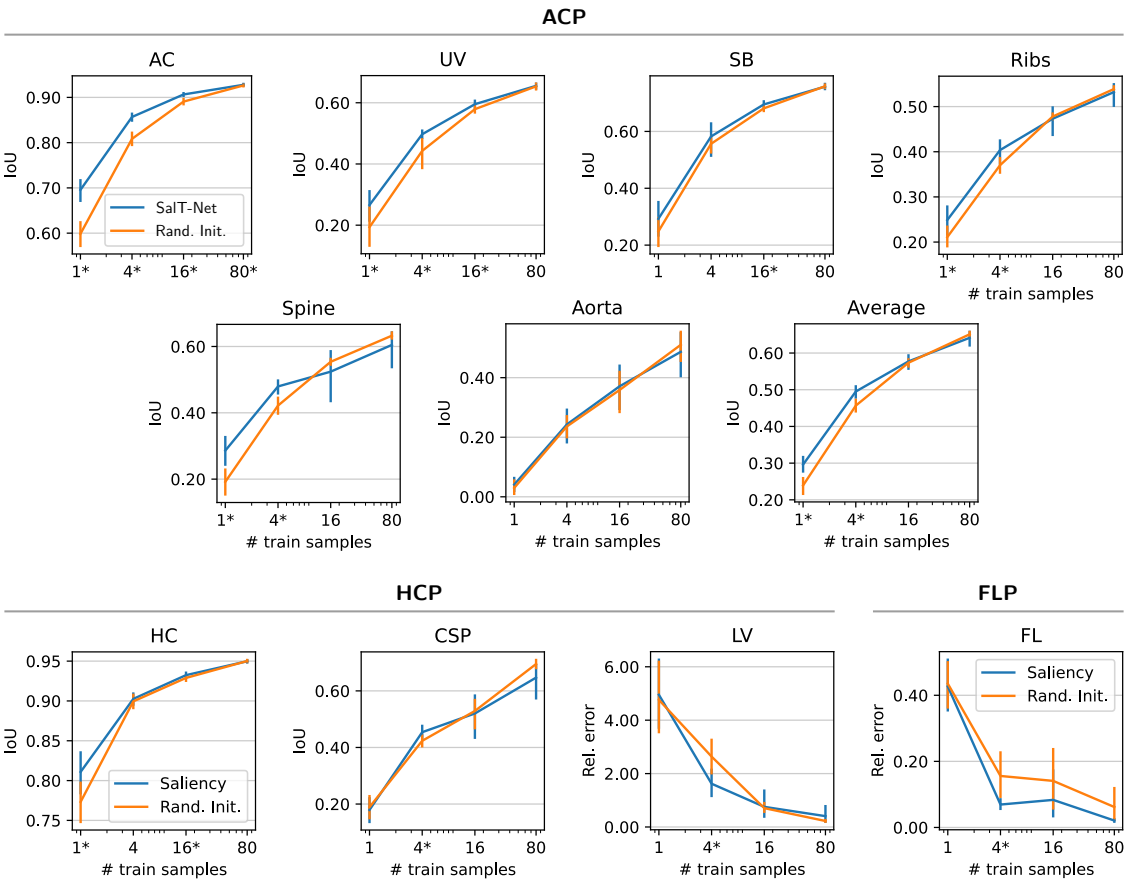

Figure 5.9: Quantitative results for the automatic biometry plane annotation downstream task. "# train samples" denotes the number of training samples per trimester. Error bars are 90% confidence intervals. Statistically significant differences of mean scores are denoted by an asterisk after the corresponding number of training samples.

sample per trimester setting. The relative error of the FL measurement is improved for all settings of 4 or more samples per trimester, with statistical significance at 4 samples per trimester. Overall, we observe a significantly better mean score of SalT-Net compared to the baseline for 19/44 settings and anatomies versus 0/44 vice versa. For the remaining anatomies the difference was not significant.

It is notable that with a single sample per trimester, IoUs of 69.6% and 81.1% are achieved for the AC and HC segmentation, respectively. For the FL, 4 samples per trimester suffice for a relative measurement error of 6.9% (vs. 15.6%), dropping to 2.1% (vs. 6.2%) at 80 samples per trimester. Compared to the FL, the relative errors for the LV are substantially larger which can be explained by the measurement

Table 5.6: Quantitative results for the standard plane detection downstream task (mean ± standard deviation [%]). The best score among the models without pre-training with manual labels (columns 2-6) is marked in bold. Scores of the fine-tuned ImageNet-21k pre-trained ResNet-50 and SonoNet that exceed these models are marked in bold as well. The literature SonoNet scores are given in parenthesis.

Metric	Rand. Init.	Video [JDD+20]	Speech [JCA+20]	Saliency [DCS+19]	Saliency (SalT-Net)	ImageNet -21k	SonoNet-FT (Lit. value)
Precision	71.5 ± 4.4	75.8 ± 1.9	72.7 ± 1.8	79.5 ± 1.7	**82.3** ± 2.4	**83.7** ± 2.2	**82.3** ± 1.3 (81)
Recall	67.2 ± 2.7	76.4 ± 2.7	73.3 ± 2.4	75.1 ± 3.4	**80.4** ± 3.8	**83.4** ± 4.0	**87.3** ± 1.1 (86)
F1-score	68.2 ± 3.5	75.7 ± 2.0	72.6 ± 1.7	76.6 ± 2.6	**80.6** ± 2.9	**82.8** ± 2.9	**84.5** ± 0.9 (83)

Table 5.7: Quantitative results per standard plane (mean ± standard deviation [%]). Score formatting is equivalent to Table 5.6.

Plane	Rand. Init.	Saliency [DCS+19]	Saliency (SalT-Net)	Δ(SalT-Net, [DCS+19]) ↓	ImageNet -21k	SonoNet-FT (Lit. value)
RVOT	38.9 ± 7.9	58.7 ± 2.7	**65.5** ± 9.3	6.8	**67.2** ± 9.0	**71.2** ± 2.8
Lips	83.3 ± 2.6	85.6 ± 4.5	**92.1** ± 2.3	6.5	92.0 ± 2.5	**92.9** ± 0.8
LVOT	28.0 ±18.5	48.6 ± 3.3	**54.7** ± 8.6	6.1	**65.8** ± 9.7	**69.9** ± 5.3
Abdominal	66.9 ±10.7	72.9 ± 2.9	**78.9** ± 4.7	6.0	78.2 ± 5.7	**85.0** ± 1.4
Kidneys	71.8 ±16.8	84.7 ± 6.3	**90.6** ±10.6	5.9	87.1 ± 9.6	81.0 ± 5.0
Profile	78.6 ± 8.8	87.2 ± 7.5	**92.0** ± 3.0	4.8	90.0 ± 3.6	88.1 ± 4.5
4CH	43.9 ± 9.5	57.3 ±10.8	**61.4** ± 3.8	4.1	**76.9** ±10.0	**75.7** ± 9.1
Femur	85.6 ± 7.8	87.6 ± 1.3	**90.2** ± 4.3	2.6	**90.9** ± 5.2	89.5 ± 1.8
3VT	45.7 ± 9.4	58.3 ± 7.1	**60.1** ±10.9	1.8	**69.8** ±14.6	**77.9** ± 1.6
Spine (cor.)	69.5 ± 7.9	79.0 ± 3.7	**80.5** ± 8.0	1.5	74.2 ±10.2	**90.3** ± 4.9
Spine (sag.)	85.6 ± 3.9	**89.1** ± 2.1	89.0 ± 4.0	-0.1	**92.0** ± 7.1	**91.6** ± 2.5
Brain (Tv.)	85.1 ± 6.1	**92.9** ± 5.0	91.6 ± 6.7	-1.3	90.7 ± 6.3	92.1 ± 4.5
Brain (Cb.)	86.1 ± 3.5	**93.7** ± 4.6	90.9 ± 3.6	-2.8	**93.8** ± 2.7	92.8 ± 1.1
Background	85.1 ± 1.2	89.0 ± 0.4	**90.3** ± 1.5	1.3	90.2 ± 1.3	90.3 ± 0.4

RVOT: right ventricular outflow tract; LVOT: left ventricular outflow tract; 4CH: four chamber view; 3VT: three vessel and trachea view; Brain (Cb.): brain cerebellum suboccipitobregmatic plane; Brain (Tv.): brain transventricular plane; Cor.: Coronal plane; Sag.: Sagittal plane.

being smaller, resulting in higher errors for the same pixel offset.

Standard Plane Detection

For the standard plane detection task we observe significant improvements compared to our previous saliency-based work [DCS+19] as well as other ultrasound image representation learning methods [JDD+20, JCA+20] across all metrics (see Table 5.6). The precision is now on par with SonoNet-FT at 82.3%, and the recall jumped

by 5.3%, halving the difference to SonoNet-FT in terms of F1-score. Looking at each class individually (Table 5.7), the F1-scores of [DCS+19] are improved upon for 11/14 classes. We also observe that the fine-tuned ImageNet-21k pre-trained ResNet-50 [KBZ+20] is a strong competitor to the SonoNet-FT benchmark reference, outperforming SonoNet-FT for 6/14 classes. SalT-Net outperforms this ImageNet-21k benchmark for 7/14 classes and the SonoNet-FT benchmark for 6/14 classes. Interestingly, both the ImageNet-21k pre-trained model and SonoNet-FT outperform SalT-Net for the cardiac classes (RVOT, LVOT, 4CH and 3VT). The difference in F1-score between SalT-Net and SonoNet-FT shrinks from 3.9% to 0.9% when these classes are excluded.

Training from random initialization is outperformed by all other methods for all classes.

5.3.4 Discussion and Conclusion

Overall the results show that the proposed method generalizes our original representation learning work to spatial downstream tasks and additionally outperforms the original work for classification. We have shown this for the spatial task of automatically annotating biometry plane images and for the classification task of standard plane detection.

For segmentation, the improvement of SalT-Net over the baseline is especially pronounced for small numbers of training samples (1, 4 and 16 samples per trimester). This confirms that the feature representations which are learned by predicting sonographer gaze serve as a strong prior for this task. Small sample sizes occur frequently in medical imaging, through the scarcity of annotator time, acquisition resources and/or natural scarcity such as rare diseases. The fact that the scores of SalT-Net and the baseline converge for larger sample sizes (here 80 samples per trimester) is to be expected since the features can be learned effectively from

sufficient annotated data.

For the 80 samples per trimester setting, we obtain IoU scores of 97.4% and 96.2% for the HC and AC segmentation. This surpasses the results of prior work by Wu et al. [WXL+17] that reports IoU scores of 94.1% and 86.5% for the HC and AC, respectively, with a U-Net based model, although it is trained with 900 (HC) and 688 (AC) samples. Moreover, our evaluation includes the segmentation of additional structures and images of both the second and third trimester. The authors boost the scores to 96.9% and 96.3% with a cascade of several neural networks and complex post-processing, but this model far surpasses ours in complexity.

The adaptation of our original representation learning framework for image annotation came at no cost to the standard plane detection performance. On the contrary, the new method further closes the gap to the SonoNet [BKM+17] model which is trained with over 22k standard plane images compared to an average of 753 standard plane images for SalT-Net. In fact, the average scores of SalT-Net are almost on par with SonoNet except for the cardiac planes. At the same time the network architecture is simplified compared to [DCS+19], building upon the standard U-Net encoder-decoder model instead of dilated convolutions that are inserted and removed for pre-training and fine-tuning. Prior work has explored the use of spatio-temporal perturbations [JDD+20] and sonographer speech [JCA+20] as supervision signals for ultrasound image representation learning, but our results confirm that sonographer gaze induces the strongest semantic features.

This has four main implications for the development of automated image analysis algorithms for clinical ultrasound. (1) The image analysis tasks included in this paper, automatic biometry plane annotation and standard plane detection, have direct clinical applications. Automated biometry plane annotation can be used for quality assurance, in real time during scanning and retrospectively, or for automated biometric measurements, which can mitigate known issues of manual measurements [DDC+20a]. Standard plane detection can also be used to provide real-time feedback

to operators during standard plane acquisition, potentially lowering the required level of expertise. (2) we show that the pre-trained SalT-Net can be fine-tuned on diverse downstream tasks quickly and with few labeled data. This can facilitate the engineering of any ultrasound image analysis application, likely including those not explicitly included in this paper, by reducing the annotation cost by orders of magnitude and enabling faster and cheaper development cycles. (3) The use of gaze-tracking data for large-scale pre-training makes the entire proposed framework accessible, not just the fine-tuning aspect. In contrast to large-scale manual annotations, the pre-training data can be acquired by attaching a low-cost gaze-tracking device to an ultrasound machine monitor. This is important since separate pre-trained models may be needed for ultrasound images of different appearance, e.g. from low-cost probes. (4) The proposed SalT-Net has learned to effectively predict the gaze of experienced sonographers on ultrasound video frames, which could be used to guide the visual attention of novice operators via a real-time visual overlay.

The strong performance of the same architecture pre-trained with the over 14 million manually labeled natural images from ImageNet-21k is in line with prior work on the importance of the pre-training dataset size [KBZ+20]. For future work it would therefore be interesting to investigate how the model performance scales with increasing number of saliency pre-training samples. The number of samples could be increased with the currently available dataset by (1) increasing the video frame sampling rate, (2) by using the entire live B-mode data instead of time windows, (4) adding data from other tasks such as Doppler imaging, and/or (3) using first trimester data in addition to the second and third trimester. Further, incorporating temporal information in the saliency pre-training may improve the performance on the cardiac planes with dynamically changing appearance. Finally, additional downstream tasks such as ultrasound probe movement guidance [DDPN20] or skills assessment [WDJ+20] can be considered.

