Medical Image Analysis 65 (2020) 101762

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Spatio-temporal visual attention modelling of standard biometry plane-finding navigation

Yifan Cai[a], Richard Droste[a], Harshita Sharma[a], Pierre Chatelain[a], Lior Drukker[b], Aris T. Papageorghiou[b], J. Alison Noble[a]

[a] *Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, OX3 7DQ, UK*
[b] *Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, OX3 9DU, UK*

ARTICLE INFO

Article history:
Received 22 October 2019
Revised 15 June 2020
Accepted 18 June 2020
Available online 20 June 2020

MSC:
41A05
41A10
65D05
65D17

Keywords:
Fetal ultrasound
Gaze tracking
Multi-task learning
Saliency prediction
Standard plane detection

ABSTRACT

We present a novel multi-task neural network called Temporal SonoEyeNet (TSEN) with a primary task to describe the visual navigation process of sonographers by learning to generate visual attention maps of ultrasound images around standard biometry planes of the fetal abdomen, head (trans-ventricular plane) and femur. TSEN has three components: a feature extractor, a temporal attention module (TAM), and an auxiliary video classification module (VCM). A soft dynamic time warping (sDTW) loss function is used to improve visual attention modelling. Variants of the model are trained on a dataset of 280 video clips, each containing one of the three biometry planes and lasting 3–7 seconds, with corresponding real-time recorded gaze tracking data of an experienced sonographer. We report the performances of the different variants of TSEN for visual attention prediction at standard biometry plane detection. The best model performance is achieved using bi-directional convolutional long-short term memory (biCLSTM) in both TAM and VCM, and it outperforms a previous spatial model on all static and dynamic saliency metrics. As an auxiliary task to validate the clinical relevance of the visual attention modelling, the predicted visual attention maps were used to guide standard biometry plane detection in consecutive US video frames. All spatio-temporal TSEN models achieve higher scores compared to a spatial-only baseline; the best performing TSEN model achieves F1 scores on these standard biometry planes of 83.7%, 89.9% and 81.1%, respectively.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

At least 60% of neonatal deaths worldwide are associated with low birth weight and therefore identification of growth restricted fetuses is clinically important (Lawn et al., 2005). Obstetric ultrasound (US) is the chosen modality for pregnancy imaging due to its non-invasiveness, absence of radiation, high accessibility, high reliability and low cost (Abramowicz, 2013). Obstetric sonography largely relies on identification and acquisition of standardized anatomical landmarks followed by performance of an accurate measurement (Salomon et al., 2006). There are three basic standard biometry measurement planes which are the head circumference plane(HCP), the abdominal circumference plane (ACP), and the femur length plane (FLP) (Papageorghiou et al., 2014). These standard planes are captured in the majority of second- and third-trimester ultrasound scans. However, the detection of these standard planes requires a high level of operator skills, leading to common problems such as intra- and inter-observer variability (Sarris et al., 2012). Inaccurate measurement can lead to erroneous detection of growth restriction and thus to unnecessary intervention, maternal anxiety and iatrogenic perinatal morbidity; or may lead to inadvertently overlooking growth-restricted fetuses and classifying them as normal (Mongelli et al., 1998). As a result, automated methods, most recently based on random forest (Yaqub et al., 2015) and deep neural network-based models (Baumgartner et al., 2017), have been investigated for standard plane detection in fetal ultrasound motivated by the potential to increase work flow efficiency as well as to reduce variability in fetal ultrasound image acquisition. In addition, the standard biometry plane *finding navigation* process requires a sonographer to choose a single frame for biometric measurement from a proximity with similar video frame contents, demanding a high level of hand-eye coordination. Plane-finding navigation process has not been studied in the medical image analysis literature to our

E-mail address: yifan.cai@eng.ox.ac.uk (Y. Cai).

https://doi.org/10.1016/j.media.2020.101762
1361-8415/© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

knowledge. This paper is the first to investigate the visual attention of sonographers in standard biometry plane-finding navigation.

In this paper, we propose an original gaze-based spatio-temporal network called **Temporal SonoEyeNet (TSEN)** with the primary task to model sonographer visual attention during standard biometry plane-finding navigation, and validate the predicted visual attention maps on an auxiliary task of gaze-guided standard biometry plane detection. The models are built using a novel dataset containing real-time screen recordings of US anomaly scans coupled with simultaneous gaze-tracking. This paper (1) investigates architectures to model temporal visual attention variations of a sonographer in 2-D US videos using a Temporal Attention Module (TAM), and the best-performing model is based on a bi-directional convolutional long-short term memory (Xingjian et al., 2015) (biCLSTM) as a recurrent module; (2) further improves spatio-temporal visual attention prediction by training TSEN with a novel loss function for saliency prediction, the soft Dynamic Time Warping (sDTW) (Cuturi and Blondel, 2017) loss, for visual attention alignment; and (3) uses the learnt visual attention maps to guide standard plane detection on all three standard biometry planes: ACP, HCP and FLP.

2. Related works

2.1. Visual attention modelling

Machine Learning methods that predict human visual attention are referred to as *saliency prediction models* (Treisman and Gelade, 1980). Early saliency prediction models followed the feature integration theory (FIT) (Treisman and Gelade, 1980), defining saliency by the fusion of several hand-crafted features, such as edges, color, disparity and direction of movement (Koch and Ullman, 1987), extracted at multiple scales (Itti et al., 1998). Recent developments of deep learning in computer vision have led to state-of-the-art performance saliency prediction models built on neural networks. Huang et al. (2015) designed a saliency prediction model using convolutional neural networks (CNNs) pre-trained on ImageNet to extract spatial information from static images; the model is fine-tuned using saliency metrics such as the Kullback-Leibler divergence (KLD), Normalized Scanpath Saliency (NSS), and Correlation Coefficient (CC) to predict human visual attention on static images. Wang et al. (2018) proposes a CNN-LSTM (long-short term memory) architecture to exploit both the spatial and temporal information for video saliency prediction. With a supervised attention mechanism, it explicitly captures static saliency information of each frame, which is then fed into an LSTM to focus on learning dynamic information to predict dynamic visual attention.

Human visual attention has been used in a number of ways for image analysis: one class of algorithms record human gaze information in order to perform inter-observer comparisons (Nodine and Kundel, 1987; James et al., 2007; Ahmed, 2014); another class uses recorded human gaze information as input, in addition to medical images, to assist medical image analysis tasks (Ramanathan et al., 2009; Xu et al., 2013; Shanmuga Vadivel et al., 2015). In ultrasound image analysis, Ahmed and Noble (2016) built a vocabulary of visual words trained on SURF descriptors (Bay et al., 2006) extracted around eye fixations to classify head, abdomen and femur image frames. Cai et al. (2018b) built a CNN model for US abdominal standard plane classification assisted by predicted visual attention maps. The visual attention maps were fine-tuned using an adversarial regulariser. Droste et al. (2019a) modelled sonographer visual attention on static ultrasound video frames through visual saliency prediction as well as gaze-point regression. Those models only learn static visual attention by treating each US video frame as an independent image. However, sonographer visual attention changes on subsequent frames, transitioning between key anatomical landmarks. Our hypothesis is that dynamic visual attention models (Xingjian et al., 2015) can learn the temporal transition of visual attention between frames to predict the attention maps on US video (Droste et al., 2019b). This paper models both spatial and temporal visual attention on US video clips using original convolutional and recurrent neural network architectures with a novel soft Dynamic Time Warping (sDTW) loss function to regularize the alignment of predicted and ground truth visual attention during training, which is the first time that sDTW has been used in the context of visual attention modelling.

2.2. Standard plane detection in fetal ultrasound

Chen et al. (2015) developed a model based on AlexNet to detect Fetal Anomaly Screening Programme (FASP) standard planes on a dataset acquired using a sweeping protocol, achieving a precision of 71.4%. Gao et al. (2016) used pre-trained weights of AlexNet on ImageNet to classify fetal US images including fetal skull, abdomen, heart, and demonstrated that the features learnt on natural images could be transferred to an US image dataset. The mean classification accuracy reached 91.5% compared to 87.9%, the performance achieved by a network of the same architecture but initialised using random weights. SonoNet (Baumgartner et al., 2017) was built on VGGNet Simonyan and Zisserman (2014) for FASP standard plane detection (Kirwan, 2010) on routine free-hand US scan on a large dataset of 2694 2D ultrasound examinations. Three variants of the SonoNet were tested. The largest was SonoNet-64, which uses the same architecture as VGG-16, while the others, SonoNet-32 and SonoNet-16, adopt architectures with halved and quartered number of kernels in all layers. The best performing model was SonoNet-64, achieving mean F1 score of 82.8%. Schlemper et al. (2018) introduced a self-gated soft-attention mechanism that allows the network to contextualise local spatial information useful for detection of US standard planes and weakly-supervised object localisation. Cai et al. (2018a) used pre-processed sonographer visual attention maps as an additional input to assist abdominal standard planes detection, and Cai et al. (2018b) further improved the potential clinical usefulness of the previous work by learning to predict attention maps on input US images without compromising standard plane classification accuracy. Droste et al. (2019a) learnt a feature extractor by predicting sonographer visual attention, and then fine-tuned the feature extractor for standard planes detection in a transfer learning manner. These works focused on learning spatial information in images to assist US image classification but forfeited the temporal information in US video.

2.3. Novelty

Similar to Droste et al. (2019a) and Cai et al. (2018b), TSEN explores the inherent information in sonographer gaze-tracking data as a strong prior to guide US standard plane navigation. However, different from Droste et al. (2019a), which takes a two-stage training scheme for plane classification, TSEN is trained end-to-end by employing a multi-task network structure. TSEN learns a feature representation of input US videos for both sonographer visual attention map prediction and standard plane detection. Different from Cai et al. (2018b), which explores spatial visual attention information of a single US image, TSEN extends into the temporal dimension by exploring spatio-temporal feature representations of US videos using variants of convolutional Recurrent Neural Networks. Specifically, it utilizes a soft Dynamic Time Warping (sDTW) loss to regularise the alignement of predicted visual attention maps with ground-truth attention maps, which proves to be an effective regularisation method for visual attention modelling.

3. Methods

3.1. Data collection and processing

This study is part of an on-going project entitled Perception Ultrasound by Learning Sonographic Experience (PULSE). A novel dataset of clinical ultrasound exams with real-time gaze-tracking data is being collected. 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. Data were stored according to local data governance rules.