In conclusion, we have introduced a general framework for learning image feature representations from gaze-tracking data with application to obstetric ultrasound scanning. The framework extends prior work for a wider range of downstream tasks while achieving more simplicity and higher performance on the original downstream task.

6

Discussion and Conclusion

6.1 Discussion

In this thesis we presented seven research items, linked by the goals of the PULSE study and the use of gaze-tracking to advance ultrasound scanning. Here, we provide a brief common discussion to relate the findings and proposed methods to the goals of the PULSE project and the broader research area.

A first central goal of the PULSE study is to develop a significantly deeper understanding of clinical ultrasound scanning than currently exists. A key aspect to this understanding is identifying potential weaknesses in the scanning process. In chapter 3 we presented two studies that uncover such weaknesses based on ultrasound video and eye-tracking data: potential bias in the measurement process in Sec. 3.1 [DDC+20a] and infrequent monitoring of safety parameters in Sec. 3.2 [DDC+20b]. Each of these shortcomings may cause a negative impact on outcomes

for mothers and newborns as long as they are not addressed.

Fetal growth restriction (FGR) is associated with increased risks for severe consequences such as stillbirth, neonatal death and neuro-developmental impairment. Consequently, detecting FGR via biometric measurements is a key component of obstetric ultrasound scanning. It is achieved by identifying small-for-gestational-age (SGA) fetuses with measurement values below the 10^{th} percentile. If FGR is detected, the fetus can be monitored closely and delivery can be induced in order to avoid fetal compromise or death [MBE+15]. However, in our study we found that operators adjust the biometric measurements towards the expected, normal-for-gestational-age (NGA) measurement. As a result, the classification into SGA and NGA is altered through the measurement adjustment in as much as 17% of scans. Classifying an SGA fetus as NGA may prevent a fetus with FGR from being monitored and risk adverse perinatal outcomes [MET98]. Classifying an NGA as SGA can have similarly negative consequences such as increased perinatal morbidity due to unnecessary intervention [MET98].

Equally important is the finding from Sec. 3.2 [DDC+20b] that sonographers rarely monitor the safety indicators that are displayed on the ultrasound machine's graphical interface. Ultrasound scanning is generally considered safe, but the safety is contingent on the adherence to the safety guidelines, which is the operator's responsibility. In addition to the guidelines which set time-limits for different signal intensities, the operators are instructed to always set the intensity as low as reasonably achievable (ALARA principle) [KJM+20, Tom06]. In our study we analyzed the scans of 637 women and found that for 610 of them (95.8%) the safety indices were never gazed at. While we also find that the time-limits specified in the guidelines are adhered to, the lacking monitoring of the indices by the operators during scanning brings into question if the ALARA principle is followed in practice.

The discoveries of our two clinical studies discussed above were enabled by the terabytes of full-length video data and eye-tracking of the PULSE study,

combined with large-scale automated data analyses to extract insights from millions of unlabeled video frames and gaze points. The studies shed further light onto the known limitations of obstetric ultrasound imaging, adding to known concerns about the graphical interface [Mar05, MBE+15]. Our methodology opens up many avenues for future research and the advancement of ultrasound imaging, some of which are mentioned in Sec. 7.1.

A second key goal of the PULSE study is to enable a new generation of assistive machine learning algorithms based on sonographer expertise which is reflected in the recorded operator-machine interaction. In Sec. 4.1 [DCS+20] we demonstrated that visual saliency prediction can capture the mechanisms of sonographer visual attention, a key aspect of this interaction. Moreover, we showed that existing visual saliency prediction approaches are outperformed by a new model that is inspired by cognitive science. Sonographer gaze is predicted by past, present and future visual information, suggesting that the sonographers are looking where they expect structures of interest to appear. Our work constitutes the first dedicated visual saliency model for ultrasound video and deploying such a model could show any operator immediately which structures the expert would pay attention to. Our ideas have already been built upon in subsequent work which employs bidirectional recurrent networks for ultrasound video saliency prediction and simultaneous anatomy classification [CDS+20].

While we focus on ultrasound video in Sec. 4.1, ultrasound data consists of an alternation of video and image (frozen video) data. To address the issue of jointly modeling both modalities, we proposed the first unified model, UNISAL (Sec. 4.2 [DJN20]. In order to compare it to the substantial amount of existing visual saliency models for image and for video saliency prediction, we evaluated the model on five computer vision benchmarks and obtained state-of-the-art performance despite the model's simplicity and lower number of parameters. The key to unified modeling of both modalities were novel methods of *domain adaptation*. This family of machine

learning algorithms has been successfully applied to adapting segmentation models for different MRI scanners and acquisition protocols [KBL+17], and for different modalities such as MRI to CT [DOC+18]. Our proposed domain adaptation between image and video data extends the capabilities of this powerful set of methods.

A common theme of our two contributions to the field of saliency prediction is that high accuracy was achieved through models that are designed based on prior knowledge and an understanding of particularities of the data. We have shown that the use of generic deep learning models, i.e., single-frame CNNs or forward-RNNs, are outperformed by models that are designed with the data acquisition and cognitive processes of the observers in mind. This is an approach that opens up many exciting research avenues of applying insights from traditional ("*non-deep*") saliency models, psychology and cognitive science for the advancement of saliency prediction.

In chapter 5 we showed that by training neural networks for ultrasound saliency prediction, they automatically learn features that are directly useful for ultrasound image analysis. In particular, we presented methods for standard plane detection, discovery and detection of salient landmarks and localization and segmentation of anatomical structures, all of which outperformed existing methods without or with limited labeled training data.

Sec. 5.1 is the first work that systematically explores visual saliency prediction for image representation learning. We found the prediction of the 2D gaze distribution, the most common setting for visual saliency, to be superior to the regression of a point-estimate of the gaze coordinates. Moreover, we provided insights into the anatomical classes that benefit from the representation learning step, the predictive power of different network layers, and the topology of the feature space. We believe that gaze as a means of learning strong features is especially important for ultrasound imaging since most of the recently proposed self-supervised learning methods for natural images do not generalize to ultrasound imaging. These self-supervised methods heavily depend on the recognition of spatial image transformations

("*augmentations*") [DSB15, GSK18], which induces semantic features for natural images, but can be a trivial task for ultrasound images (e.g. recognizing image rotation). The transformations that remain for self-supervised ultrasound image representation learning like shuffling video frames empirically perform worse than saliency-based learning [JDD+20]. The same applies to other automatically acquired supervision signals like raw sonographer speech [JCA+20].

With the concept of *visually salient landmarks*, introduced in Sec. 5.2, we extend the idea of saliency-based feature learning to the detection of anatomical landmarks, which makes it applicable to downstream tasks like registration. Besides providing a basis for image analysis tasks, salient anatomical landmarks also enable insights into the expert operator's visual attention and cognition. In particular, our method revealed which anatomical landmarks are attended to in practice, and which ones are perceived only with peripheral vision.

Finally, in Sec. 5.3 we extended the saliency-based ultrasound image representation learning method of sections 5.1 and 5.2 into a generalized, simplified and more powerful framework that covers all downstream tasks, including purely semantic tasks like standard plane detection and spatial tasks such as segmentation.

Overall, we have shown how gaze-tracking data can advance the tasks of standard plane detection, landmark detection and semantic segmentation, in addition to introducing the idea of discovering salient landmarks. Due to the large body of literature regarding accurate saliency prediction [Bor18, BI13] and the comparatively small existing work on leveraging saliency prediction to understand gaze patterns and for image analysis (see Sec. 2.1), we believe that our work can inspire further research in the latter two areas.

6.2 Limitations

The most important limitations of each study of this thesis are presented in the individual discussions of the respective sections. Overall, there are some limitations that concern several of the presented studies.

First, the PULSE dataset is acquired with one ultrasound machine so that operator-machine interactions can be studied without uncontrolled variations of the ultrasound machine or environment. As a result, some of the results of sections 3.1 and 3.2 may need to be verified for different graphical user interfaces, hospitals or locations. However, our findings are consistent across the 16 to 17 sonographers that were available for our analyses, suggesting good generalization.

The fact that all data is acquired with one ultrasound machine may also impact the results of sections 5.1, 5.2 and 5.3 on ultrasound image representation learning. However, in all three studies our contribution is to demonstrate benefit of visual saliency-based models compared to existing approaches. While the absolute values would likely change for ultrasound data from other vendors or locations, there is no reason to assume that the observed relative improvements would change. In Sec. 5.1 and 5.3, where we also compare to a gold-standard method which is trained with a different dataset [BKM+17], we achieve a fair comparison by fine-tuning it on the PULSE dataset, increasing its performance above the values provided by the authors.

A third aspect to consider is that all proposed methods depend on gaze-tracking data, which is not generally available in typical ultrasound scanning setups. However, this is less a limitation than it is by design since the overarching goal is to advance ultrasound imaging by studying sonographer perception. Moreover, gaze-tracking devices have become very affordable, with the device used in our study being available for little over one hundred British Pounds. This is only a fraction of the cost of even the least expensive ultrasound machines. What is more, algorithms have been developed to track gaze with front-facing smartphone cameras [KKKK16], which may

be relevant for increasingly popular portable ultrasound probes. Finally, the image analysis algorithms described in chapter 5 only require gaze data during the model training phase and can therefore be deployed on any device without a gaze tracker.

6.3 Conclusion

In conclusion, we have shown that the capturing of sonographer gaze patterns can advance the understanding of ultrasound imaging and enable new, powerful ultrasound image analysis methods. We have presented seven research contributions, structured into the three chapters, spanning the analysis of sonographer gaze patterns, capturing gaze patterns through visual saliency modeling, and deployment of saliency models for various ultrasound image analysis tasks. We have exposed measurement bias and lacking attention to safety indices as shortcomings of routine ultrasound scanning, introduced the first dedicated ultrasound visual saliency model, set a new state-of-the-art for image and video saliency prediction with the first unified model, established saliency modeling as a feature representation learning method that boosts performance for limited amounts of training data, and introduced the concept of visually salient anatomical landmarks. These contributions form a significant step towards the goals of the PULSE project of developing a better understanding of obstetric ultrasound scanning, capturing sonographer expertise through machine learning models and employing these models for better assistive algorithms.

7

Outlook

Contents

7.1 Analysis of Sonographer Gaze Patterns

Regarding the "*big data*" analysis of sonographer gaze-tracking data in order to identify limitations of the current procedures and graphical interfaces, as presented in chapter 3, there is vast untapped research potential. First, the interaction with machine parameters that determine the ultrasound image quality such as gain, focus, depth, fan angle, zoom *etc.* could be examined. For example, sonographers can set the ultrasound beam focus to a certain depth in order to enhance the image quality locally. It could be examined if the focus is set to the region that is gazed at in practice. Second, the biometry measurement process could be further examined. For instance, it could be studied which anatomical structures and image features sonographers use when placing the measurement calipers, and to compare these between sonographers, between trimesters and against the guidelines.

Other avenues that can be tackled with similar methodology include the study of sonographer visual perception strategies and quantification of sonographer expertise. For the former, we have presented preliminary analyses [DDNP20] that identified the frequency with which anatomical landmarks are looked at during standard plane acquisition. Possible future analyses could include the identification of spatio-temporal visual search strategies and the inclusion of probe motion data. For the quantification of sonographer expertise, we have published initial work based on probe motion data [WDJ+20] which could be extended with gaze-tracking data into a unified framework. As most of our work, this study is based on the PULSE dataset which includes qualified sonographers with different levels of expertise. A possible future extension is to include unqualified novices or students in the process of becoming qualified.

7.2 Visual Saliency Modeling for Image and Video Data

Even though the field of visual saliency prediction has existed for over 30 years [IKN98], the space for future research directions is as large as ever. We demonstrated the importance of two approaches: application-specific saliency models inspired by cognitive science (Sec. 4.1) and domain adaptation to unify different sources of data (Sec. 4.2). Each of these directions opens up new research questions. First, while our cognition-inspired model consisting of a bidirectional recurrent neural network is a first attempt to capture the prediction-driven nature of sonographer visual attention, it remains a simplistic representation of the actual underlying cognitive processes of the observer. In contrast to the passive observation of an image or video, probe movement is an inherent part of sonographers' visual search. Hence, incorporating it into ultrasound visual saliency prediction would likely capture the sonographer cognition more comprehensively. On the direction of multi-domain saliency prediction, a possible next step could be transferring the method from the benchmark datasets considered in the paper to more realistic settings like point-of-view gaze-tracking (*e.g.* [LLR20]) or different ultrasound scanning tasks such as the different trimesters.

7.3 Ultrasound Image Analysis with Visual Saliency Models

While we have demonstrated the merit of visual saliency prediction for a range of ultrasound image analysis tasks, we have limited our studies to the analysis of frozen ultrasound images. The video frames were only used for representation learning with the gaze-tracking data. However, there are many video-based tasks

that could potentially benefit from our methods, such as classifying ultrasound video clips [SDC+19] or the related analysis of sonographer workflow [SDC+21]. First experiments in this direction have been conducted [CDS+20] but focus on saliency prediction while the downstream task is limited to clip classification with three anatomical categories. Moreover, it remains to study the benefit of the proposed methods during deployment as real-time assistive systems. This is the goal of the PURFECT (Perception Ultrasound for Reassuring Fetal Echo Clinical Trainees) project, a new effort which will develop prototypes of the methods developed within the PULSE project.

7.4 Automatic Probe Movement Guidance for Freehand Obstetric Ultrasound

Authors. Richard Droste, Lior Drukker, Aris T. Papageorghiou, J. Alison Noble

Conference. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* 2020.

Background. Here I present an additional paper which is not included in the research contribution chapters because it does not make use of gaze-tracking data, the overarching theme of this thesis. Nonetheless, it is added to this extended outlook chapter since it describes a machine learning model that learns from experienced sonographers in order to be able to guide novice operators, a central aspect of the PULSE project. The freehand acquisition of standard anatomical views is arguable the most difficult aspect of obstetric ultrasound scanning. While it is possible that image-based methods like standard plane detection could facilitate this by allowing a trial-and-error approach of finding the correct views, we present a system that provides explicit guidance for the movement of the ultrasound probe.

Statement of Authorship. I was the lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft. J. Alison Noble was the main responsible for supervision and funding acquisition and contributed to conceptualization, methodology and editing the draft. Aris T. Papageorghiou contributed to supervision, funding acquisition and editing the draft. Lior Drukker was responsible for data acquisition and edited the draft.

Recognition. The paper was nominated and shortlisted for the MICCAI 2020 Young Scientist Award.

Abstract

We present the first system that provides real-time probe movement guidance for acquiring standard planes in routine freehand obstetric ultrasound scanning. Such a system can contribute to the worldwide deployment of obstetric ultrasound scanning by lowering the required level of operator expertise. The system employs an artificial neural network that receives the ultrasound video signal and the motion signal of an inertial measurement unit (IMU) that is attached to the probe, and predicts a guidance signal. The network termed US-GuideNet predicts either the movement towards the standard plane position (goal prediction), or the next movement that an expert sonographer would perform (action prediction). While existing models for other ultrasound applications are trained with simulations or phantoms, we train our model with real-world ultrasound video and probe motion data from 464 routine clinical scans by 17 accredited sonographers. Evaluations for 3 standard plane types show that the model provides a useful guidance signal with an accuracy of 88.8 % for goal prediction and 90.9 % for action prediction.

7.4.1 Introduction

Ultrasound scanning is an indispensable diagnostic tool in obstetrics due to its safety, real-time results and low cost. At the same time, many women in developing countries do not receive a single ultrasound examination throughout their pregnancy due to a lack of skilled operators [SBA+15]. The main tasks of ultrasound scanning are the acquisition, examination/verification and interpretation of predefined standard anatomical planes that enable the detection of fetal abnormalities. Systems that provide assistance for or automate these tasks have the potential of enabling worldwide access to ultrasound scanning by reducing the level of necessary expertise. Standard plane examination/verification and interpretation are largely standardized [SAB+11], can be performed remotely [BMS+19], and can be facilitated through automated image analysis [YKPN17]. Freehand standard plane acquisition, on the other hand, is harder to facilitate/automate since it is not standardized and requires interaction with the mother. It demands years of training and is the rate-limiting step even for experienced sonographers [BBB+16].

To address this issue, we present the first system that provides real-time probe movement guidance for fetal standard plane acquisition in routine freehand obstetric ultrasound scanning. An overview of the system is presented in Fig. 7.1. An artificial neural network termed *US-GuideNet* receives the ultrasound video signal alongside the signal of a motion sensor that is attached to the ultrasound probe, and outputs probe movement guidance that directs the operator towards the desired standard plane. No specialized equipment is required: The motion sensor is a common inertial measurement unit (IMU) that is attached to the probe of a standard clinical ultrasound machine. Further, the *US-GuideNet* neural network is designed to be extremely lightweight and can run real time inference on a CPU. Behavioral cloning (BC), a type of imitation learning, has emerged as a powerful technique to train neural networks to perform complex real-world tasks

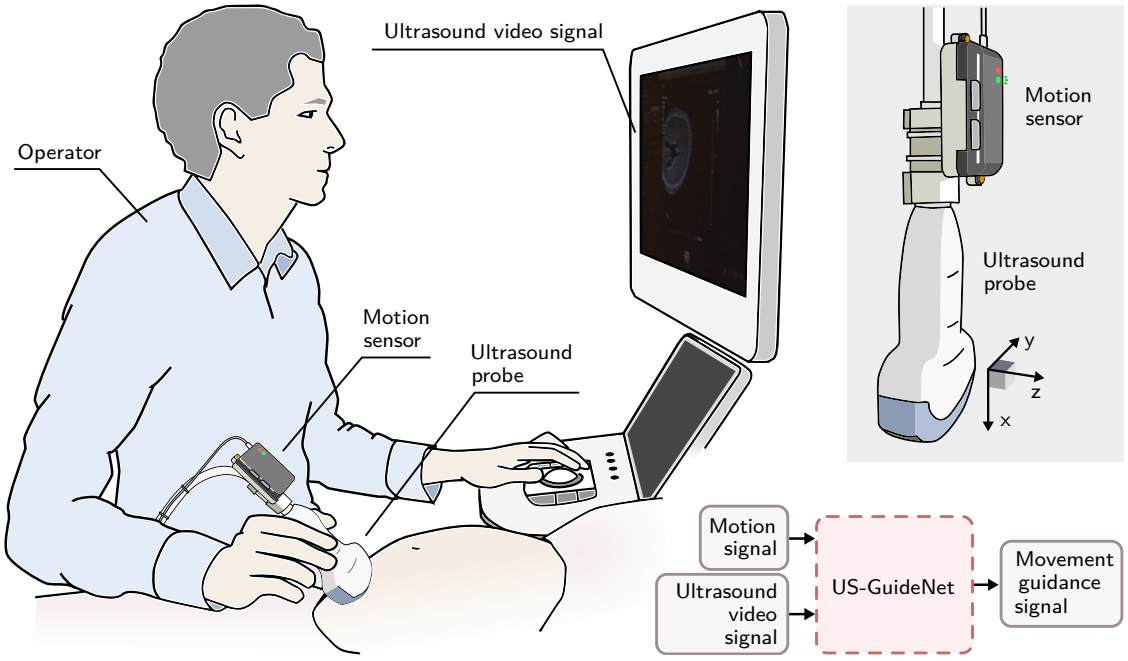

Figure 7.1: System overview. *Left*: An operator performs ultrasound scanning with a routine clinical setup while the motion of the probe is recorded with an IMU. *Bottom right*: The *US-GuideNet* receives the IMU motion signal and the ultrasound video signal as inputs and outputs a real-time probe movement guidance signal. *Top right*: Attachment of the IMU to the ultrasound probe and IMU coordinate system.

such as autonomous driving [PCS+18]. Here, we collect 5079 demonstrations of standard plane acquisitions from 464 2^{nd}- and 3^{rd}-trimester scans acquired by 17 accredited sonographers and implement two settings of BC: 1) For *goal prediction*, the network predicts the movement that leads directly to the estimated position of the standard plane. 2) For *action prediction*, the network predicts the next movement that the expert would perform.

7.4.2 Related Work

Various approaches have been proposed to address the difficulty of ultrasound standard plane acquisition.

Robotic Ultrasound. Human-controlled robotic systems have been developed that allow experienced sonographers to perform obstetric ultrasound exams remotely

[VTC+03]. Automated robotic systems have been proposed for highly structured tasks such as finding planes of motionless objects [MKC10, LRL+10] or the human liver [MIM+13]. However, despite on-going efforts [WHN+19], no robotic system has been proposed that can automate the complex task of obstetric ultrasound scanning.

Simplified Acquisition Protocols. Instead of assisting operators to acquire typical freehand 2D standard planes, previous work has proposed to automatically extract standard planes from data that are acquired with a simplified protocol, such as 3D ultrasound volumes [RPN11, LKH+18] or linear sweeps over the maternal abdomen [MBN+17]. Moreover, IMUs have been used to acquire 3D ultrasound volumes with 2D probes [HTGP08, PSJ+18]. However, these methods are applicable only for a subset of standard planes (fetal abdomen and head) and the standard plane quality is not up to par with typical freehand scanning.

Phantoms and Simulated Environments. Recent studies have proposed learning based systems that are trained to acquire ultrasound planes in simplified environments. One study proposes an algorithm that learns to find a view of the adult heart in a grid of pre-acquired ultrasound images [MBS19]. Moreover, learning-based systems have been proposed in which a robotic actuator finds predefined views of simple tissue phantoms [JL19] or a fetal US phantom [TWBX18]. However, a fetus in the mother's womb is a dynamic and highly variant object that can not be well-represented with static simulations or a phantom. Furthermore, these algorithms are purely image-based and therefore rely on the exact execution of the predicted actions, which is only possible within a simulation or with a robotic system. Here, we train a guidance algorithm with video and probe motion data from a large number of real-world expert demonstrations from routine scanning. Moreover, our algorithm receives the real-time probe motion signal and can therefore react to the movements of a human operator.

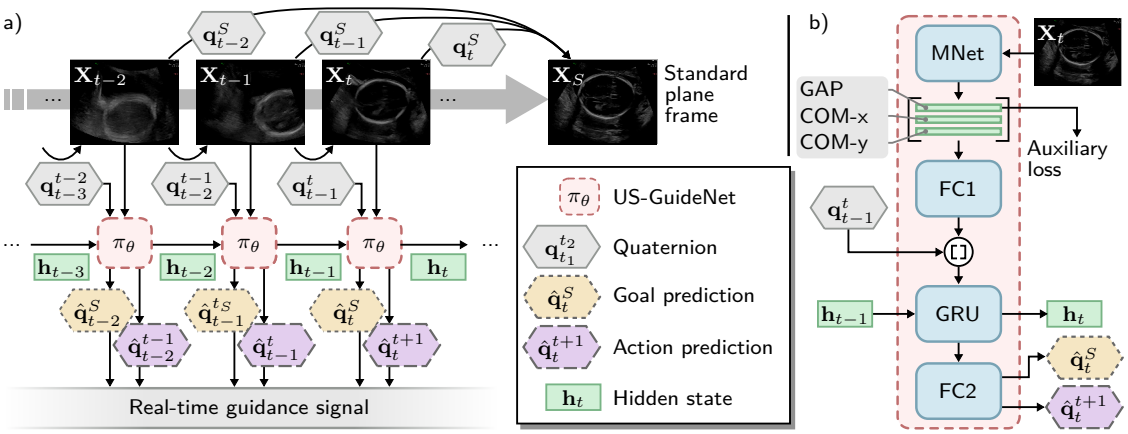

Figure 7.2: a) Proposed behavioral cloning framework. b) *US-GuideNet* architecture.

7.4.3 Method

Fig. 7.1 presents an overview of the proposed system. An operator performs routine obstetric ultrasound scanning with a standard clinical machine. An inertial measurement unit (IMU) motion sensor is attached to the ultrasound probe and an on-board attitude and heading reference system (AHRS) estimates the sensor's orientation in the earth coordinate system. The motion sensor signal and the machine video signal are input into a neural network, *US-GuideNet*, that outputs a 3D rotation of the probe that guides the operator towards the standard plane. The network training method is described in Sec. 7.4.3, the network architecture is detailed in Sec. 7.4.3 and implementation details are provided in Sec. 7.4.3.

Learning from Expert Demonstrations

We pose the problem of training a neural network to predict a probe guidance signal as a behavioral cloning problem. That is, we record standard plane acquisition demonstrations from several experts for a large number of patients and train the network to replicate the demonstrated behavior. In general, a standard plane acquisition consists of live B-mode scanning followed by *freezing* the ultrasound video and optionally selecting a previous frame with the desired appearance from

a *cine-buffer*. We define the finally selected frame as the standard plane.

Problem Formulation. Fig. 7.2 a) presents the formulation of the learning problem. Let $\{\mathbf{X}_t \in \mathbb{R}^{H\times W} \,|\, t \in \mathcal{T}\}$ be ultrasound video frames of a standard plane acquisition, temporally downsampled to 6 Hz, with resolution $H \times W$ and frame indices $\mathcal{T} = \{i\}_{i=0}^{F}$, where F is the *freeze* frame index. Moreover, let $\mathbf{X}_S$ be the standard plane with index $S \in \mathcal{T} \setminus \{0\}$. Finally, let $\mathbf{q}_t = [q_w, q_x, q_y, q_z]^\top$ be the probe orientation quaternion of frame $\mathbf{X}_t$ and $\mathbf{q}_t^{t_2} := \mathbf{q}_t^* \mathbf{q}_{t_2}$ the probe rotation quaternion from frame $\mathbf{X}_t$ to frame $\mathbf{X}_{t_2}$, where $\mathbf{q}^*$ is the conjugate. We represent orientations with quaternions since they can be smoothly interpolated without discontinuities or singularities, and are numerically stable and computationally efficient [PFAG19]. Euler angles, in contrast, another popular representation of rotations, suffer from discontinuities such as the *gimbal lock*. We do not consider probe translation in this work since the IMU is not suitable to estimate it accurately.

Behavioral Cloning. We train a policy network $\pi_\theta : s_t \mapsto u_t$ termed *US-GuideNet* with parameters θ that maps the state s_t at time step $t \in \mathcal{T}$ to an action u_t. We define the state as the tuple $s_t := (\mathbf{X}_t, \mathbf{q}_{t-1}^t, \mathbf{h}_{t-1})$, where $\mathbf{h}_{t-1}$ is the hidden state of a recurrent neural network within π_θ. We explore the two different settings for the action u_t: *goal prediction* and *action prediction*. For *goal prediction*, the policy $\pi_\theta^g : s^t \mapsto \hat{\mathbf{q}}_t^S$ estimates the rotation from the current orientation to the orientation of the standard plane. If the estimated standard plane orientation is accurate, this policy is optimal, *i.e.*, it guides the operator directly to the standard plane. However, it is not guaranteed that enough information has been seen at time t for an accurate estimation of the standard plane orientation. Therefore, we explore a second setting, *action prediction*, where the policy $\pi_\theta^a : s_t \mapsto \hat{\mathbf{q}}_t^{t+1}$ estimates the next rotation that the operator would perform. This policy aims to closely mimic the expert sonographer behavior.