3.1.1. Data collection

All free-hand ultrasound exams were performed on a GE 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; its LCD monitor has a resolution of 1920 × 1080 pixels and refreshes at a frequency of 60 *Hz*, while the video signal is recorded lossless at 30 Hz. Synchronized eye tracking was undertaken using an eye-tracker (Tobii Eye-tracking Eye Tracker 4C, Danderyd, Sweden) that records the point-of-gaze (relative *x* and *y* coordinates with corresponding timestamps) and 3-D eye position of each eye at a rate of 90 Hz, effectively recording 3 gaze points per frame. The eye tracker was rigidly attached under the display area with a magnetic mounting bracket as per the instruction of the product. The calibration of the eye-tracker was previously studied in Chatelain et al. (2018).

Sonographers were free to adjust the height of the chair and the inclination of the monitor, and operate the ultrasound probe in order to perform ultrasound examinations without being affected by the presence of an eye tracker so that authentic and clinically relevent gaze data were recorded. The eye tracker was calibrated for each sonographer following a 9-point calibration protocol as described in Chatelain et al. (2018). During ultrasound examination, sonographers identify key fetal anatomies and make corresponding measurements according to the UK FASP guidelines for mid-pregnancy ultrasound examinations (Kirwan, 2010). The video signal of the scanner and sonographer gaze tracking data are stored for later processing. Video were stored as.mp4 video files, and each video file was converted to individual video frames stored as.png image files. A normal examination takes 34 ± 14 minutes.

3.1.2. US Scan video processing

A subset of the PULSE dataset containing 93 anomaly scans (280 video clips in total) performed by a single sonographer was selected. An optical character recognition algorithm was used to identify the standard biometry planes on which sonographers made measurements, including the standard ACP, HCP and FLP. Since it is common for sonographers to hold the probe still or make very small movements when a standard plane is found, multiple consecutive frames could contain the same standard plane. It is also common for a sonographer to move the probe away from the already found standard plane to confirm that it is the best available plane before entering freeze frame, a static video frame on which the sonographer makes biometric measurements. Thus, in each anomaly scan, an experienced biomedical engineer manually annotated the location of all such standard biometry planes. Each frame in this dataset was assigned one of the following seven *frame-level* labels: standard AC Plane (std ACP), non-standard abdomen (bg Ab), standard HC plane (std HCP), non-standard head (bg Head), standard FL plane (std FLP), non-standard Femur (bg Femur), and Others. A sonographer's navigation and plane-finding decision-making process for a particular standard biometry plane was represented in a video clip C_x (where x is the plane type), starting from a video frame containing a non-standard view of that anatomy up until the frame before a freeze frame that captures the standard biometry plane. For each anomaly scan, C_A, C_H and C_F were sampled to reflect the navigation refinement process of finding standard ACP, HCP and FLP. An “Others” class video clip C_O that does not contain any clear anatomical structures was also sampled from the scans. A total of 89 C_A, 71 C_H, 76 C_F, and 44 C_O were selected. Each of these clips contains between 200 and 500 frames (approximately corresponding to 6–17 seconds), of which between 10 and 40 frames are standard biometry planes. In total, this dataset contains 1910 ACPs, 2359 HCPs and 2151 FLPs from 22,927 abdomen frames, 24,437 head frames, 12,762 femur frames, and 8982 Other frames. The distribution of frame count for all clips is summarized in Fig. 1. All frames containing Doppler overlay, 3-D/4-D, split-screens, or freeze frames containing bounding boxes/circles were excluded from the dataset. Text information and the Graphical User Interface (GUI) on each frame was cropped out. Examples of the standard planes are shown in Fig. 2.

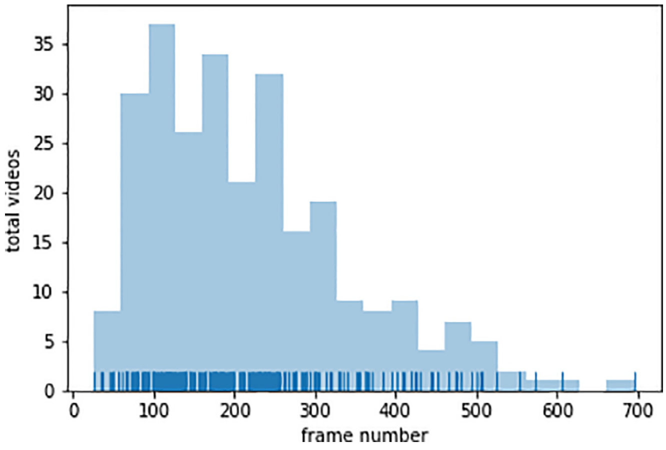


Fig. 1. Frame count distribution for all US video clips used in this dataset.

3.1.3. Gaze data processing

Using the gaze data, binary maps B of the same dimensions as the corresponding video frames were generated, with pixels corresponding to gaze points labelled as 1 and others labelled as zero (0). A *sonographer visual attention map* A was generated for each binary map by convolving it with a truncated Gaussian Kernel $G(\sigma_{x,y})$: $A = B * G(\sigma_{x,y})$, where $G(\sigma_{x,y})$ has 30 *pixels* along *x*, *y*-dimensions corresponding to visual angle of 1.5° with an observer-to-screen distance of 0.5m. A was further normalised such that all pixel values add up to 1. Examples of sonographer visual attention maps overlaid on their corresponding B-mode image for 6 consecutive frames of C_A, C_F, C_H and C_O can be seen in Fig. 3.

3.1.4. Training sample (snippet) generation

We define an ultrasound video snippet as a short video segment extracted from a video clip with defined time depth and skip size to train TSEN. A number of ultrasound video snippets were sampled from the aforementioned video clips, as illustrated schematically in Fig. 4. Time depth is defined as the number of frames in a video snippet that forms an input to the network, and skip size is the number of frames skipped in the original video clip C between consecutive sampled frames. Arrows of different colors and line styles indicate 4 possible ways of sampling from a clip C with the same skip size (*left*), while the stacked frames with different colors and line styles shows the 4 resultant training samples with the same time depth (*right*). In order to model temporal attention of different time scales, time depth of 5, 10, 15, and

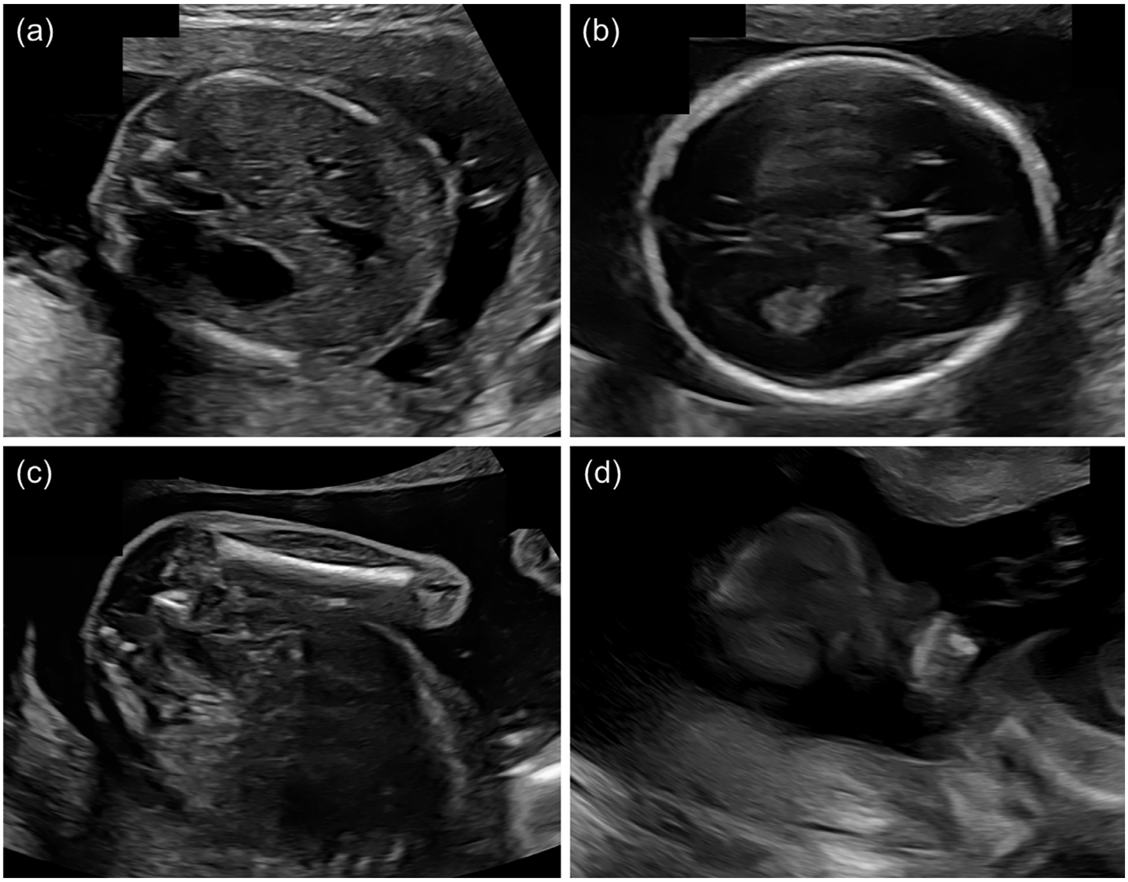


Fig. 2. Standard Biometry Planes and an example of background frame. (a) Abdominal Circumference Plane; (b) Head Circumference Plane; (c) Femur Length Plane; (d) Background.

20 frames were tested. Sampling with skip sizes of 5, 10, 15 and 20 was also considered. Based on empirical analysis, this study uses time depth of 10 for most efficient use of GPU memory, and a skip size of 10, as this skip size strikes a good balance between reducing temporal redundancy and not losing temporal information. Each video snippet was coupled with frame-wise anatomy labels as well as frame-wise sonographer visual attention maps. In addition, a *snippet-level* label (one label per snippet) was assigned for each snippet: if the snippet contained one or more instances of a standard biometry plane, it was labelled as a standard snippet; otherwise a non-standard snippet of that anatomy.

It is worth noticing that not all snippets have corresponding sonographer visual attention maps for every frame in the snippet. The reason for this is that sonographers are not guaranteed to be looking at the screen and they can check the position of ultrasound probe, talk to patients, or move away from the receptive field of the eye tracker. In these cases, no gaze-data are recorded, thus no valid sonographer visual attention maps are generated; a sampled snippet containing frames with no corresponding sonographer visual attention maps was discarded.

3.2. Temporal SonoEyeNet (TSEN) architecture

The previously reported multi-task SonoEyeNet (MSEN) (Cai et al., 2018b) captures spatial information in 2-D US images by predicting sonographer visual attention maps to assist a binary classification task of detecting standard Abdominal Circumference Planes (ACP) from non-standard abdominal images. However, MSEN does not utilize temporal information inherent in 2-D US videos, which is hypothesized to assist sonographers in standard biometry plane detection. This subsection describes Temporal SonoEyeNet (TSEN), which expands on the idea of Multi-task SonoEyeNet by modelling temporal visual attention variations of a sonographer and, in addition, expands detection targets from ACP to all three standard biometry planes: ACP, HCP, and FLP.

3.2.1. Convolutional recurrent neural networks

Vanilla (ungated) Recurrent Neural Networks (RNNs) (Pearlmutter, 1989; Giles et al., 1994) are commonly used to encode temporal or spatio-temporal information. Comparing to vanilla RNNs, *Long-Short Term Memory* (LSTM) (Hochreiter and Schmidhuber, 1997) and *Gated Recurrent Units* (GRUs) (Cho et al., 2014) use gating mechanisms with additional internal recurrence to control the flow of information so that they can solve the long-term dependency problem of vanilla RNNs. LSTMs and GRUs have been successfully used in computer vision (Vinyals et al., 2015) and natural language processing (Bahdanau et al., 2014), but the dot product operation in all gating mechanisms is especially redundant for spatial information such as images with high dimensional feature representations (Xingjian et al., 2015). In input-to-state and state-to-state transitions, no spatial information is encoded. To address this problem, Xingjian et al. (2015) extended the LSTM concept by replacing the dot product operation by a convolution operation in input-to-state and state-to-state transitions, as shown in Fig. 5. Similarly, an extension to GRUs was proposed by Cho et al. (2014) to incorporate convolutional operations.

3.2.2. Network architecture

The architecture of TSEN is described in Fig. 6. The network takes a sample input snippet $\mathbf{X} = [\mathbf{X}^1; \ldots; \mathbf{X}^T] \in [|0, 255|]^{H\times W\times T}$,

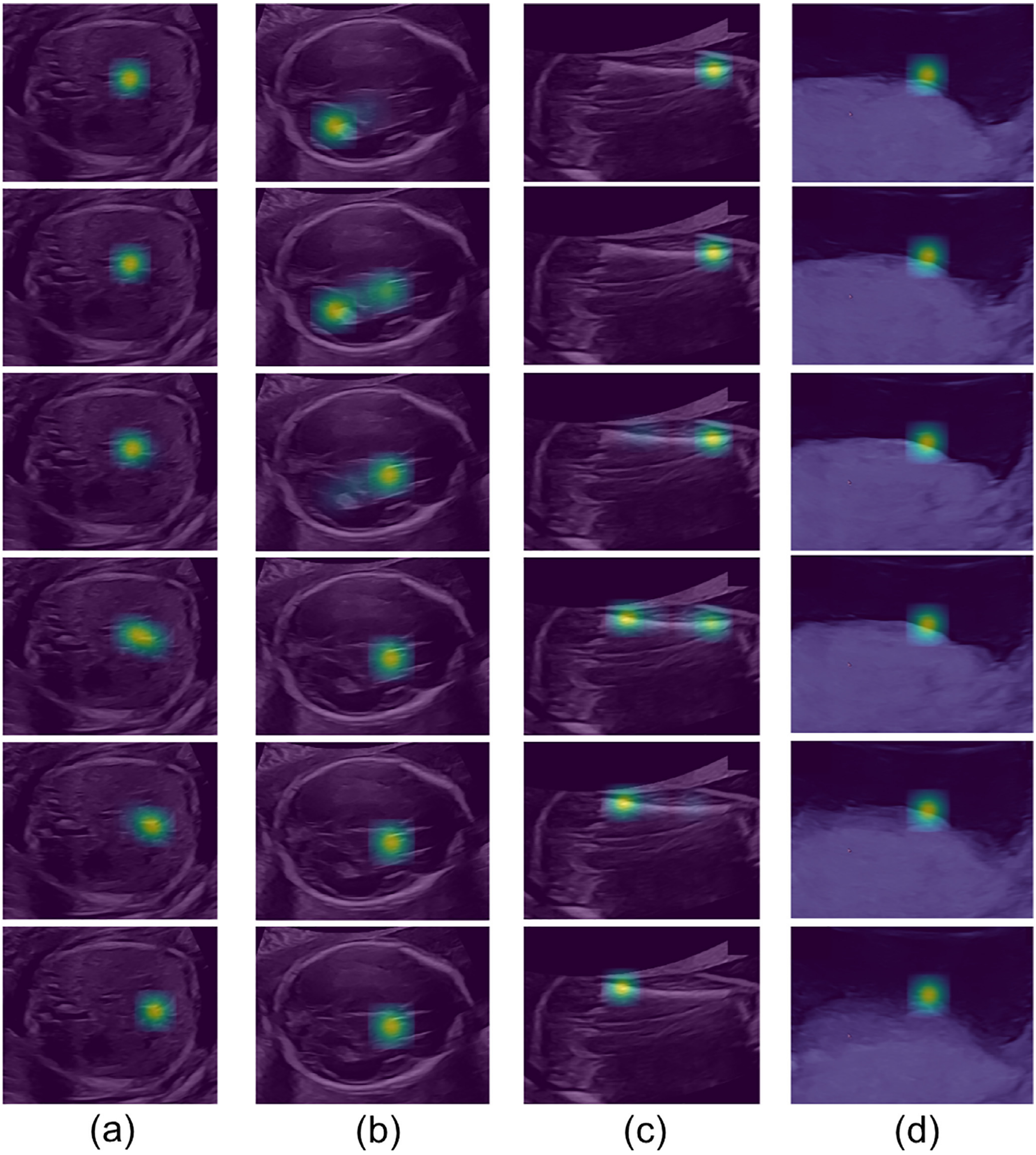


Fig. 3. Sonographer visual attention maps on 6 consecutive frames of (a) Standard Abdomen snippet (b) Standard Head snippet (c) Standard Femur snippet and (d) non-standard snippet.

where $\mathbf{X}^t$ is a frame, H, W represent the height and width of input frames, and T the number of frames in a sample; the notation $|a, b|$ indicates integer intervals between a and b, with a and b included. For each $\mathbf{X}^t$, a CNN feature extractor is used to extract spatial feature representations $\boldsymbol{\phi}^t$, which are subsequently fed into a Temporal Attention Module (TAM), a recurrent module that produces Dynamic Attention Maps (DAMs) $\mathbf{M} = [\mathbf{M}^1; \ldots; \mathbf{M}^T] \in [0, 1]^{H \times W \times T}$ for each input video frame, where H, W represent the height and width of the predicted attention maps. The generated attention maps are then fed into the Video Classification Module (VCM), which is also recurrent, that predicts frame-wise class label $k \in [|1, K|]$ for each frame $\mathbf{X}^t$, $t \in [|1, T|]$, where $K = 7$ (bg Ab, std ACP, bg Head, std HCP, bg Femur, std FLP, and an additional "other" class). The labels are one-hot encoded so that for a class k, the corresponding target is $\mathbf{y} = (y_i)_1^K$ with $y_k = 1$ and $\forall i : i \neq k, y_i = 0$. In any trained module compared below, TAM and VCM use the same

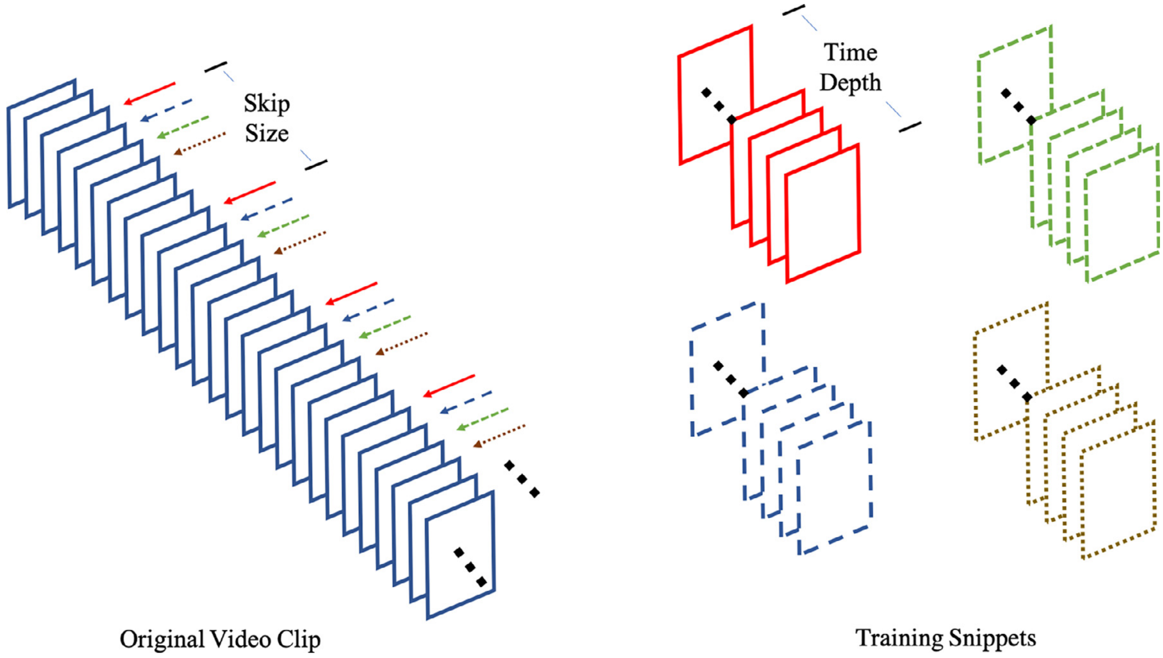


Fig. 4. Illustration showing sampling video snippets with defined Time Depth and Skip Size from the original video clip. Different colors and line styles indicate 4 different training samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

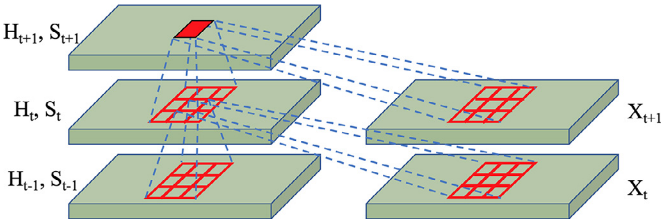


Fig. 5. Schematic of convolution operation for input-to-state and state-to-state transitions.

recurrent module, both in terms of bi-direction/uni-directional, and CLSTM/CGRU. In the following sections, schematics of the TAM and VCM are presented in the case of bi-directional RNNs.

3.2.3. Spatial feature extractor

The Spatial Feature Extractor is derived from VGG-16 (Simonyan and Zisserman, 2014) with a quarter of the number of convolutional kernels in each layer, as can be seen in Fig. 7. The Spatial Feature Extractor consists of three convolutional blocks, the first two of which consist of two convolutional layers, and the third block consists of three convolutional layers. All convolutional layers use 3 × 3 convolutional kernels; the number of convolutional kernels used in the three blocks are 16, 32, and 64, respectively.