Loss Function. During training, a demonstration is constructed from a subset of indices $\mathcal{T}_D \subset \mathcal{T}$ with start and end indices t_0 and T. Let $\hat{\mathbf{Q}} := [\hat{\mathbf{q}}_t^{t_2}]_{t=t_0}^{T}$ and

$\mathbf{Q} := [\mathbf{q}_t^{t_2}]_{t=t_0}^{T}$ be the predicted and ground truth rotation quaternion sequences, with $t_2 \in \{S, t+1\}$ for *goal prediction* and *action prediction* respectively. We add an auxiliary output after the MNet in order to facilitate and regularize its training: Since we want the MNet to recognize the appearance of standard planes, we input the average pooled MNet features into a softmax layer that predicts the class probabilities of the SonoNet standard plane classifier [BKM+17] for each frame. Let $\hat{\mathbf{P}} = [\hat{\mathbf{p}}_t]_{t=t_0}^{T}$ be the auxiliary softmax output and $\mathbf{Y} = [\mathbf{y}_t]_{t=t_0}^{T}$ the SonoNet class probabilities, with $\hat{\mathbf{p}}_t, \mathbf{y}_t \in \mathbb{R}_{\geq 0}^{14}$. The total training loss $\mathcal{L}$ is

$$\mathcal{L} = \sum_{t=t_0+W}^{T} \Bigg\{ \underbrace{-\frac{1}{\|\hat{\mathbf{q}}_t^{t_2}\|} \hat{\mathbf{q}}_t^{t_2} \cdot \mathbf{q}_t^{t_2}}_{\text{Similarity loss}} + \alpha \underbrace{(1 - \|\hat{\mathbf{q}}_t^{t_2}\|^2)^2}_{\text{Norm loss}} \Bigg\} - \beta \sum_{t=t_0}^{T} \underbrace{\mathbf{y}_t^{\top} \operatorname{diag}(\mathbf{w}) \log(\mathbf{p}_t)}_{\text{Auxiliary loss}}$$

where $\cdot$ denotes the dot-product, $\alpha, \beta \in \mathbb{R}_{>0}$ are scalar weighting factors, $\mathbf{w} \in \mathbb{R}_{\geq 0}^{14}$ is a weight vector that balances the SonoNet class probabilities, and W is a warm up time for the rotation prediction.

US-GuideNet Architecture

The *US-GuideNet* policy network receives the ultrasound video and probe motion signals and outputs predicted expert probe rotations as described in Sec. 7.4.3. We design the architecture for small time and space computational complexity (runtime and model size) such that it can run real-time inference on the CPU of an inexpensive computer. The network architecture is illustrated in Fig. 7.2 b). At each time step t, the ultrasound video frame X_t is fed into a MobileNet V2 (MNet) convolutional neural network [SHZ+18], which consists of a cascade of lightweight depthwise-separable and pointwise convolutions. We use MNet with a width-multiplier of 0.5, *i.e.*, 50% reduced number of channels. Next, the dimensionality of the MNet output is reduced with a custom *ConcatPool* operation that preserves both semantic and spatial information by concatenating global average pooled (GAP) features with

the x and y coordinates of the centers of mass (COM-x/y) of the feature maps. After reducing the features to 128 channels with a fully-connected layer *FC1*, they are concatenated with the current probe rotation quaternion $\mathbf{q}_{t-1}^{t}$ and input into a gated recurrent unit [CvBB14] (GRU) with 132 input channels and 128 hidden channels. Finally *FC2*, a fully-connected layer with one 128-channel hidden layer, outputs the 4-dimensional probe rotation quaternion.

Experimental Setup

Data Acquisition. The data were acquired as part of the PULSE (Perception Ultrasound by Learning Sonographic Experience) project, a prospective study of routine fetal ultrasound scans performed in all trimesters by accredited sonographers and fetal medicine doctors at the maternity ultrasound unit, Oxford University Hospitals NHS Foundation Trust, Oxfordshire, United Kingdom. The exams were performed on a GE Voluson E8 scanner (General Electric, USA) while the video signal of the machine monitor was recorded lossless at 30 Hz. The motion of each of two curved linear array transducer (2D) probes was recorded with a NGIMU IMU/AHRS (x-io Technologies Ltd., UK). Each motion sensor was attached to the cable outlet of the probe with a custom 3D-printed mount as shown in Fig. 7.1. The probe orientation quaternions were sampled at 400 Hz. This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051) and written informed consent was given by all participating pregnant women and operators. In this paper we use ultrasound video and corresponding gaze data of 464 second and third trimester scans acquired by 17 accredited sonographers between May 2018 and February 2020.

Data Processing. We extracted the standard plane acquisitions from the ultrasound scans with a purpose-built program based on optical character recognition. For each of the 5079 resulting acquisitions, the program outputs the corresponding live B-Mode scanning segment, the *freeze* frame and the *cine-buffer*-corrected

standard plane. In addition, the program labels acquisitions of the biometry standard planes: the femur standard plane (FSP), the abdominal standard plane (ASP) and the trans-ventricular plane (TVP) [SAB+11], which we use for evaluation. The acquisition duration was limited to 10 s before the standard plane. We automatically corrected any lag between the video and motion signals by correlating frame differences with probe motion, and manually verified the synchronization. The video frames were cropped such that the ultrasound machine's graphical user interface was removed, and normalized to zero-mean and unit-variance. The scans are divided into 80% for training and 20% for testing.

Training. For each training epoch, a demonstration of 32 frames is randomly selected from each standard plane acquisition, which corresponds to a duration 5.3 s at 6 Hz. It is ensured that $\min_t\{\mathbf{q}_t \cdot \mathbf{q}_S\} \geq 0.7$ for each demonstration. The frames are augmented by randomly changing of the brightness, contrast and gamma by $\pm 10\%$ and randomly symmetrically cropping up to 20% of the frame border. The frames are then down-sampled to the network input resolution of 224×288. The MNet is pre-trained via the auxiliary loss with a large number of ultrasound frames. The entire *US-GuideNet* neural network is then trained from the demonstrations for 20 epochs with the AdamW optimizer [LH19] with weight decay of 10^{-2} and initial learning rate of 0.001, which is decayed by a factor of 0.1 every 8 epochs. The batch size is set to 8 and the warm up time for the rotation loss to 1 s. After training with all demonstrations, the model is fine-tuned for each evaluation plane (FSP, ASP and TVP) separately for 16 epochs.

Evaluation and Baseline. We evaluate the trained model on the full-length standard plane acquisitions (clipped to 10 s before the standard plane). For each time step, we classify the predicted probe rotation as correct if and only if $\angle(\mathbf{q}_t \hat{\mathbf{q}}_t^{t_2}, \mathbf{q}_{t_2}) < \angle(\mathbf{q}_t, \mathbf{q}_{t_2}), t_2 \in \{S, t+1\}$, *i.e.*, if applying the predicted rotation reduces the angle to the target orientation. As before, $t_2 = S$ for *goal prediction* and $t_2 = t+1$ for *action prediction*. As a baseline rotation prediction we use $\mathbf{q}_{t-1}^t$, *i.e.*, continuing

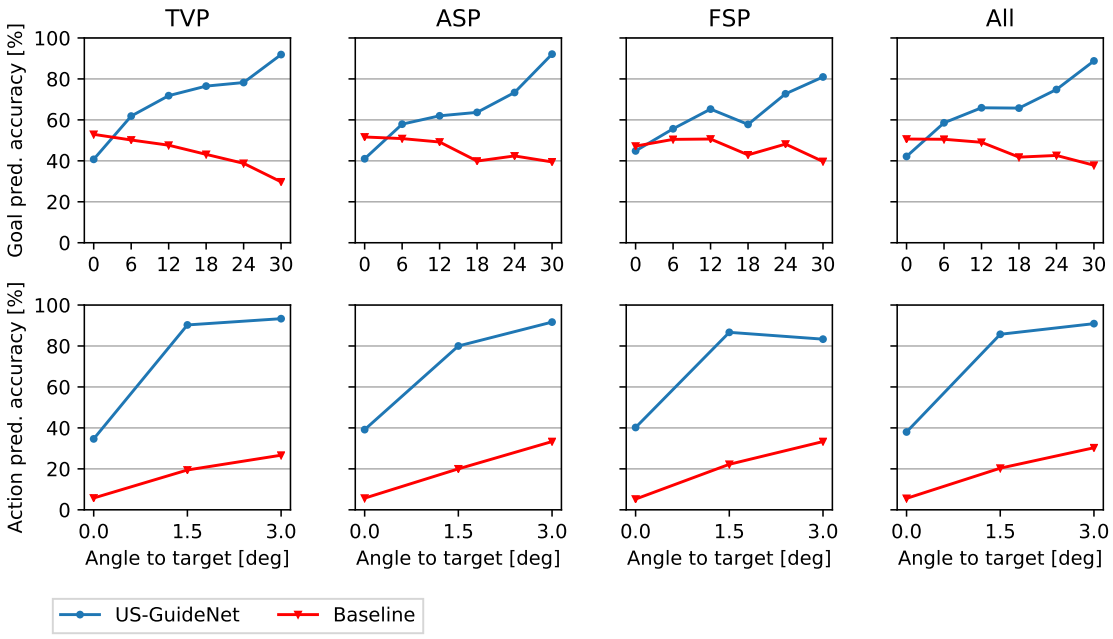

Figure 7.3: Experimental results for the evaluated standard planes: TVP (head), ASP (abdomen) and FSP (femur). In addition, the overall accuracies are provided.

in the current direction of rotation at each time step.

7.4.4 Results

The experimental results are shown in Fig. 7.3. The average accuracy of the guidance signal and baseline is evaluated for different ranges of the angular distance to the target (standard plane orientation for *goal prediction* or next probe position for *action prediction*). This enables the separation of the performance for coarse (large angular distance) and fine (low angular distance) adjustments. The x-axis of the individual plots provides the lower limits of the ranges, which extend to the next-higher x-axis value. Across the *action prediction* and *goal prediction* settings and all evaluated standard plane types, a common pattern can be observed that the accuracy of the guidance signal tends to increase with increasing angular distance to the standard plane.

Goal Prediction. The *goal prediction* accuracies are given in the upper row of Fig. 7.3. The guidance signal performs better than the baseline for any angle

range >6°. The accuracy of the guidance signal increases with increasing angular distances to the standard plane, ranging from 42.2 % for angles 0° to 6° to 88.8 % for angles >30°, with 81.0 % for the FSP and 92.1 % for the ASP. The average baseline accuracy slightly declines towards higher angular distances.

Action Prediction. The guidance signal accuracy is higher than the baseline accuracy for all target distance ranges. The average guidance signal accuracy increases from 38.0 % for angles 0° to 1.5° to 90.9 % for distances >3°. At angles >3°, the largest accuracy is observed for the TVP with 93.3 % and the lowest for the FSP with 83.3 %. The average baseline accuracy slightly increases with increasing angular distances.

7.4.5 Discussion and Conclusion

The results presented in subsection 7.4.4 demonstrate that the proposed probe guidance system for obstetric ultrasound scanning indeed provides a useful navigation signal towards the respective target, which is the standard plane orientation for *goal prediction* and the next expert movement for *action prediction*. The accuracy of *US-GuideNet* increases for larger differences to the target orientation, which shows that the algorithm is robust for guiding the operator towards the target orientation from distant starting points. For small distances, it is difficult to predict an accurate guidance signal since the exact target orientation may be subject to inter- and intra-sonographer variations or sensor uncertainty. The accuracy is similar for *goal prediction* and *action prediction* but slightly higher for action prediction at intermediate angles to the target, which can be explained by the fact that the action is always based on the previously seen frames, while the target position might be yet unknown.

The guidance signal accuracies are generally the highest for the abdominal and head standard planes (ASP and TVP). The accuracy for the femur standard plane

(FSP) is slightly lower, which can be explained by the fact the femur is part of an extremity and therefore subject to more fetal movement, which can make its final position unpredictable. Moreover, the FSP is defined via two anatomical landmarks—the distal and proximal ends of the femur—while the ASP and TVP are determined by the appearance of more anatomical structures [SAB+11]. This might make it more difficult for the model to predict the FSP position that was chosen by the operator, since it is subject to more degrees of freedom.

A limitation of our study is that we test our algorithm with pre-acquired data. However, in contrast with previous work [MBS19, JL19, TWBX18] which uses simulations or phantoms, our proposed system is trained and evaluated on data from real-world routine ultrasound scanning. Moreover, instead of relying on the exact execution of the probe guidance as in previous work [MBS19, JL19, TWBX18], our system reacts to the actual operator probe movements that are sensed with an IMU. This suggests that the system will perform well in future tests on volunteer subjects. In general, the accuracy of the predictions of *US-GuideNet* is evident from the large improvements over the baseline of simply continuing the current direction of rotation. While probe translation is not predicted due to IMU limitations, only the through-plane sweeping translation would usually change the view of the fetus while the sideways sliding and downwards/upwards translations would shift the fetal structure within the ultrasound image. In combination with the rotation guidance, this leaves one degree of freedom to be determined by the operator.

In conclusion, this paper presents the first probe movement guidance system for the acquisition of standard planes in freehand obstetric ultrasound scanning. Moreover, it is the first guidance system for any application of ultrasound standard plane acquisition that is trained with video and probe motion data from routine clinical scanning. Our experiments have shown that the proposed *US-GuideNet* network and behavioral cloning framework result in an accurate guidance system. These results will serve as a foundation for subsequent validation studies with novice

operators. The proposed algorithm is lightweight which facilitates the deployment for existing ultrasound machines.

Bibliography

[BBB+16] David P. Bahner, J. Matthew Blickendorf, Marcia Bockbrader, Eric Adkins, Amar Vira, Creagh Boulger, and Ashish R. Panchal. Language of Transducer Manipulation. *Journal of Ultrasound in Medicine*, 35(1):183–188, 2016.

[BKM+17] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. *IEEE Trans. Med. Imag.*, 36(11):2204–2215, 2017.

[BMS+19] Noel Britton, Michael A. Miller, Sami Safadi, Ariel Siegel, Andrea R. Levine, and Michael T. McCurdy. Tele-Ultrasound in Resource-Limited Settings: A Systematic Review. *Front. Public Health*, 7, 2019.

[CvBB14] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014.

[HTGP08] Richard Housden, Graham M Treece, Andrew H Gee, and Richard W Prager. Calibration of an orientation sensor for freehand 3D ultrasound and its use in a hybrid acquisition system. *BioMed Eng OnLine*, 7(1):5, 2008.

[JL19] Piotr Jarosik and Marcin Lewandowski. Automatic Ultrasound Guidance Based on Deep Reinforcement Learning. In *IEEE International Ultrasonics Symposium (IUS)*, pages 475–478, 2019.

[LH19] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[LKH+18] Yuanwei Li, Bishesh Khanal, Benjamin Hou, Amir Alansary, Juan J. Cerrolaza, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, and Daniel Rueckert. Standard Plane Detection in 3D Fetal Ultrasound Using an Iterative Transformation Network. In *MICCAI*, pages 392–400, 2018.

[LRL+10] Kaicheng Liang, Albert J. Rogers, Edward D. Light, Daniel von Allmen, and Stephen W. Smith. Three-Dimensional Ultrasound Guidance of Autonomous Robotic Breast Biopsy: Feasibility Study. *Ultrasound in Medicine & Biology*, 36(1):173–177, January 2010.

[MBN+17] M. A. Maraci, C. P. Bridge, R. Napolitano, A. Papageorghiou, and J. A. Noble. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. *Med. Image Anal.*, 37:22–36, 2017.

[MBS19] Fausto Milletari, Vighnesh Birodkar, and Michal Sofka. Straight to the Point: Reinforcement Learning for User Guidance in Ultrasound. In *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis (MICCAI Workshops)*, pages 3–10, 2019.

[MIM+13] Ammar Safwan Bin Mustafa, Takashi Ishii, Yoshiki Matsunaga, Ryu Nakadate, Hiroyuki Ishii, Kouji Ogawa, Akiko Saito, Motoaki Sugawara, Kiyomi Niki, and Atsuo Takanishi. Development of robotic system for autonomous liver screening using ultrasound scanning device. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 804–809, 2013.

[MKC10] R. Mebarki, A. Krupa, and F. Chaumette. 2-D Ultrasound Probe Complete Guidance by Visual Servoing Using Image Moments. *IEEE Trans. Robot.*, 26(2):296–306, 2010.

[PCS+18] Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile Autonomous Driving using End-to-End Deep Imitation Learning. In *Proceedings of Robotics: Science and Systems*, 2018.

[PFAG19] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling Human Motion with Quaternion-based Neural Networks. In *BMVC*, 2019.

[PSJ+18] Raphael Prevost, Mehrdad Salehi, Simon Jagoda, Navneet Kumar, Julian Sprung, Alexander Ladikos, Robert Bauer, Oliver Zettinig, and Wolfgang Wein. 3D freehand ultrasound without external tracking using deep learning. *Medical Image Analysis*, 48:187–202, August 2018.

[RPN11] Bahbibi Rahmatullah, Aris Papageorghiou, and J. Alison Noble. Automated Selection of Standardized Planes from Ultrasound Volume. In *Machine Learning in Medical Imaging (MICCAI Workshop)*, volume 7009, pages 35–42, 2011.

[SAB+11] L. J. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. L. Johnsen, K. Kalache, K. Y. Leung, G. Malinger, H. Munoz, F. Prefumo, A. Toi, and W. Lee. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound Obstet. Gynecol.*, 37(1):116–126, 2011.

[SBA+15] Sachita Shah, Blaise A. Bellows, Adeyinka A. Adedipe, Jodie E. Totten, Brandon H. Backlund, and Dana Sajed. Perceived barriers in the use of ultrasound in developing countries. *Crit. Ultrasound J.*, 7:11, 2015.

[SHZ+18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 2018.

[TWBX18] Grzegorz Toporek, Haibo Wang, Marcin Balicki, and Hua Xie. Autonomous image-based ultrasound probe positioning via deep learning. In *Hamlyn Symposium on Medical Robotics*, 2018.

[VTC+03] A. Vilchis, J. Troccaz, P. Cinquin, K. Masuda, and F. Pellissier. A new robot architecture for tele-echography. *IEEE Trans. Robot. Autom.*, 19(5):922–926, 2003.

[WHN+19] Shuangyi Wang, James Housden, Yohan Noh, Davinder Singh, Anisha Singh, Emily Skelton, Jacqueline Matthew, Cornelius Tan, Junghwan Back, Lukas Lindenroth, Alberto Gomez, Nicolas Toussaint, Veronika Zimmer, Caroline Knight, Tara Fletcher, David Lloyd, John Simpson, Dharmintra Pasupathy, Hongbin Liu, Kaspar Althoefer, Joseph Hajnal, Reza Razavi, and Kawal Rhode. Robotic-Assisted Ultrasound for Fetal Imaging: Evolution from Single-Arm to Dual-Arm System. In *Towards Autonomous Robotic Systems*, pages 27–38, 2019.

[YKPN17] Mohammad Yaqub, Brenda Kelly, Aris T. Papageorghiou, and J. Alison Noble. A Deep Learning Solution for Automatic Fetal Neurosonographic Diagnostic Plane Verification Using Clinical Standard Constraints. *Ultrasound in Medicine & Biology*, 43(12):2925–2933, 2017.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Automatic Probe Movement Guidance for Freehand Obstetric Ultrasound
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Richard Droste, Lior Drukker, Aris T. Papageorghiou, J. Alison Noble. "Automatic Probe Movement Guidance for Freehand Obstetric Ultrasound". In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020.

Student Confirmation

Student Name:	Richard Droste		
Contribution to the Paper	I was the lead author responsible for conceptualization, methodology, implementation, experiments, data analysis/visualization and the original draft		
Signature	Richard Droste	Date	31.05.2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. J. Alison Noble			
I confirm that the description above is accurate.			
Signature	J Alison Noble	Date	01/06/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Bibliography

[Abr12] Jacques S. Abramowicz. Ultrasound and autism: Association, link, or coincidence? *J Ultrasound Med*, 31(8):1261–1269, August 2012.

[Abr13] Jacques S. Abramowicz. Benefits and risks of ultrasound in pregnancy. *Seminars in Perinatology*, 37(5):295–300, October 2013.

[AES+20] Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. A Curriculum Learning Based Approach to Captioning Ultrasound Images. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis (MICCAI Workshops)*, 2020.

[AGD+06] Eugenius S. B. C. Ang, Vicko Gluncic, Alvaro Duque, Mark E. Schafer, and Pasko Rakic. Prenatal exposure to ultrasound waves impacts neuronal migration in mice. *PNAS*, 103(34):12903–12910, August 2006.

[AHM13] S. D. Adhikary, A. Hadzic, and P. M. McQuillan. Simulator for teaching hand–eye coordination during ultrasound-guided regional anaesthesia. *British Journal of Anaesthesia*, 111(5):844–845, November 2013.

[AKM+03] J. S. Abramowicz, G. Kossoff, K. Marsal, G. Ter Haar, and International Society of Ultrasound in Obstetrics and Gynecology Bioeffects and Safety Committee. Executive Board of the International Society of Ultrasound in Obstetrics and Gynecology. Safety Statement, 2000 (reconfirmed 2003). International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). *Ultrasound Obstet Gynecol*, 21(1):100, January 2003.

[Ame16] American Institute of Ultrasound in Medicine. Official Statement: Recommended Maximum Scanning Times for Displayed Thermal Index (TI) Values. https://www.aium.org/officialStatements/65, 2016.

[AN16a] M. Ahmed and J. A. Noble. An eye-tracking inspired method for standardised plane extraction from fetal abdominal ultrasound volumes. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2016.

[AN16b] M. Ahmed and J. A. Noble. Fetal Ultrasound Image Classification Using a Bag-of-words Model Trained on Sonographers' Eye Movements. *Procedia Computer Science*, 90(July):157–162, 2016.

[ASD+19] Mohammad Alsharid, Harshita Sharma, Lior Drukker, Pierre Chatelain, Aris T. Papageorghiou, and J. Alison Noble. Captioning

Ultrasound Images Automatically. In *Medical Image Computing and Computer Assisted Intervention*, 2019.

[ATG+17] S. Kate Alldred, Yemisi Takwoingi, Boliang Guo, Mary Pennant, Jonathan J. Deeks, James P. Neilson, and Zarko Alfirevic. First trimester ultrasound tests alone or in combination with first trimester serum tests for Down's syndrome screening. *Cochrane Database Syst Rev*, 3:CD012600, March 2017.

[AZ18] Relja Arandjelović and Andrew Zisserman. Objects that Sound. In *European Conference on Computer Vision*, 2018.

[BBB+16] David P. Bahner, J. Matthew Blickendorf, Marcia Bockbrader, Eric Adkins, Amar Vira, Creagh Boulger, and Ashish R. Panchal. Language of Transducer Manipulation. *Journal of Ultrasound in Medicine*, 35(1):183–188, 2016.

[BHK+18] Lindsay K. Borg, T. Kyle Harrison, Alex Kou, Edward R. Mariano, Ankeet D. Udani, T. Edward Kim, Cynthia Shum, and Steven K. Howard. Preliminary Experience Using Eye-Tracking Technology to Differentiate Novice and Expert Image Interpretation for Ultrasound-Guided Regional Anesthesia. *Journal of Ultrasound in Medicine*, 37(2):329–336, 2018.

[BI13] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[BI15] Ali Borji and Laurent Itti. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. In *CVPR 2015 Workshop on "Future of Datasets"*, May 2015.

[Bis94] Christopher M Bishop. Mixture Density Networks. 1994.

[BJB+12] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT Saliency Benchmark. http://saliency.mit.edu/, 2012.

[BJO+19] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757, 2019.

[BKEE18] Cagdas Bak, Aysun Kocak, Erkut Erdem, and Aykut Erdem. Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction. *IEEE Trans. Multimed.*, 20(7):1688–1698, 2018.

[BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *NIPS - Deep Learning Symposium*, 2016.

[BKM+17] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. *IEEE Trans. Med. Imag.*, 36(11):2204–2215, 2017.

[BLO01] G. Beevers, G. Y. Lip, and E. O'Brien. ABC of hypertension. Blood pressure measurement. Part I-sphygmomanometry: Factors common to all techniques. *BMJ*, 322(7292):981–985, April 2001.

[BLT17] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani. Recurrent Mixture Density Network for Spatiotemporal Visual Attention. *ICLR*, 2017.

[BMS+19] Noel Britton, Michael A. Miller, Sami Safadi, Ariel Siegel, Andrea R. Levine, and Michael T. McCurdy. Tele-Ultrasound in Resource-Limited Settings: A Systematic Review. *Front. Public Health*, 7, 2019.

[BN00] L Bricker and J P Neilson. Routine doppler ultrasound in pregnancy. *Cochrane database of systematic reviews (Online)*, (2):CD001450, 2000.

[Bor18] Ali Borji. Saliency Prediction in the Deep Learning Era: An Empirical Investigation. *arXiv:1810.03716*, 2018.

[BRB+16] Zoya Bylinskii, Adriá Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *Lecture Notes in Computer Science*, volume 9909 LNCS, pages 809–824, 2016.

[Bry95] Marc Brysbaert. Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, 124(4):434–452, 1995.

[BS93] Heiner C Bucher and Johannes G Schmidt. Does routine ultrasound scanning improve outcome in pregnancy? Meta-analysis ofvarious outcome measures. *BMJ*, 307:13–17, 1993.

[BSFT14] Bryann Bromley, Jean Spitz, Karin Fuchs, and Loralei L. Thornburg. Do clinical practitioners seeking credentialing for nuchal translucency measurement demonstrate compliance with biosafety recommendations? Experience of the Nuchal Translucency Quality Review Program. *J Ultrasound Med*, 33(7):1209–1214, July 2014.

[BSI13a] Ali Borji, Dicky N. Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013.

[BSI13b] Ali Borji, Dicky N. Sihite, and Laurent Itti. What stands out in a scene? A study of human explicit saliency judgment. *Vision Research*, 91:62–77, October 2013.

[BTHZ+00] S. B. Barnett, G. R. Ter Haar, M. C. Ziskin, H. D. Rott, F. A. Duck, and K. Maeda. International recommendations and guidelines for the safe use of diagnostic ultrasound in medicine. *Ultrasound Med Biol*, 26(3):355–366, March 2000.

[BTS+16] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *International Conference on Neural Information Processing Systems*, 2016.

[BTSI13] Ali Borji, Hamed R Tavakoli, Dicky N Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 921–928, 2013.

[BTVG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision*, 2006.

[Bus35] Guy Thomas Buswell. *How People Look at Pictures*. University of Chicago Press Chicago, 1935.

[BYPC16] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving Deeper into Convolutional Networks for Learning Video Representations. *ICLR*, 2016.

[CAB+17] Marcella Cornia, Davide Abati, Lorenzo Baraldi, Andrea Palazzi, Simone Calderara, and Rita Cucchiara. Attentive models in vision: Computing saliency maps in the deep learning era. *LNAI*, 10640:387–399, 2017.

[CB18] Marco Cuturi and Mathieu Blondel. Soft-DTW: A Differentiable Loss Function for Time-Series. *arXiv:1703.01541 [stat]*, February 2018.