3.2.4. Temporal attention module

A detailed architecture of the Temporal Attention Module (TAM) is given in Fig. 8, where yellow cubes represent tensors of feature maps from convolution operations, orange stripes represent activation function (ReLUs), and blue cubes represent tensors of hidden state of recurrent neural networks (Convolutional GRU or Convolutional LSTM). Since the recurrent modules are bi-directional, feature maps extracted by CNN ϕ^t, $t \in [|1, T|]$ from sample video clips are fed into TAM in both positive (bottom) and reverse order (top), as demonstrated in Fig. 8. Extracted spatial features ϕ^t (bottom-left and top-right) are passed through several convolutional layers to generate a Static Attention Map (SAM) $\tilde{\mathbf{M}}^t$ (bottom-middle and top-middle), which is then processed through a residual operation to generate $\tilde{\phi}^t$:

$$\tilde{\phi}^t = \phi^t \odot \tilde{\mathbf{M}}^t + \phi^t. \qquad (1)$$

$\tilde{\phi}^t$ from both the positive order ($\tilde{\phi}^{t+}$) and reverse order ($\tilde{\phi}^{t-}$) are each fed into a convolutional recurrent neural network to generate $\tilde{h}^{t+}$ and $\tilde{h}^{t-}$, respectively. These two hidden-states are concatenated to generate $\tilde{h}^t$, which, after further convolution and sigmoid activation, generates Dynamic Attention Map $\mathbf{M}^t$. Loss function between ground truth visual attention maps $\mathbf{A}$ and $\mathbf{M}$ are defined in Section 3.3.

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

3.2.5. Video classification module

A detailed architecture of the Video Classification Module (VCM) can be seen in Fig. 9. Feature maps ϕ^t from the t^{th} frame are fed into a bi-directional RNN to generate $\hat{h}^{t+}$ and $\hat{h}^{t-}$ from the positive and reverse order, which are subsequently concatenated to form $\hat{h}^t$. After three convolution layers, the resultant feature maps are merged with $\mathbf{M}^t$ through element-wise multiplication to produce $\hat{\phi}^t$. Class prediction is performed on each $\hat{\phi}^t$ through three convolutional layers, two adaptation layers, and a global average pooling layer before producing video-wise class prediction $\hat{y}$. The loss function for classification is discussed in Section 3.3.

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

3.3. Loss functions

3.3.1. Classification loss

For an input video snippet $\mathbf{X} = [\mathbf{X}^1; \ldots; \mathbf{X}^T] \in [|0, 255|]^{h\times w\times T}$, the first loss function is a **classification loss** L_c between class-prediction $\hat{\mathbf{y}} = \{\hat{\mathbf{y}}^1, \ldots, \hat{\mathbf{y}}^T\} \in [0, 1]^{C\times T}$ and ground truth class label $\mathbf{y} = [\mathbf{y}^1; \ldots; \mathbf{y}^T] \in \{0, 1\}^{C\times T}$, with C being the number of classes. To

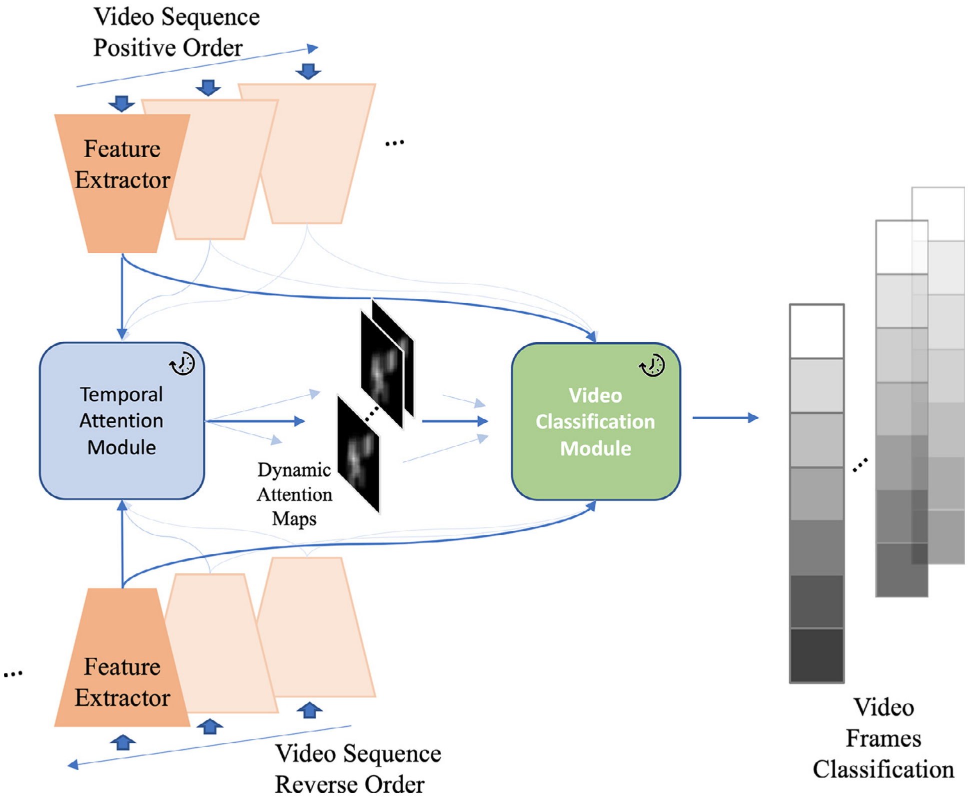


Fig. 6. Architecture of Temporal SonoEyeNet, consisting of a Feature Extractor, Temporal Attention Module (TAM), and a Video Classification Module (VCM). The clock symbol in the figure indicates it is a recurrent module, as temporal information is encoded.

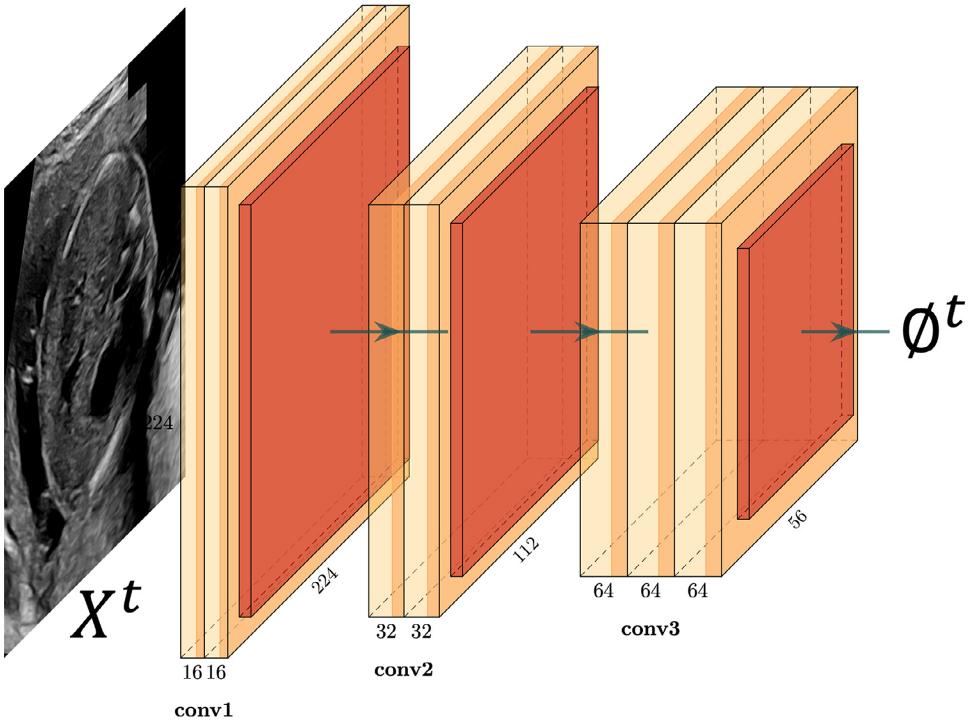


Fig. 7. Schematic of the Spatial Feature Extractor used in TSEN. $\mathbf{X}^t$ represents the t^{th} frame in an input US video clip; ϕ^t represents a tensor of extracted features from $\mathbf{X}^t$.

tackle class imbalance, the Focal Loss (Lin et al., 2017), a variant of cross-entropy loss, is used. Focal Loss allows hard (less well-classified) samples to contribute more to total loss, and down weights the contribution from easy (well-classified) samples by adding a modulating factor $(1 - \hat{y}_p)^{\xi}$ to cross-entropy loss:

$$FL(p_t) = -(1 - \hat{\mathbf{y}}^t_p)^{\xi} \log(\hat{\mathbf{y}}^t_p). \tag{2}$$

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

$$L_c = -\sum_{t=1}^{T} (1 - \hat{\mathbf{y}}^t_p)^{\xi} \log(\hat{\mathbf{y}}^t_p). \tag{3}$$

3.3.2. Saliency loss

The Kullback-Leibler Divergence (KLD) loss is used for Saliency Loss L_s. The KLD between predicted dynamic visual attention maps $\mathbf{M} = \{\mathbf{M}^1, \dots, \mathbf{M}^T\} \in [0, 1]^{h \times w \times T}$ and a ground truth sonographer visual attention map $\mathbf{A} = \{\mathbf{A}^1, \dots, \mathbf{A}^T\} \in [0, 1]^{h \times w \times T}$ can be written as:

$$L_s = D_{KL}(\mathbf{A}||\mathbf{M}) = -\sum_{t=1}^{T}\sum_{i=1}^{h}\sum_{j=1}^{w} \mathbf{A}^t_{i,j} \log \frac{\mathbf{M}^t_{i,j}}{\mathbf{A}^t_{i,j}}. \tag{4}$$

3.3.3. Temporal regularisation loss

Soft Dynamic Time Warping (sDTW) (Cuturi and Blondel, 2017) is used as the temporal regularisation Loss L_t. sDTW is a differentiable function, derived from Dynamic Time Warping (DTW) (Sakoe et al., 1990), that can be used as a loss function to enforce alignment of two time series. **Dynamic Time Warping (DTW)** measures the discrepancy between two time series by computing the best possible alignment between the two time series x and y with lengths n and m. It first computes the $n \times m$ pairwise cost matrix $\mathbf{\Delta}(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{n \times m}$ between points (Fig. 10(A)), and then finds an Alignment Matrix Ψ that defines a path on a $n \times m$ matrix that connects the upper-left (1, 1) matrix entry to the lower-right (n, m) one using only down $\downarrow$, right $\rightarrow$ and down-right $\searrow$ moves. The DTW score is defined as the minimised sum of cost; it is found by solving a dynamic program (DP) problem using Bellman's recursion (Bellman, 1952) with a quadratic (nm) cost. A soft-minimum function was used to guarantee the differentiability of the score so that it can be used as a loss function.