[CBH16] Souad Chaabouni, Jenny Benois-pineau, and Ofer Hadar. Deep Learning for Saliency Prediction in Natural Video. *arXiv:1604.08010*, 2016.

[CBP+15] Ann J. Carrigan, Patrick C. Brennan, Mariusz Pietrzyk, Jillian Clarke, and Eugene Chekaluk. A 'snapshot' of the visual search behaviours of medical sonographers. *Australasian Journal of Ultrasound in Medicine*, 18(2):70–77, 2015.

[CBSC16] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2016.

[CBSC17a] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. *Proceedings - International Conference on Pattern Recognition*, pages 3488–3493, 2017.

[CBSC17b] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Visual saliency for image captioning in new multimedia services. In *ICMEW*, 2017.

[CDP+19] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*, 28(3):231–237, March 2019.

[CDS+20] Yifan Cai, Richard Droste, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Spatio-temporal visual attention modelling of standard biometry plane-finding navigation. *Medical Image Analysis*, 65:101762, October 2020.

[CFL+09] L. W. Chan, T. Y. Fung, T. Y. Leung, D. S. Sahota, and T. K. Lau. Volumetric (3D) imaging reduces inter- and intraobserver variation of fetal biometry measurements. *Ultrasound in Obstetrics and Gynecology*, 33(4):447–452, 2009.

[CGCB14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *NIPS*, 2014.

[CGGC08] G. Carneiro, B. Georgescu, S. Good, and D. Comaniciu. Detection and Measurement of Fetal Anatomies from Ultrasound Images using a Constrained Probabilistic Boosting Tree. *IEEE Transactions on Medical Imaging*, 27(9):1342–1355, September 2008.

[Cla13] Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03):181–204, 2013.

[CMG17] Chengqian Che, Tejas Sudharshan Mathai, and John Galeotti. Ultrasound registration: A review. *Methods*, 115:128–143, February 2017.

[CNQ+15] Hao Chen, Dong Ni, Jing Qin, Shengli Li, Xin Yang, Tianfu Wang, and Pheng Ann Heng. Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1627–1636, 2015.

[Cou04] Jennifer T. Coull. fMRI studies of temporal attention: Allocating attention within, or towards, time. *Cognitive Brain Research*, 21(2):216–226, 2004.

[Cou18] Rachel Courtland. Bias detectives: The researchers striving to make algorithms fair. *Nature*, 558(7710):357–360, June 2018.

[CSCN18a] Y. Cai, H. Sharma, P. Chatelain, and J. A. Noble. SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking. In *IEEE International Symposium on Biomedical Imaging*, 2018.

[CSCN18b] Yifan Cai, Harshita Sharma, Pierre Chatelain, and J. Alison Noble. Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[CSD+18] P. Chatelain, H. Sharma, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies. *IEEE Trans. Cybern.*, pages 1–11, 2018.

[CvBB14] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014.

[CvG+14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *EMNLP*, 2014.

[CWD+17] Hao Chen, Lingyun Wu, Qi Dou, Jing Qin, Shengli Li, Jie-Zhi Cheng, Dong Ni, and Pheng-Ann Heng. Ultrasound Standard Plane Detection Using a Composite Neural Network Framework. *IEEE Transactions on Cybernetics*, 47(6):1576–1586, June 2017.

[CYS+19] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, 2019.

[DA00] S. J. Day and D. G. Altman. Statistics notes: Blinding in clinical trials and other studies. *BMJ*, 321(7259):504, August 2000.

[DCD+20] R. Droste, P. Chatelain, L. Drukker, H. Sharma, A. T. Papageorghiou, and J. A. Noble. Discovering Salient Anatomical Landmarks by Predicting Human Gaze. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1711–1714, April 2020.

[DCS+19] Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention. In *Information Processing in Medical Imaging*, volume 11492, pages 592–604. 2019.

[DCS+20] Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Aris T. Papageorghiou, and J. Alison Noble. Towards Capturing Sonographic Experience: Cognition-Inspired Ultrasound Video Saliency Prediction. In *Medical Image Understanding and Analysis*, 2020.

[DDC+20a] L. Drukker, R. Droste, P. Chatelain, J. A. Noble, and A. T. Papageorghiou. Expected-value bias in routine third-trimester growth scans. *Ultrasound in Obstetrics & Gynecology*, 55(3):375–382, 2020.

[DDC+20b] Lior Drukker, Richard Droste, Pierre Chatelain, J. Alison Noble, and Aris T. Papageorghiou. Safety Indices of Ultrasound: Adherence to Recommendations and Awareness During Routine Obstetric Ultrasound Scanning. *Ultraschall Med*, 41(2):138–145, April 2020.

[DDNP20] L. Drukker, R. Droste, A. Noble, and A. T. Papageorghiou. VP40.20: Standard biometric planes: What are the salient anatomical landmarks? *Ultrasound in Obstetrics & Gynecology*, 56(S1):235–235, 2020.

[DDPN20] Richard Droste, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Automatic Probe Movement Guidance for Freehand Obstetric Ultrasound. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020.

[DDS+09] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[DFI+15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, December 2015.

[DFS+14] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. In *NIPS*, 2014.

[DGE15] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*, 2015.

[DJN20] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified Image and Video Saliency Modeling. In *European Conference on Computer Vision*, 2020.

[DL00] C. Deane and C. Lees. Doppler obstetric ultrasound: A graphical display of temporal changes in safety indices. *Ultrasound Obstet Gynecol*, 15(5):418–423, May 2000.

[DM16] Ian L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis, with Applications in R*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, September 2016.

[DOC+18] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised Cross-Modality Domain Adaptation of ConvNets for Biomedical Image Segmentations with Adversarial Loss. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 691–697, July 2018.

[DSB15] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1538–1546, 2015.

[DSD+20] L. Drukker, H. Sharma, R. Droste, J. A. Noble, and A. T. Papageorghiou. OC10.11: The data science of obstetric ultrasound: Automatic analysis of full-length anomaly scans using machine learning algorithms. *Ultrasound in Obstetrics & Gynecology*, 56(S1):31–31, 2020.

[Dud05] N. J. Dudley. A systematic review of the ultrasound estimation of fetal weight. *Ultrasound Obstet Gynecol*, 25(1):80–89, January 2005.

[ECF+93] Bernard G. Ewigman, James P. Crane, Fredric D. Frigoletto, Michael L. LeFevre, Raymond P. Bain, and Donald McNellis. Effect of Prenatal Ultrasound Screening on Perinatal Outcome. *New England Journal of Medicine*, 329(12):821–827, 1993.

[ESS94] Janet L. Engstrom, Claudia P. Sittler, and Karen E. Swift. Fundal height measurement: Part 5—The effect of clinician bias on fundal height measurements. *Journal of Nurse-Midwifery*, 39(3):130–141, May 1994.

[FWLF14] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE TIP*, 23(9):3910–3921, 2014.

[GC18] Siavash Gorji and James J Clark. Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *CVPR*, 2018.

[GEV+14] Yury Gitman, Mikhail Erofeev, Dmitriy Vatolin, Bolshakov Andrey, and Fedorov Alexey. Semiautomatic visual-attention modeling and its application to video compression. *IEEE International Conference on Image Processing*, pages 1105–1109, 2014.

[GFD06] Guorong Wu, Feihu Qi, and Dinggang Shen. Learning-based deformable registration of MR brain images. *IEEE Transactions on Medical Imaging*, 25(9):1145–1157, September 2006.

[GG16] Yarin Gal and Zoubin Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *NIPS*, 2016.

[GMZ08] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, 2008.

[GN17] Yuan Gao and J Alison Noble. Detection and Characterization of the Fetal Heartbeat in Free-hand Ultrasound Sweeps with Weakly-supervised Two-streams Convolutional Networks. *MICCAI*, 2017.

[GPM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[GSC01] E. Garne, C. Stoll, and M. Clementi. Evaluation of prenatal diagnosis of congenital heart diseases by ultrasound: Experience from 20 European registries. *Ultrasound in Obstetrics and Gynecology*, 17(5):386–391, 2001.

[GSK18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*, 2018.

[GZ09] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP*, 19(1):185–198, 2009.

[GZ10] C. Guo and L. Zhang. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE Transactions on Image Processing*, 19(1):185–198, January 2010.

[HA17] Roxane Holt and Jacques S. Abramowicz. Quality and Safety of Obstetric Practices Using New Modalities- Ultrasound, MR, and CT. *Clin Obstet Gynecol*, 60(3):546–561, September 2017.

[HAM11] Laura E. Houston, Jenifer Allsworth, and George A. Macones. Ultrasound is safe... right?: Resident and maternal-fetal medicine

fellow knowledge regarding obstetric ultrasound safety. *J Ultrasound Med*, 30(1):21–27, January 2011.

[HBNZ17] Weilin Huang, Christopher P. Bridge, J. Alison Noble, and Andrew Zisserman. Temporal HeartNet: Towards Human-Level Automatic Analysis of Fetal Cardiac Screening Video. *MICCAI*, 2017.

[HCSA+16] L. Höglund Carlsson, S. Saltvedt, B.-M. Anderlid, J. Westerlund, C. Gillberg, M. Westgren, and E. Fernell. Prenatal ultrasound and childhood autism: Long-term follow-up after a randomized controlled trial of first- vs second-trimester ultrasound. *Ultrasound Obstet Gynecol*, 48(3):285–288, September 2016.

[Hen03] John M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504, 2003.

[HHM91] F P Hadlock, R B Harrist, and J Martinez-Poyer. In utero analysis of fetal growth: A sonographic weight standard. *Radiology*, 181(1):129–133, October 1991.

[HKP07] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NeurIPS*, 2007.

[HKVBS15] Sayed Hossein Khatoonabadi, Nuno Vasconcelos, Ivan V Bajic, and Yufeng Shan. How many bits does it take for a stimulus to be salient? In *CVPR*, 2015.

[HS97] Sepp Hochreiter and Jurgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[HSBZ15a] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015.

[HSBZ15b] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015.

[HSLF17] Yasmin Halwani, Septimiu E. Salcudean, Victoria A. Lessoway, and Sidney S. Fels. Enhancing Zoom and Pan in Ultrasound Machines with a Multimodal Gaze-based Interface. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1648–1654, May 2017.

[HSS17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *CVPR*, 2017.

[HTE+13] Asbjørn Hróbjartsson, Ann Sofia Skou Thomsen, Frida Emanuelsson, Britta Tendal, Jørgen Hilden, Isabelle Boutron, Philippe Ravaud, and Stig Brorson. Observer bias in randomized clinical trials with measurement scale outcomes: A systematic review of trials with both blinded and nonblinded assessors. *CMAJ*, 185(4):E201–E211, March 2013.

[HTGP08] Richard Housden, Graham M Treece, Andrew H Gee, and Richard W Prager. Calibration of an orientation sensor for freehand 3D ultrasound and its use in a hybrid acquisition system. *BioMed Eng OnLine*, 7(1):5, 2008.

[HXAN18] Ruobing Huang, Weidi Xie, and J. Alison Noble. VP-Nets : Efficient automatic localization of key brain structures in 3D fetal neurosonography. *Med Image Anal*, 47:127–139, July 2018.

[HZ09] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. In *NeurIPS*, 2009.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. In *arXiv:1603.05027 [Cs]*, March 2016.

[IKN98] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998.

[IS15] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ICML*, 2015.

[ISN+13] Christos Ioannou, Ippokratis Sarris, Raffaele Napolitano, Eric Ohuma, M. Kassim Javaid, and Aris T. Papageorghiou. A longitudinal study of normal fetal femur volume. *Prenat Diagn*, 33(11):1088–1094, November 2013.

[ITO+12] C. Ioannou, K. Talbot, E. Ohuma, I. Sarris, J. Villar, A. Conde-Agudelo, and A. T. Papageorghiou. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. *BJOG*, 119(12):1425–1439, November 2012.

[JCA+20] Jianbo Jiao, Yifan Cai, Mohammad Alsharid, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Self-Supervised Contrastive Video-Speech Representation Learning for Ultrasound. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020.

[JDD+20] J. Jiao, R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Self-Supervised Representation Learning for Ultrasound Video. In

2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pages 1847–1850, April 2020.

[JDT12] Tilke Judd, Frédo Durand, and Antonio Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations. *Mit-Csail-Tr-2012*, 1:1–7, 2012.

[JEDT09] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to Predict Where Humans Look. In *ICCV*, 2009.

[JHDZ15a] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, 2015.

[JHDZ15b] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in Context. In *CVPR*, June 2015.

[Jia18] Sen Jia. EML-NET:An Expandable Multi-Layer NETwork for Saliency Prediction. *arXiv:1805.01047 [cs]*, May 2018.

[JL19] Piotr Jarosik and Marcin Lewandowski. Automatic Ultrasound Guidance Based on Deep Reinforcement Learning. In *IEEE International Ultrasonics Symposium (IUS)*, pages 475–478, 2019.

[JMV16] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-End Saliency Mapping via Probability Distribution Prediction. *CVPR*, 2016.

[JPK+18] J. Jang, Y. Park, B. Kim, S. M. Lee, J.-Y. Kwon, and J. K. Seo. Automatic Estimation of Fetal Abdominal Circumference From Ultrasound Images. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1512–1520, September 2018.

[JXL+18] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. DeepVS: A Deep Learning Based Video Saliency Prediction Approach. In *European Conference on Computer Vision*, 2018.

[JXW18] Lai Jiang, Mai Xu, and Zulin Wang. Predicting Video Saliency with Object-to-Motion CNN and Two-layer Convolutional LSTM. In *European Conference on Computer Vision (ECCV)*, 2018.

[JXZ14] Ming Jiang, Juan Xu, and Qi Zhao. Saliency in crowd. *Lecture Notes in Computer Science*, 8695 LNCS(7):17–32, 2014.