In the context of visual attention map prediction, we want to minimise the sDTW score between predicted dynamic visual attention maps $\mathbf{M} = \{\mathbf{M}^1, \dots, \mathbf{M}^T\} \in [0, 1]^{h \times w \times T}$ and ground truth sonographer visual attention maps $\mathbf{A} = \{\mathbf{A}^1, \dots, \mathbf{A}^T\} \in [0, 1]^{h \times w \times T}$. Let $\mathbf{\Psi}_{T,T} \subset \{0, 1\}^{T \times T}$ be a set of *binary alignment matrices* describing the path connecting top-left to bottom-right of the *cost matrix* $\mathbf{\Delta}$ using only right, down, or right-down connections, as presented in Fig. 10(B). The sDTW score is defined by:

$$\mathbf{dtw}_{\gamma}(\mathbf{M}, \mathbf{A}) = \min{}^{\gamma}\{\langle \Psi, \mathbf{\Delta}(\mathbf{M}, \mathbf{A})\rangle, \Psi \in \mathbf{\Psi}_{T,T}\}, \tag{5}$$

$$\min{}^{\gamma}\{a_1, \dots, a_n\} = -\gamma \log \sum_{i=1}^{n} e^{\frac{-a_i}{\gamma}}, \tag{6}$$

where $\langle \cdot,\cdot \rangle$ represents the inner product operator, γ the smoothing parameter, and $a_1, \dots, a_n$ represent elements in a vector. To find the optimal alignment matrix Ψ^* that minimises DTW(**M, A**), an *intermediary alignment cost matrix* $\mathbf{R}$ (Fig. 11(B)) is constructed using Bellman's Recursion (Fig. 11(A)) to calculate the minimum summed cost (*i.e.* sDTW) achieved by Ψ^*. Bellman's recursion to construct $\mathbf{R}$ is calculated through the following equations:

$$r_{i,j} = \delta_{i,j} + \min{}^{\gamma}(r_{i-1,j}, r_{i,j-1}, r_{i-1,j-1}), \tag{7}$$

$$\delta_{i,j} = MSE(\mathbf{M}^i, \mathbf{A}^j), \tag{8}$$

where $r_{i,j}$ represents an element in $\mathbf{R}$ and $\delta_{i,j}$ an element in $\mathbf{\Delta}$; *MSE* indicates the *Mean Squared Error*. When $\mathbf{R}$ is complete through Bellman's Recursion, the final sDTW score is:

$$\mathbf{dtw}_{\gamma}(\mathbf{M}, \mathbf{A}) = r_{T,T}. \tag{9}$$

The algorithm for calculating $\mathbf{dtw}_{\gamma}$(**M, A**) as well as the intermediate alignment cost matrix R^{γ} are summarised in Algorithm 1.

Algorithm 1 Forward recursion to compute $\mathbf{dtw}_{\gamma}$(**M, A**) and $\mathbf{R}^{\gamma}$.

```
Require:
    Predicted dynamic visual attention maps $\mathbf{M} \in [0, 1]^{h \times w \times T}$
    Ground truth sonographer visual attention maps $\mathbf{A} \in [0, 1]^{h \times w \times T}$
    Smoothing temperature term $\gamma > 0$
    Empty intermediary alignment cost matrix $\mathbf{R} \in \mathbb{R}^{T \times T}$
for $j = 1, \dots, T$ do
    for $i = 1, \dots, T$ do
        $\delta_{i,j} = MSE(\mathbf{M}^i, \mathbf{A}^j)$
    end for
end for
$r_{0,0} = 0;\ r_{0,i} = r_{j,0} = \infty;\ i \in [|1, T|], j \in [|1, T|]$    ▷ Initialisation
for $j = 1, \dots, T$ do
    for $i = 1, \dots, T$ do
        $r_{i,j} = \delta_{i,j} + \min^{\gamma}(r_{i-1,j}, r_{i,j-1}, r_{i-1,j-1})$
    end for
end for
return $(r_{T,T}, \mathbf{R})$
```

3.3.4. Overall loss

The network is trained with all three losses: the **classification Loss** L_c, , the **saliency loss** L_s, and the **temporal regularisation loss** L_t that encourages alignment of dynamic attention maps over time. The total loss L is represented as:

$$L = \alpha L_c + \beta L_s + \lambda L_t \tag{10}$$

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

3.4. Training details

All images were resized to 240 × 240 pixels and randomly cropped into size 224 × 224 pixels on-the-fly for data augmentation during training. In addition, all image frame intensities were normalised to zero-mean and unit variance. All TSEN variants were initialised using a zero-mean Gaussian distribution with standard deviation of 0.01. They were trained using adaptive moment estimation (Adam) (Kingma and Ba, 2014) with an initial learning rate of 2×10^{-4} for 100 epochs and weight decay of 5×10^{-4}. All video snippets $\mathbf{X}$ were sampled from video clips with a skip size of 10 and time depth T of 10 with a mini-batch size of 16. TSEN models were trained with three losses: Classification loss L_c, Saliency loss L_s, and temporal regularisation loss L_t. α and β for L_c and L_s were set to 0.5 to give the two tasks equal weighting. λ for L_t was set at 0.01 after searching through 5 logarithmically spaced weights between 10^{-4} and 1. The best performing model used focal loss for L_c and the weight ξ was set at a value of 2 to suppress gradients of easy negative samples; other variants used cross-entropy loss for L_c. The dataset was split at scan-level into 5 folds for cross-validation.

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

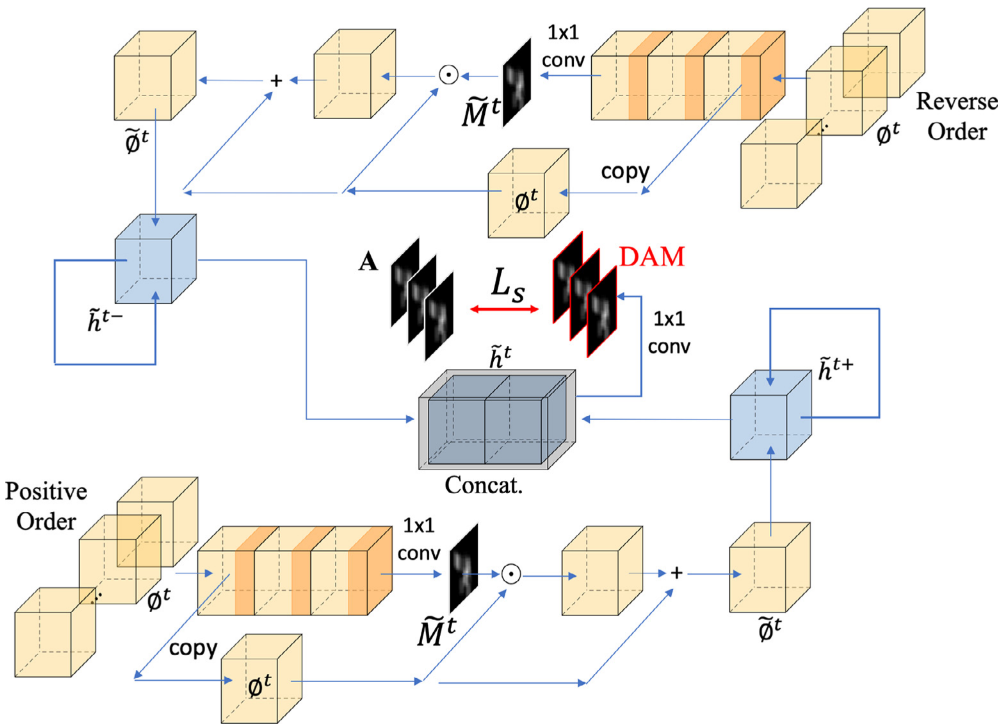


Fig. 8. Schematic of the Temporal Attention Module (TAM).

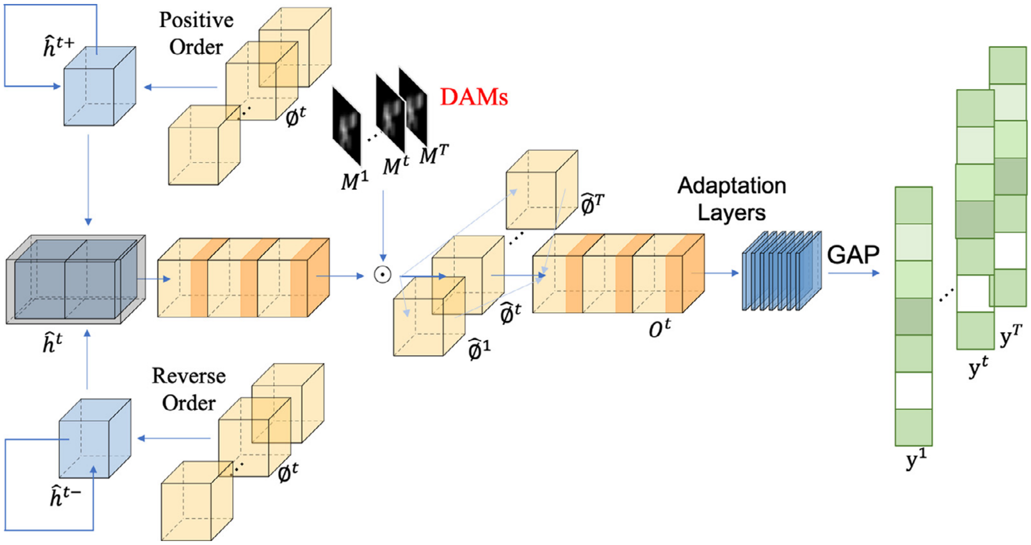


Fig. 9. Schematic of the Video Classification Module (VCM).

SonoNet models (Baumgartner et al., 2017) and MSEN (Cai et al., 2018b) were trained as baseline comparisons; SonoNet was randomly initialised and trained for 100 epochs using Adam optimiser with a learning rate of 10^{-4} and weight decay of 5×10^{-4}; MSEN was also trained from randomly-initialised weights for 100 epochs. Video frames were treated as independent images and sampled at the inverse of the frequencies of their corresponding frame-wise label.

3.5. Performance metrics

3.5.1. Classification metrics

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

3.5.2. Static saliency metrics

Five static saliency metrics discussed in Bylinskii et al. (2018) were used to quantify the similarity between ground truth and predicted visual attention maps: Area Under ROC Curve (*AUC*), Normalized Scanpath Saliency (*NSS*), Information Gain (*IG*), Similarity (*Sim*), and Pearson's Correlation Coefficient (*CC*). In addition, the Kullback-Leibler divergence (*KLD*) is also reported. All metrics are reported on a per-anatomy basis as well as on a macro-average basis.