[KAB15] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations. *IEEE Trans. Image Process.*, 26(9):4446–4456, 2015.

[KB01] Timor Kadir and Michael Brady. Saliency, Scale and Image Description. *Int. J. Comput. Vision*, 45(2):83–105, November 2001.

[KBJ+12] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT/Tübingen saliency benchmark. 2012.

[KBL+17] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks. In *Information Processing in Medical Imaging*, 2017.

[KBV+15] A. Kosevoi-Tichie, F. Berghea, V. Vlad, M. Abobului, M. Trandafir, T. Gudu, A. Peltea, M. Duna, L. Groseanu, C. Patrascu, and R. Ionescu. THU0583 Does Eye Gaze Tracking Have the Ability to Assess How Rheumatologists Evaluate Musculoskeletal Ultrasound Images? *Annals of the Rheumatic Diseases*, 74(Suppl 2):411–412, June 2015.

[KBZ+20] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big Transfer (BiT): General Visual Representation Learning. In *European Conference on Computer Vision (ECCV)*, 2020.

[KJM+20] Christian Kollmann, Klaus-Vitold Jenderka, Carmel M. Moran, Ferdinando Draghi, J. F. Jimenez Diaz, and Ragnar Sande. EFSUMB Clinical Safety Statement for Diagnostic Ultrasound - (2019 revision). *Ultraschall Med*, 41(4):387–389, August 2020.

[KKKK16] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, and Harini Kannan. Eye Tracking for Everyone. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016.

[KS17] Nitish Shirish Keskar and Richard Socher. Improving Generalization Performance by Switching from Adam to SGD. *arXiv:1712.07628*, 2017.

[KSL18] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? In *arXiv:1805.08974*, 2018.

[KSWF07] Wolf Kienzle, Bernhard Schölkopf, Felix A. Wichmann, and Matthias O. Franz. How to Find Interesting Locations in Video: A Spatiotemporal Interest Point Detector Learned from Human Eye Movements. *Pattern Recognition*, pages 405–414, 2007.

[KT86] F. W. Kremkau and K. J. Taylor. Artifacts in ultrasound imaging. *Journal of Ultrasound in Medicine*, 5(4):227–237, 1986.

[KTB15] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on

ImageNet. In *International Conference on Learning Representations (ICLR)*, 2015.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–227, 1985.

[KWB16] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563*, October 2016.

[KWGB17] Matthias Kummerer, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding Low- and High-Level Contributions to Fixation Prediction. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:4799–4808, 2017.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[LFDA15] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-End Training of Deep Visuomotor Policies. *J. Mach. Learn. Res.*, 17(1):1334–1373, 2015.

[LGDFVP16] Victor Leboran, Anton Garcia-Diaz, Xosé R Fdez-Vidal, and Xosé M Pardo. Dynamic whitening saliency. *IEEE TPAMI*, 39(5):893–907, 2016.

[LH01] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41:3559–3565, 2001.

[LH19] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[LKH+18] Yuanwei Li, Bishesh Khanal, Benjamin Hou, Amir Alansary, Juan J. Cerrolaza, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, and Daniel Rueckert. Standard Plane Detection in 3D Fetal Ultrasound Using an Iterative Transformation Network. In *MICCAI*, pages 392–400, 2018.

[LLR20] Yin Li, Miao Liu, and James M. Rehg. In the Eye of the Beholder: Gaze and Actions in First Person Video. *arXiv:2006.00626 [cs]*, October 2020.

[LMLCBT06] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE TPAMI*, 28(5):802–817, 2006.

[LMN+19] Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Kevin McGuinness, Xavier Giro-i Nieto, and Noel E. O'Connor. Simple vs complex temporal recurrences for video saliency prediction. In *BMVC*, 2019.

[LRL+10] Kaicheng Liang, Albert J. Rogers, Edward D. Light, Daniel von Allmen, and Stephen W. Smith. Three-Dimensional Ultrasound Guidance of Autonomous Robotic Breast Biopsy: Feasibility Study. *Ultrasound in Medicine & Biology*, 36(1):173–177, January 2010.

[LSXW16] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016.

[LTM+18] Z. Li, I. Tong, L. Metcalf, C. Hennessey, and S. E. Salcudean. Free Head Movement Eye Gaze Contingent Ultrasound Interfaces for the da Vinci Surgical System. *IEEE Robotics and Automation Letters*, 3(3):2137–2143, July 2018.

[LWS+17] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting Batch Normalization For Practical Domain Adaptation. In *ICLR Workshops*, 2017.

[LWSS19] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE TIP*, 2019.

[Mar05] K. Marsál. The output display standard: Has it missed its target? *Ultrasound Obstet Gynecol*, 25(3):211–214, March 2005.

[MBE+15] I. Monier, B. Blondel, A. Ego, M. Kaminiski, F. Goffinet, and J. Zeitlin. Poor effectiveness of antenatal detection of fetal growth restriction and consequences for obstetric management and neonatal outcomes: A French national study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 122(4):518–527, 2015.

[MBN+17] M. A. Maraci, C. P. Bridge, R. Napolitano, A. Papageorghiou, and J. A. Noble. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. *Med. Image Anal.*, 37:22–36, 2017.

[MBS19] Fausto Milletari, Vighnesh Birodkar, and Michal Sofka. Straight to the Point: Reinforcement Learning for User Guidance in Ultrasound. In *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis (MICCAI Workshops)*, pages 3–10, 2019.

[MBSK16] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 574–582, Las Vegas, NV, USA, June 2016. IEEE.

[MC19] Kyle Min and Jason J. Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *ICCV*, 2019.

[MET98] M. Mongelli, S. Ek, and R. Tambyrajia. Screening for fetal growth restriction: A mathematical model of the effect of time interval and ultrasound error. *Obstet Gynecol*, 92(6):908–912, December 1998.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[MIM+13] Ammar Safwan Bin Mustafa, Takashi Ishii, Yoshiki Matsunaga, Ryu Nakadate, Hiroyuki Ishii, Kouji Ogawa, Akiko Saito, Motoaki Sugawara, Kiyomi Niki, and Atsuo Takanishi. Development of robotic system for autonomous liver screening using ultrasound scanning device. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 804–809, 2013.

[MKC10] R. Mebarki, A. Krupa, and F. Chaumette. 2-D Ultrasound Probe Complete Guidance by Visual Servoing Using Image Moments. *IEEE Trans. Robot.*, 26(2):296–306, 2010.

[MPG+09] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International journal of computer vision*, 82(3):231, 2009.

[MS12] Stefan Mathe and Cristian Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. *Lecture Notes in Computer Science*, 7573 LNCS(PART 2):842–856, 2012.

[MS15] Stefan Mathe and Cristian Sminchisescu. Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1408–1424, 2015.

[MSBH18] Kamal Mahtani, Elizabeth A. Spencer, Jon Brassey, and Carl Heneghan. Catalogue of bias: Observer bias. *BMJ Evidence-Based Medicine*, 23(1):23–24, February 2018.

[MSHH11] Parag K. Mital, Tim J. Smith, Robin L. Hill, and John M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011.

[MV09] Vijay Mahadevan and Nuno Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE TPAMI*, 32(1):171–177, 2009.

[MZH16] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *ECCV*, 2016.

[NBO+15] Dragos Nemescu, Anca Berescu, Mircea Onofriescu, Dan Bogdan Navolan, and Cristian Rotariu. Safety Indices during Fetal Echocardiography at the Time of First-Trimester Scan Are Machine Dependent. *PLoS One*, 10(5):e0127570, 2015.

[NDO+14] R. Napolitano, J. Dhami, E. O. Ohuma, C. Ioannou, A. Conde-Agudelo, S. H. Kennedy, J. Villar, and A. T. Papageorghiou. Pregnancy dating by fetal crown-rump length: A systematic review of charts. *BJOG*, 121(5):556–565, April 2014.

[NFAC09] Thomas R. Nelson, J. Brian Fowlkes, Jacques S. Abramowicz, and Charles C. Church. Ultrasound biosafety considerations for the practicing sonographer and sonologist. *J Ultrasound Med*, 28(2):139–150, February 2009.

[NHS18] NHS. Fetal Anomaly Screening Programme Handbook, 2018.

[NM17] Thuyen Ngo and B S Manjunath. Saccade gaze prediction using a recurrent neural network. In *ICIP*, 2017.

[OSPD11] Yangming Ou, Aristeidis Sotiras, Nikos Paragios, and Christos Davatzikos. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. *Medical Image Analysis*, 15(4):622–639, August 2011.

[PCS+18] Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile Autonomous Driving using End-to-End Deep Imitation Learning. In *Proceedings of Robotics: Science and Systems*, 2018.

[PFAG19] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling Human Motion with Quaternion-based Neural Networks. In *BMVC*, 2019.

[PFM+17] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv:1701.01081*, 2017.

[PMS+16] Junting Pan, Kevin McGuinness, Elisa Sayrol, Noel O'Connor, and Xavier Giro-i-Nieto. Shallow and Deep Convolutional Networks for Saliency Prediction. *arXiv:1603.00845 [cs]*, March 2016.

[POA+14] Aris T. Papageorghiou, Eric O. Ohuma, Douglas G. Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A. Jaffer, Enrico

Bertino, Michael G. Gravett, Manorama Purwar, J. Alison Noble, Ruyan Pang, Cesar G. Victora, Fernando C. Barros, Maria Carvalho, Laurent J. Salomon, Zulfiqar A. Bhutta, Stephen H. Kennedy, and José Villar. International standards for fetal growth based on serial ultrasound measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*, 384(9946):869–879, September 2014.

[PSGiN+16] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016.

[PSI+13] A. T. Papageorghiou, I. Sarris, C. Ioannou, T. Todros, M. Carvalho, G. Pilu, L. J. Salomon, and International Fetal and Newborn Growth Consortium for the 21st Century. Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project. *BJOG*, 120 Suppl 2:27–32, v, September 2013.

[PSJ+18] Raphael Prevost, Mehrdad Salehi, Simon Jagoda, Navneet Kumar, Julian Sprung, Alexander Ladikos, Robert Bauer, Oliver Zettinig, and Wolfgang Wein. 3D freehand ultrasound without external tracking using deep learning. *Medical Image Analysis*, 48:187–202, August 2018.

[PUL] Perception Ultrasound by Learning Sonographic Experience, European Research Council (ERC) Advanced Grant. https://cordis.europa.eu/project/id/694581.

[QWL+20] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-Batch Training with Batch-Channel Normalization and Weight Standardization. *arXiv:1903.10520 [cs]*, August 2020.

[Ray78] Keith Rayner. Eye movements in reading and information processing. *Psychological Bulletin*, 85(3):618–660, 1978.

[Ray98] K Rayner. Eye movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422, 1998.

[RCH+16] A. Rempen, R. Chaoui, M. Häusler, K.-O. Kagan, P. Kozlowski, C. von Kaisenberg, and J. Wisser. Quality Requirements for Ultrasound Examination in Early Pregnancy (DEGUM Level I) between 4+0 and 13+6 Weeks of Gestation. *Ultraschall Med*, 37(6):579–583, December 2016.

[RDGF15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. 2015.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Miccai*, pages 234–241, 2015.

[RFK+14] S. Rueda, S. Fathima, C. L. Knight, M. Yaqub, A. T. Papageorghiou, B. Rahmatullah, A. Foi, M. Maggioni, A. Pepe, J. Tohka, R. V. Stebbing, J. E. McManigle, A. Ciurte, X. Bresson, M. B. Cuadra, C. Sun, G. V. Ponomarev, M. S. Gelfand, M. D. Kazanov, C. Wang, H. Chen, C. Peng, C. Hung, and J. A. Noble. Evaluation and Comparison of Current Fetal Ultrasound Image Segmentation Methods for Biometric Measurements: A Grand Challenge. *IEEE Transactions on Medical Imaging*, 33(4):797–813, April 2014.

[RGSZM13] Dmitry Rudoy, Dan B Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, 2013.

[RKK+17] K. Retz, S. Kotopoulis, T. Kiserud, K. Matre, G. E. Eide, and R. Sande. Measured acoustic intensities for clinical diagnostic ultrasound transducers and correlation with thermal index. *Ultrasound Obstet Gynecol*, 50(2):236–241, August 2017.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[Roy13] Royal College of Obstetricians and Gynaecologists. The Investigation and Management of the Small–for–Gestational–Age Fetus (Green-top Guideline No. 31). https://www.rcog.org.uk/globalassets/documents/guidelines/gtg_31.pdf, 2013.

[RPN11] Bahbibi Rahmatullah, Aris Papageorghiou, and J. Alison Noble. Automated Selection of Standardized Planes from Ultrasound Volume. In *Machine Learning in Medical Imaging (MICCAI Workshop)*, volume 7009, pages 35–42, 2011.

[RSF19] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):801–814, April 2019.

[RW16] Md Atiqur Rahman and Yang Wang. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. In *Advances in Visual Computing*, volume 10072, pages 234–244. 2016.

[SA08] Eyal Sheiner and Jacques S. Abramowicz. Clinical end users worldwide show poor knowledge regarding safety issues of ultrasound during pregnancy. *J Ultrasound Med*, 27(4):499–501, April 2008.

[SA09] Eyal Sheiner and Jacques S. Abramowicz. Acoustic output as measured by thermal and mechanical indices during fetal nuchal

translucency ultrasound examinations. *Fetal Diagn Ther*, 25(1):8–10, 2009.

[SAB+11] L. J. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. L. Johnsen, K. Kalache, K. Y. Leung, G. Malinger, H. Munoz, F. Prefumo, A. Toi, and W. Lee. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound Obstet. Gynecol.*, 37(1):116–126, 2011.