Table 1
Static saliency scores of different models, including area under curve (AUC), Pearson cross-correlation (CC), similarity (SIM), information gain (IG), normalized saliency scan path (NSS), and Kullback-Leibler Divergence (KLD) (Bylinskii et al., 2018).

	AUC [%]	CC[%]	SIM [%]	IG	NSS	KLD ↓
biCLSTM+sDTW	**64.8** ± 0.6	**69.9** ± 1.1	**57.7** ± 2.6	**1.50** ± 0.02	**2.76** ± 0.05	**1.06** ± 0.08
biCLSTM	62.3 ± 0.6	41.2 ± 2.2	36.8 ± 3.1	1.27 ± 0.04	2.16 ± 0.01	1.79 ± 0.12
biCGRU+sDTW	54.2 ± 0.4	51.9 ± 2.8	36.7 ± 2.5	0.71 ± 0.02	2.18 ± 0.05	1.50 ± 0.04
biCGRU	47.5 ± 0.5	29.6 ± 2.4	19.8 ± 3.0	0.12 ± 0.04	1.32 ± 0.07	2.08 ± 0.05
uniCLSTM	61.4 ± 0.2	41.7 ± 1.0	33.0 ± 2.7	0.18 ± 0.03	1.53 ± 0.05	1.62 ± 0.07
uniCGRU	43.9 ± 0.8	27.5 ± 2.0	20.3 ± 2.9	0.10 ± 0.03	1.30 ± 0.10	2.21 ± 0.07
MSEN (Cai et al., 2018b)	40.1 ± 0.3	27.4 ± 1.8	18.4 ± 2.4	0.13 ± 0.03	1.45 ± 0.08	1.82 ± 0.05

Table 2
Scanpath similarity scores of different models, including Vector Similarity (VecSim), Length Similarity (LenSim), Direction Similarity (DirSim), and Position Similarity (PosSim) (Dewhurst et al., 2012).

	VecSim	DirSim	LenSim	PosSim
biCLSTM+sDTW	**97.6** ± 0.4	**75.9** ± 0.2	**97.1** ± 0.3	**95.9** ± 0.6
biCLSTM	96.1 ± 0.8	69.3 ± 0.3	94.5 ± 0.3	88.8 ± 0.8
biCGRU+sDTW	92.7 ± 0.7	70.6 ± 0.8	87.9 ± 0.3	79.5 ± 0.2
biCGRU	95.1 ± 0.6	68.5 ± 0.7	93.2 ± 0.4	84.5 ± 0.5
uniCLSTM	96.1 ± 0.3	69.5 ± 0.2	94.8 ± 0.6	87.8 ± 0.7
uniCGRU	92.5 ± 0.2	66.1 ± 0.6	89.6 ± 0.4	83.0 ± 0.5
MSEN (Cai et al., 2018b)	93.0 ± 0.3	69.1 ± 0.5	92.9 ± 0.5	86.7 ± 0.4

Table 3
Classification results of different models.

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

Table 4
F1 scores of different baseline models by anatomy.

	MSEN	SonoNet-64	SonoNet-32	SonoNet-16
bg Ab	87.2 ± 3.5	87.3 ± 1.6	**87.6** ± 2.1	85.3 ± 4.3
std ACP	**68.3** ± 2.6	40.2 ± 8.7	39.8 ± 9.2	40.4 ± 8.0
bg Head	**92.1** ± 3.2	91.0 ± 2.1	91.4 ± 1.9	91.4 ± 1.7
std HCP	**68.1** ± 4.1	49.7 ± 5.6	45.0 ± 13.7	44.5 ± 14.0
bg Femur	76.7 ± 2.4	**77.8** ± 4.7	76.8 ± 6.0	75.9 ± 6.0
std FLP	**60.0** ± 2.6	43.5 ± 12.3	44.8 ± 12.1	44.5 ± 10.5
Others	**87.1** ± 2.7	72.8 ± 3.2	74.0 ± 4.6	73.4 ± 4.5

3.5.3. *Scanpath metrics*

A *scanpath* is defined as a set of fixation points ((*x*, *y*)-coordinate of the maxima on a visual attention map) on consecutive video frames and the transitions between each pair of fixation points, as can be seen in Fig. 12(A) where red dots represent fixation points and the dotted arrows represent transitions; different color-coding for visual attention maps indicate different time points. Following *MultiMatch* (Dewhurst et al., 2012; Jarodzka et al., 2010), a set of metrics is used to measure scanpath similarities, four different metrics are calculated for two scanpaths: Vector Similarity (*VecSim*), Length Similarity (*LenSim*), Direction Similarity (*DirSim*), and Position Similarity (*PosSim*). The MultiMatch metrics represent the scanpath as a set of vectors on a 2-D plane, as can be seen in Fig Fig. 12(B); each vector originates from the fixation point at time point t and points to the fixation point at time point $t+1$.

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

Four metrics are calculated, as illustrated in Fig. 12(C):

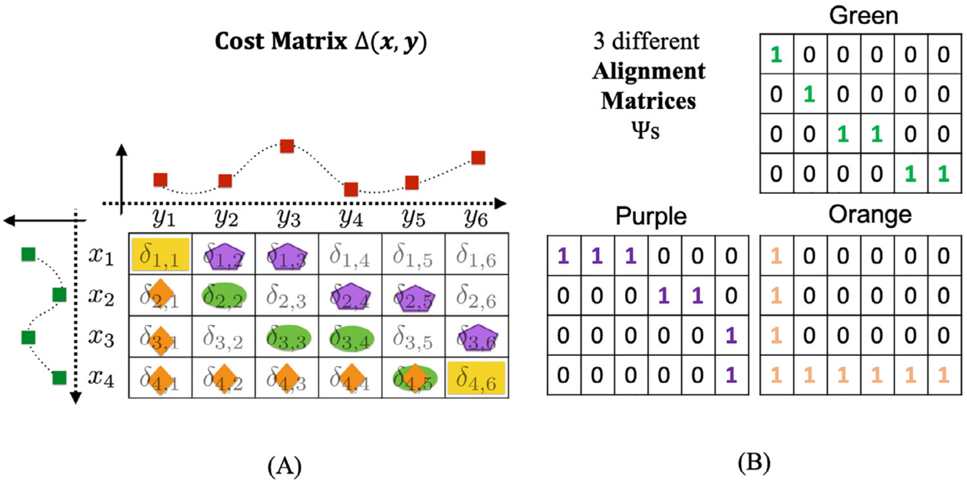
$$VecSim = 1 - \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\|\mathbf{v_1^t} - \mathbf{v_2^t}\|}{2 \times d} \tag{11}$$

$$LenSim = 1 - \frac{1}{T-1}\sum_{t=1}^{T-1}\|\mathbf{v_1^t}\| - \|\mathbf{v_2^t}\| \tag{12}$$

$$DirSim = 1 - \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{1}{\pi}\cos^{-1}\frac{\mathbf{v_1^t} \cdot \mathbf{v_2^t}}{\|\mathbf{v_1^t}\|\|\mathbf{v_2^t}\|} \tag{13}$$

$$PosSim = 1 - \frac{1}{T}\sum_{t=1}^{T}\frac{\|\mathbf{p}_1^t - \mathbf{p}_2^t\|}{2 \times d} \tag{14}$$

where d represents the diagonal size of the visual attention maps. It is worth noting that Duration Similarity (*DurSim*) in the original paper is not calculated in this study, because the input video frames are sampled at 30 *Hz* and the duration of each fixation is thus approximately 0.033 seconds.


Cost Matrix $\Delta(x, y)$
y_1 y_2 y_3 y_4 y_5 y_6
x_1 x_2 x_3 x_4
3 different Alignment Matrices Ψs
Green
Purple
Orange
(A)
(B)

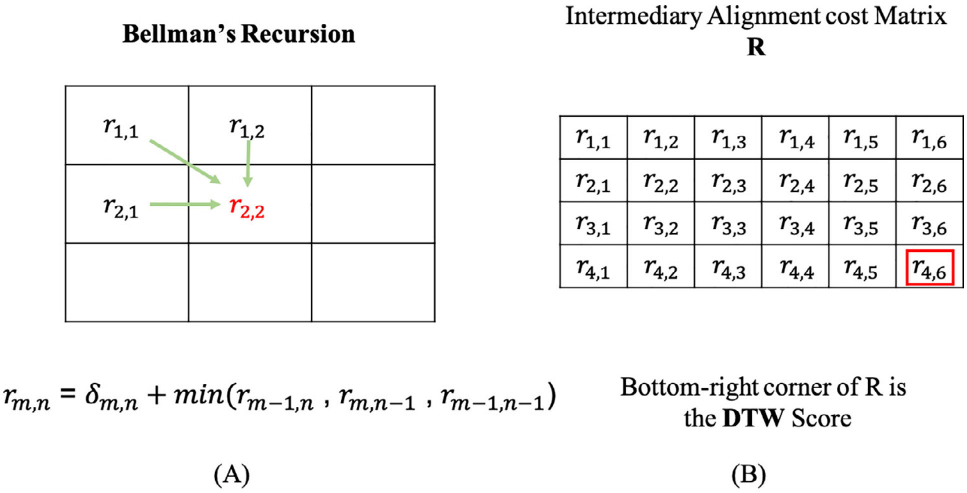
$$\begin{array}{c|cccccc} & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \\ \hline x_1 & \delta_{1,1} & \delta_{1,2} & \delta_{1,3} & \delta_{1,4} & \delta_{1,5} & \delta_{1,6} \\ x_2 & \delta_{2,1} & \delta_{2,2} & \delta_{2,3} & \delta_{2,4} & \delta_{2,5} & \delta_{2,6} \\ x_3 & \delta_{3,1} & \delta_{3,2} & \delta_{3,3} & \delta_{3,4} & \delta_{3,5} & \delta_{3,6} \\ x_4 & \delta_{4,1} & \delta_{4,2} & \delta_{4,3} & \delta_{4,4} & \delta_{4,5} & \delta_{4,6} \end{array}$$

Green
$$\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{matrix}$$

Purple
$$\begin{matrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

Orange
$$\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{matrix}$$

Fig. 10. Schematic showing (A) A cost matrix Δ between time series **x, y** with orange, green and purple as color codes for three different possible connections between top-left and bottom right elements. (B) Binary Alignment matrices with corresponding color codes. The figure is adapted from (Cuturi and Blondel, 2017). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)


Bellman's Recursion
$r_{1,1}$ $r_{1,2}$
$r_{2,1}$ $r_{2,2}$
Intermediary Alignment cost Matrix R
Bottom-right corner of R is the DTW Score
(A)
(B)

$$\begin{matrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & r_{1,5} & r_{1,6} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & r_{2,5} & r_{2,6} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & r_{3,5} & r_{3,6} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & r_{4,5} & \boxed{r_{4,6}} \end{matrix}$$

$$r_{m,n} = \delta_{m,n} + min(r_{m-1,n} , r_{m,n-1} , r_{m-1,n-1})$$

Fig. 11. Schematic showing (A) Bellman's Recursion. (B) A complete Intermediary alignment cost matrix R.