[SAB+13] L. J. Salomon, Z. Alfirevic, C. M. Bilardo, G. E. Chalouhi, T. Ghi, K. O. Kagan, T. K. Lau, A. T. Papageorghiou, N. J. Raine-Fenning, J. Stirnemann, S. Suresh, A. Tabor, I. E. Timor-Tritsch, A. Toi, and G. Yeo. ISUOG practice guidelines: Performance of first-trimester fetal ultrasound scan. *Ultrasound Obstet Gynecol*, 41(1):102–113, January 2013.

[Sac79] David L. Sackett. Bias in analytic research. *Journal of Chronic Diseases*, 32(1):51–63, January 1979.

[Saf09] Safety Group of the British Medical Ultrasound Society. Guidelines for the safe use of diagnostic ultrasound equipment. https://www.bmus.org/static/uploads/resources/BMUS-Safety-Guidelines-2009-revision-FINAL-Nov-2009.pdf, 2009.

[Sal11] K. Å Salvesen. Ultrasound in pregnancy and non-right handedness: Meta-analysis of randomized trials. *Ultrasound Obstet Gynecol*, 38(3):267–271, September 2011.

[SBA+15] Sachita Shah, Blaise A. Bellows, Adeyinka A. Adedipe, Jodie E. Totten, Brandon H. Backlund, and Dana Sajed. Perceived barriers in the use of ultrasound in developing countries. *Crit. Ultrasound J.*, 7:11, 2015.

[SBD+06] L. J. Salomon, J. P. Bernard, M. Duyme, B. Doris, N. Mas, and Y. Ville. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound Obstet Gynecol*, 27(1):34–40, January 2006.

[SBM+18] Matthew Sinclair, Christian F. Baumgartner, Jacqueline Matthew, Wenjia Bai, Juan Cerrolaza Martinez, Yuanwei Li, Sandra Smith, Caroline L. Knight, Bernhard Kainz, Jo Hajnal, Andrew P. King, and Daniel Rueckert. Human-level Performance On Automatic Head Biometrics In Fetal Ultrasound Using Fully Convolutional Neural Networks. In *EMBC*, 2018.

[SCW+15] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In

International Conference on Neural Information Processing Systems, 2015.

[SDC+19] H Sharma, R Droste, P Chatelain, L Drukker, A Papageorghiou, and J Alison Noble. Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans. In *IEEE ISBI*, 2019.

[SDC+21] Harshita Sharma, Lior Drukker, Pierre Chatelain, Richard Droste, Aris T. Papageorghiou, and J. Alison Noble. Knowledge Representation and Learning of Operator Clinical Workflow from Full-length Routine Fetal Ultrasound Scan Videos. *Medical Image Analysis*, page 101973, January 2021.

[SDD+20] H. Sharma, L. Drukker, R. Droste, P. Chatelain, A. T. Papageorghiou, and J. A. Noble. OC10.02: Task-evoked pupillary response as an index of cognitive workload of sonologists undertaking fetal ultrasound. *Ultrasound in Obstetrics & Gynecology*, 56(S1):28–28, 2020.

[SF03] Yaoru Sun and Robert Fisher. Object-based visual attention for computer vision. *Artificial intelligence*, 146(1):77–123, 2003.

[SG00] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78, November 2000.

[SG02] Kenneth F. Schulz and David A. Grimes. Blinding in randomised trials: Hiding who got what. *Lancet*, 359(9307):696–700, February 2002.

[SHZ+18a] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[SHZ+18b] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 2018.

[SIO+13] I. Sarris, C. Ioannou, E. O. Ohuma, D. G. Altman, L. Hoch, C. Cosgrove, S. Fathima, L. J. Salomon, and A. T. Papageorghiou. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics & Gynaecology*, 120(s2):33–37, 2013.

[SLA+11] K. Salvesen, C. Lees, J. Abramowicz, C. Brezinka, G. Ter Haar, K. Maršál, and Board of International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). ISUOG statement on the safe use of Doppler in the 11 to 13 +6-week fetal ultrasound examination. *Ultrasound Obstet Gynecol*, 37(6):628, June 2011.

[SM09] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009.

[SMEK13] Ragnar K. Sande, Knut Matre, Geir E. Eide, and Torvid Kiserud. The effects of reducing the thermal index for bone from 1.0 to 0.5 and 0.1 on common obstetric pulsed wave Doppler measurements in the second half of pregnancy. *Acta Obstet Gynecol Scand*, 92(7):790–796, July 2013.

[SO19] A. Sotiriadis and A. O. Odibo. Systematic error and cognitive bias in obstetric ultrasound. *Ultrasound in Obstetrics & Gynecology*, 53(4):431–435, 2019.

[SOC+18] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention-Gated Networks for Improving Ultrasound Scan Plane Detection. In *MIDL*, 2018.

[SWZ+18] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection. *ECCV*, 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. *NIPS*, 2014.

[SZ15] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. pages 4489–4497. IEEE, December 2015.

[TC17] Jen-Chieh Tsai and Jen-Tzung Chien. Adversarial domain separation and adaptation. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, September 2017.

[tH12] G. ter Haar. *The Safe Use of Ultrasound in Medical Diagnosis*. British Institute of Radiology, London, 2012.

[TKS+18] Nicolas Toussaint, Bishesh Khanal, Matthew Sinclair, Alberto Gomez, Emily Skelton, Jacqueline Matthew, and Julia A. Schnabel. Weakly Supervised Localisation for Fetal Ultrasound Images. In *4th Workshop on Deep Learning for Medical Image Analysis (MICCAI Workshops)*, 2018.

[Tom06] David A. Toms. The mechanical index, ultrasound practices, and the ALARA principle. *J Ultrasound Med*, 25(4):560–561; author reply 561–562, April 2006.

[TVM+09] M. R. Torloni, N. Vedmedovska, M. Merialdi, A. P. Betrán, T. Allen, R. González, L. D. Platt, and ISUOG-WHO Fetal Growth Study Group. Safety of ultrasonography in pregnancy: WHO systematic review of the literature and meta-analysis. *Ultrasound Obstet Gynecol*, 33(5):599–608, May 2009.

[TWBX18] Grzegorz Toporek, Haibo Wang, Marcin Balicki, and Hua Xie. Autonomous image-based ultrasound probe positioning via deep learning. In *Hamlyn Symposium on Medical Robotics*, 2018.

[UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arxiv:1607.08022*, 2016.

[vCH+16] C. von Kaisenberg, R. Chaoui, M. Häusler, K. O. Kagan, P. Kozlowski, E. Merz, A. Rempen, H. Steiner, S. Tercanli, J. Wisser, and K.-S. Heling. Quality Requirements for the early Fetal Ultrasound Assessment at 11-13+6 Weeks of Gestation (DEGUM Levels II and III). *Ultraschall Med*, 37(3):297–302, June 2016.

[VDC12] Eleonora Vig, Michael Dorr, and David Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. *Lecture Notes in Computer Science*, 7578 LNCS(PART 7):84–97, 2012.

[VDC14a] E. Vig, M. Dorr, and D. Cox. Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, June 2014.

[VDC14b] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014.

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9(Nov):2579–2605, 2008.

[VGB+18] Jose Villar, Francesca Giuliani, Fernando Barros, Paola Roggero, Irma Alejandra Coronado Zarco, Maria Albertina S. Rego, Roseline Ochieng, Maria Lorella Gianni, Suman Rao, Ann Lambert, Irina Ryumina, Carl Britto, Deepak Chawla, Leila Cheikh Ismail, Syed Rehan Ali, Jane Hirst, Jagjit Singh Teji, Karim Abawi, Jacqueline Asibey, Josephine Agyeman-Duah, Kenny McCormick, Enrico Bertino, Aris T. Papageorghiou, Josep Figueras-Aloy, Zulfiqar Bhutta, and Stephen Kennedy. Monitoring the Postnatal Growth of Preterm Infants: A Paradigm Change. *Pediatrics*, 141(2), February 2018.

[VSJR17] Sepehr Valipour, Mennatullah Siam, Martin Jagersand, and Nilanjan Ray. Recurrent Fully Convolutional Networks for Video Segmentation. In *ArxiV*, 2017.

[VTC+03] A. Vilchis, J. Troccaz, P. Cinquin, K. Masuda, and F. Pellissier. A new robot architecture for tele-echography. *IEEE Trans. Robot. Autom.*, 19(5):922–926, 2003.

[WCL+17] L. Wu, J. Cheng, S. Li, B. Lei, T. Wang, and D. Ni. FUIQA: Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks. *IEEE Transactions on Cybernetics*, 47(5):1336–1349, May 2017.

[WDJ+20] Yipei Wang, Richard Droste, Jianbo Jiao, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Differentiating Operator Skill During Routine Fetal Ultrasound Scanning Using Probe Motion Tracking. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis (MICCAI Workshops)*, 2020.

[WH18] Yuxin Wu and Kaiming He. Group Normalization. *ECCV*, 2018.

[WHN+19] Shuangyi Wang, James Housden, Yohan Noh, Davinder Singh, Anisha Singh, Emily Skelton, Jacqueline Matthew, Cornelius Tan, Junghwan Back, Lukas Lindenroth, Alberto Gomez, Nicolas Toussaint, Veronika Zimmer, Caroline Knight, Tara Fletcher, David Lloyd, John Simpson, Dharmintra Pasupathy, Hongbin Liu, Kaspar Althoefer, Joseph Hajnal, Reza Razavi, and Kawal Rhode. Robotic-Assisted Ultrasound for Fetal Imaging: Evolution from Single-Arm to Dual-Arm System. In *Towards Autonomous Robotic Systems*, pages 27–38, 2019.

[WS18] Wenguan Wang and Jianbing Shen. Deep Visual Attention Prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2018.

[WSG+18] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting Video Saliency: A Large-scale Benchmark and a New Model. *CVPR*, 2018.

[WSS18] Wenguan Wang, Jianbing Shen, and Ling Shao. Video Salient Object Detection via Fully Convolutional Networks. *IEEE Transactions on Image Processing*, 27(1):38–49, January 2018.

[WSX+19] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE TPAMI*, 2019.

[WWP14] Chia-Chien Wu, Farahnaz Ahmed Wick, and Marc Pomplun. Guidance of visual attention by semantic information in real-world scenes. *Front. Psychol.*, 5:54, 2014.

[WXL+17] L. Wu, Y. Xin, S. Li, T. Wang, P. Heng, and D. Ni. Cascaded Fully Convolutional Networks for automatic prenatal ultrasound image segmentation. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 663–666, April 2017.

[WYD+19] X. Wang, X. Yang, H. Dou, S. Li, P. Heng, and D. Ni. Joint Segmentation and Landmark Localization of Fetal Femur in Ultrasound Volumes. In *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 1–5, May 2019.

[XGD+17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*, 2017.

[XLOW16] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. In *arXiv:1604.07528 [Cs]*, 2016.

[YK16] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*, 2016.

[YKPN15] Mohammad Yaqub, Brenda Kelly, A.T Papageorghiou, and J. Alison Noble. Guided Random Forests for Identification of Key Fetal Anatomy and Image Categorization in Ultrasound Scans. *Lecture Notes in Computer Science*, 9351:687–694, 2015.

[YKPN17a] Mohammad Yaqub, Brenda Kelly, Aris T. Papageorghiou, and J. Alison Noble. A Deep Learning Solution for Automatic Fetal Neurosonographic Diagnostic Plane Verification Using Clinical Standard Constraints. *Ultrasound in Medicine & Biology*, 2017.

[YKPN17b] Mohammad Yaqub, Brenda Kelly, Aris T. Papageorghiou, and J. Alison Noble. A Deep Learning Solution for Automatic Fetal Neurosonographic Diagnostic Plane Verification Using Clinical Standard Constraints. *Ultrasound in Medicine & Biology*, 43(12):2925–2933, 2017.

[YLJL19] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A dilated inception network for visual saliency prediction. *IEEE TMM*, 2019.

[YMG10] Yu Yuanlong, G K I Mann, and R G Gosine. An Object-Based Visual Attention Model for Robotic Applications. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(5):1398–1412, 2010.

[YRK+16] Mohammad Yaqub, Sylvia Rueda, Anil Kopuri, Pedro Melo, A. T. Papageorghiou, Peter B. Sullivan, Kenneth McCormick, and J. Alison Noble. Plane Localization in 3-D Fetal Neurosonography for Longitudinal Analysis of the Developing Brain. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1–9, 2016.

[YYL+18] X. Yang, L. Yu, S. Li, H. Wen, D. Luo, C. Bian, J. Qin, D. Ni, and P. Heng. Towards Automated Semantic Segmentation in Prenatal Volumetric Ultrasound. *IEEE Transactions on Medical Imaging*, pages 1–1, 2018.

[ZJCL18] Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson WH Lau. Task-driven webpage saliency. In *ECCV*, 2018.

[ZLR+13] Sheng-hua Zhong, Yan Liu, Feifei Ren, Jinghuan Zhang, and Tongwei Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *AAAI*, 2013.

[ZRS20] Hongzhi Zhu, Robert N. Rohling, and Septimiu E. Salcudean. Hand-eye coordination-based implicit re-calibration method for gaze tracking on ultrasound machines: A statistical approach. *Int J CARS*, April 2020.

[ZSR19] Hongzhi Zhu, Septimiu E. Salcudean, and Robert N. Rohling. A novel gaze-supported multimodal human–computer interaction for ultrasound machines. *Int J CARS*, 14(7):1107–1115, July 2019.

[ZSV14] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent Neural Network Regularization. *arXiv:1409.2329*, 2014.

[ZTW+21] Yan Zeng, Po-Hsiang Tsui, Weiwei Wu, Zhuhuang Zhou, and Shuicai Wu. Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net. *J Digit Imaging*, January 2021.