4. Results

4.1. Temporal visual attention modelling

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

4.1.1. Qualitative assessment

Predicted visual attention maps generated by different TSEN variants on three video snippets in the test set are shown in Fig. 13 for the fetal abdomen, Fig. 14 for the fetal head, and Fig. 15 for the fetal femur. In general, by visual inspection, the best performing model in all three anatomies is "biCLSTM+sDTW", demonstrating good synchronisation of saccadic transitions with the sonographer visual attention map ground truth. In Fig. 13, as the ACP gradually appears, the predicted visual attention moves from the middle of the view to the area between the stomach bubble and umbilical vein in the same fashion as the ground truth; in Fig. 14, the predicted visual attention follows the ground truth

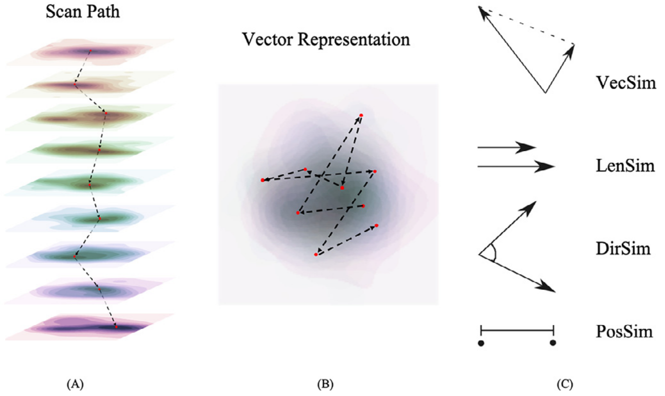


Fig. 12. Schematic outlining the MultiMatch scanpath similarity metrics. (A) Cartoon representation of a scanpath through 9 visual attention maps on consecutive US video frames. (B) Vector representation of the scanpath on 2-D plane (C) Cartoon of the 4 similarity metrics. The figure is adapted from (Dewhurst et al., 2012).

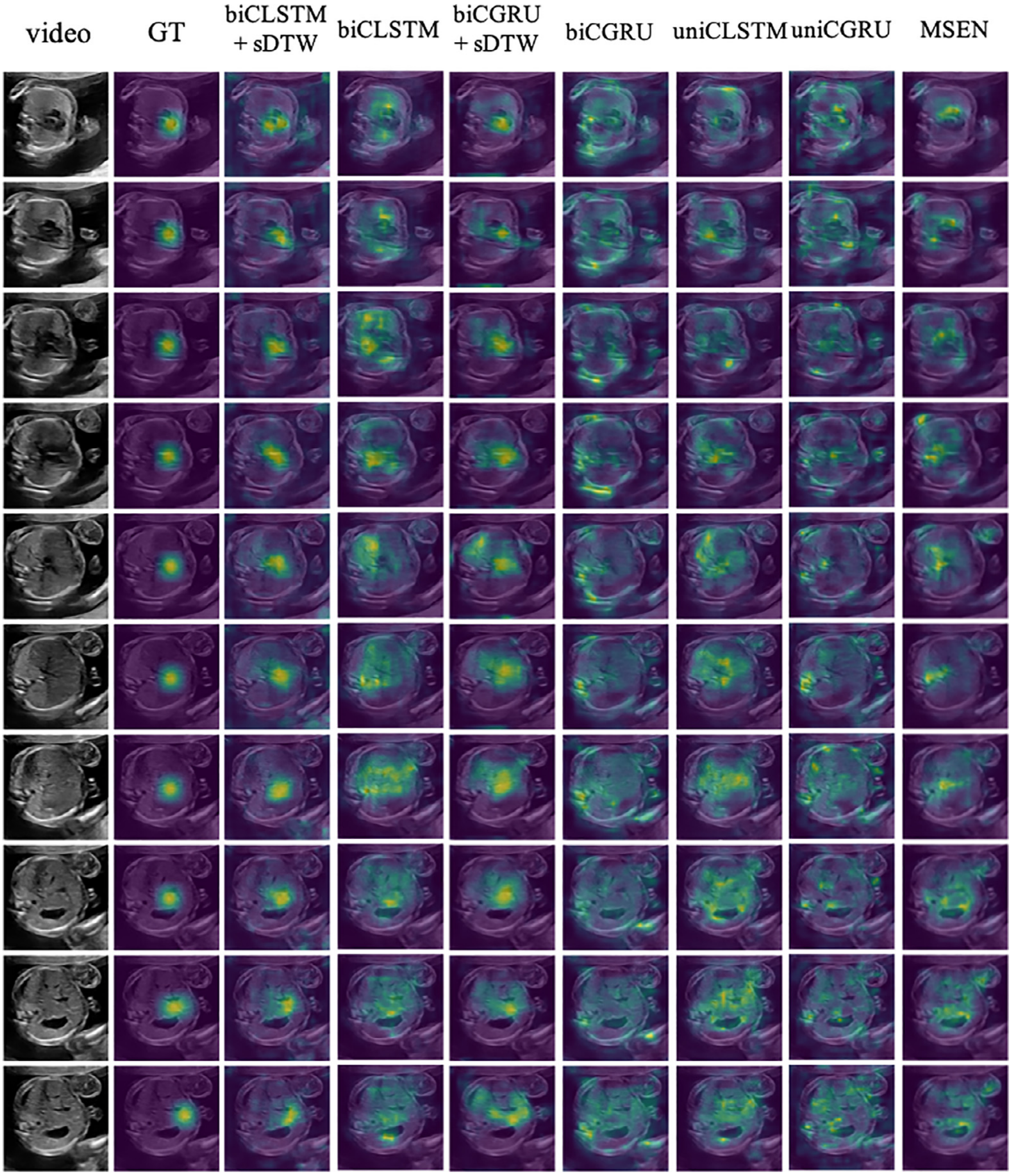


Fig. 13. Visual attention maps generated by different TSEN variants on an example of fetal abdomen clip. From left to right: US video frames, ground truth (sonographer's actual attention map), biCLSTM+sDTW, biCLSTM, biCGRU+sDTW, biCGRU, uniCLSTM, uniCGRU, MSEN.

by scanning along the centerline of the brain before focusing on the *cavum septum pellucidum*; in Fig. 15, the prediction replicates the scanning behavior along the femur bone, even with the transient appearance of other structures. The model "biCLSTM", trained without temporal regularisation, demonstrated less-focused attention and the fixations are not well-synchronised with the ground truth. Similar improvement can be seen by comparing the prediction of "biCGRU+sDTW" with "biCGRU".

In general, CLSTM models demonstrate better capacity to learn temporal visual attention transitions than CGRU models. Overall, the predicted visual attention of CGRU models are more "spread out" with no clear points of fixation compared to CLSTM models. In addition, it is observed that bi-directional models are able to predict visually higher-quality visual attention maps compared to uni-directional models.

4.1.2. Quantitative assessment

In order to quantitatively assess visual attention prediction performance of different TSEN variants, static saliency scores (Table 1) and scanpath similarity scores (Table 2) were computed on test set for each variant. Higher scores in all metrics with the exception of *KLD* indicate a higher performance; for KLD, the lower the score, the better a model performs.

Consistent with qualitative assessment findings, the "biCLSTM+sDTW" model outperforms other models in all **static saliency scores**, reaching mean scores across all classes of 64.8% for *AUC*, 69.9% for *CC*, 57.7% for *SIM*, 1.50 for *IG*, 2.76 for *NSS*, and 1.06 for *KLD*, as seen in Table 1. Using sDTW as a temporal regulariser significantly improves models for all all performance metrics; this improvement is more prominent for biCGRU models than for "biCLSTM". "biCLSTM" outperforms "uniCLSTM" in 3 out of 6 metrics (*SIM*, *IG*, and *NSS*), while "biCGRU" outperforms "uniCGRU" in 4 out of 6 metrics (*AUC*, *CC*, *IG* and *NSS*), though the improvement in *IG* and *NSS* are not significant. All TSEN models outperform the baseline "MSEN" for all metrics except for *IG* and *NSS*, where "MSEN" slightly outperforms "uniCGRU".

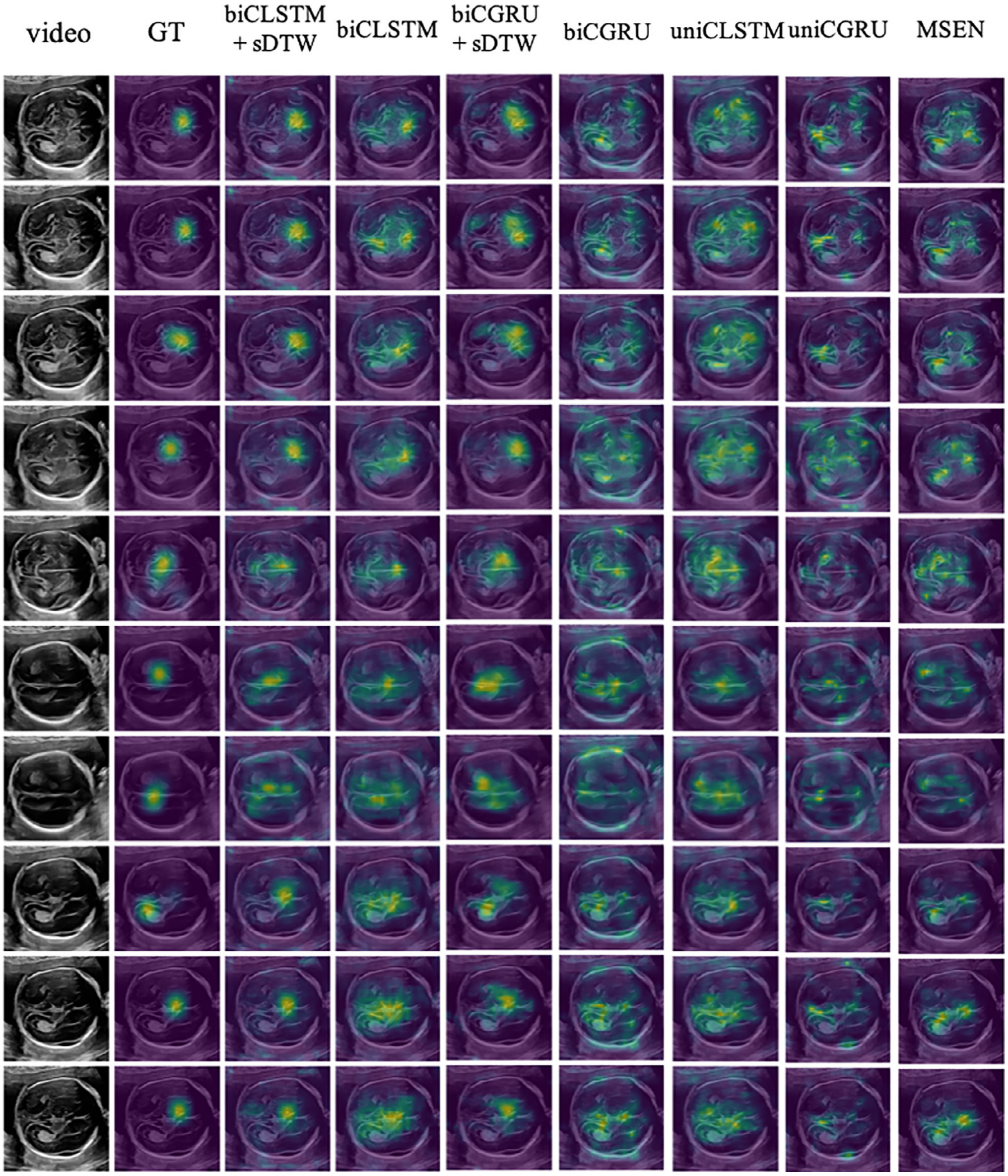


Fig. 14. Visual attention maps generated by different TSEN variants on an example of fetal head clip. From left to right: US video frames, ground truth (sonographer's actual attention map), biCLSTM+sDTW, biCLSTM, biCGRU+sDTW, biCGRU, uniCLSTM, uniCGRU, MSEN.

Similar to the results observed by using static saliency scores, the performance of "biCLSTM+sDTW" exceeds those of other variants in all of the **scanpath similarities scores**. Specifically, it reaches 97.7% for *VecSim*, 75.9% for *DirSim*, 97.1% for *LenSim*, and 95.9% for *PosSim*. However, sDTW loss did not significantly improve biCGRU models, as "biCGRU+sDTW" only exceeds "biCGRU" in *DirSim*. "biCLSTM" achieved higher scores in all metrics compared to "biCGRU", and "uniCLSTM" outperforms "uniCGRU" in all metrics, indicating that CLSTM models have higher capability to model saccadic transitions. Finally, there is no significant difference in performance between "biCLSTM" and "uniCLSTM" except in *PosSim*, while "biCGRU" outperforms "uniCGRU" in all metrics. "biCLSTM+sDTW" and "biCLSTM" exceeds the performance of the baseline, "MSEN", in all scanpath similarities metrics.

4.1.3. Anatomy-specific performances

In order to gain insight on each TSEN variant's visual attention prediction performance on standard and non-standard snippets, the static saliency scores and scanpath similarity scores are broken down according to snippet-level labels, as can be seen in Fig. 17 (static saliency scores) and Fig. 18 (scanpath similarity scores). For each type of anatomy along the $x-$axis, box plots of the performance metric are shown for 4 different models.

"biCLSTM+sDTW" (blue boxes) remains the best-performing TSEN model for all anatomies across all metrics, with the exception of the *IG* score for non-standard Abdomen (Fig. 17(C)), the *NSS* score for standard HCP (Fig. 17(E)), and *VecSim* score for standard HCP (Fig. 18(D)), where "biCLSTM" slightly outperforms.

It can also be observed that "biCLSTM+sDTW" generally performs better on standard snippets than on non-standard snippets for each biometry. Specifically, it performs better on standard abdomen snippets than on non-standard abdomen snippets on all static saliency scores with the exception of AUC; it also performs better on standard head snippets and standard femur snippets on 3 out of 6 static saliency scores comparing to their respective counter-parts. Standard head snippets achieve better scores

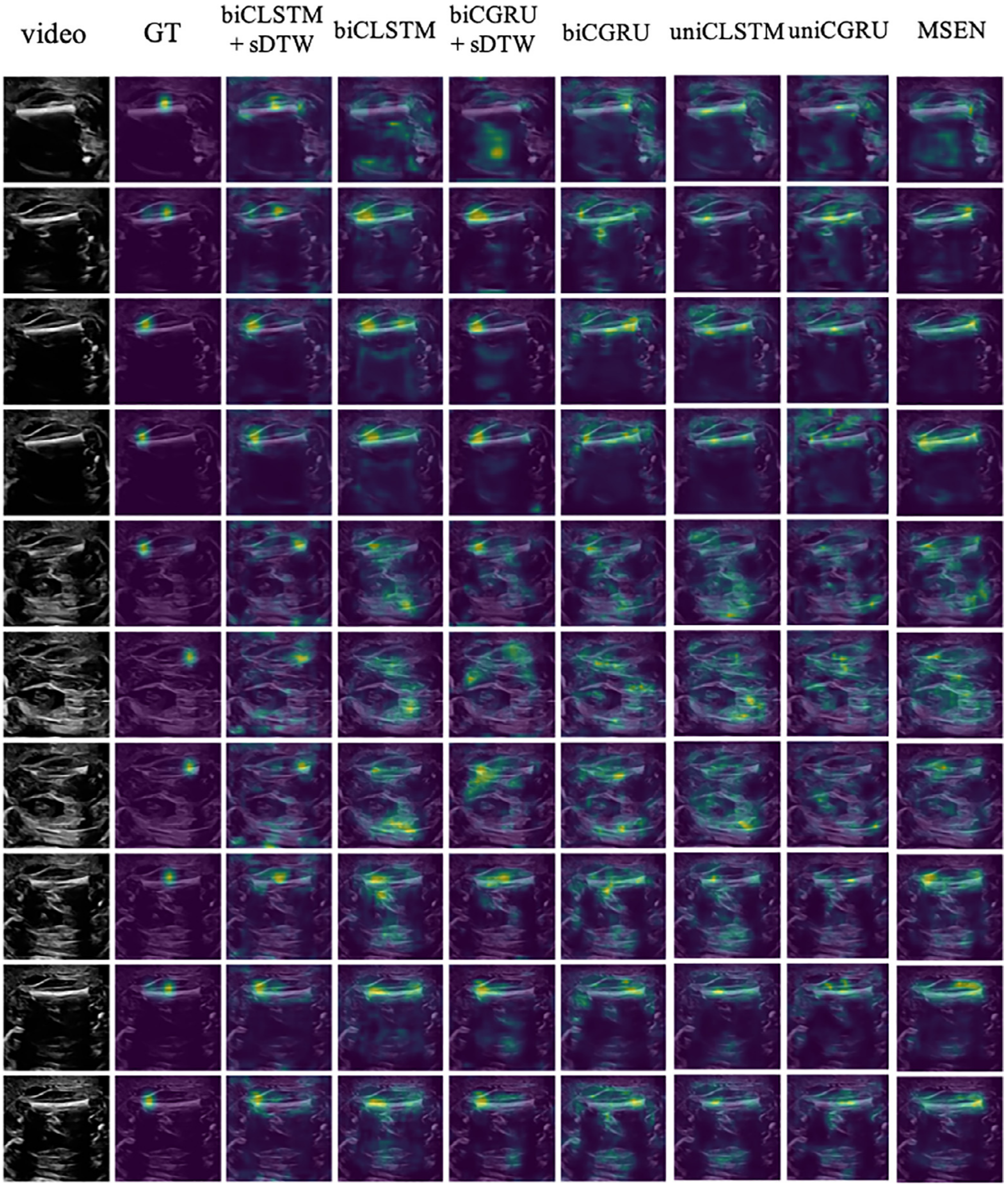


Fig. 15. Visual attention maps generated by different TSEN variants on an example of fetal femur clip. From left to right: US video frames, ground truth (sonographer's actual attention map), biCLSTM+sDTW, biCLSTM, biCGRU+sDTW, biCGRU, uniCLSTM, uniCGRU, MSEN.

in all scanpath similarity scores, while standard abdomen snippets achieve better scores in 3 out of all 4 scores. Fig. 16 summarizes comparison results based on static saliency scores and scanpath similarity scores of "biCLSTM+sDTW". Similar trends were observed for other model variants.

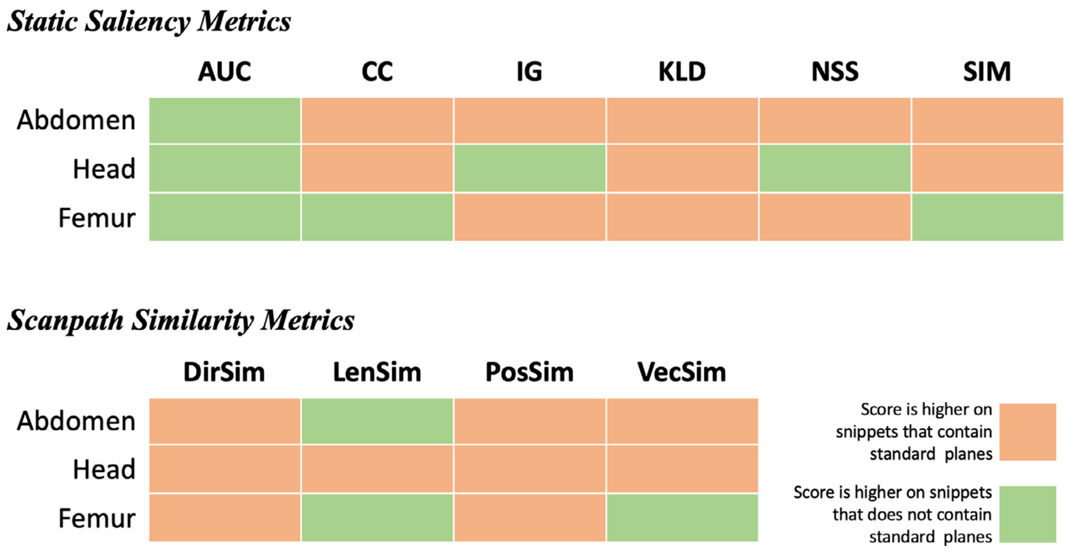
4.2. Frame classification performance

Frame-level classification results of all TSEN variants are presented in Table 3 and results of baseline models are presented in Table 4. Macro-averaged *Precision*, *Recall* and *F1 score* are reported for each variant and baseline model; performances per class are reported by the F1 score. TSEN variants are named by the specifications of the RNNs used in the VCM. Similar to the nomenclature used in the previous section, the model "biCLSTM" indicates that the model used bi-directional convolutional LSTM in the VCM which was trained using cross-entropy loss as L_c, and "biCLSTM+FL" indicates that that the choice of classification loss L_c was Focal Loss, instead of cross-entropy. As mentioned before, all variants' TAM share the same RNNs specifications with VCM; the TAM of all variants reported in this section are trained with Kullback-Leibler loss as L_s with additional temporal regularisation loss L_t using sDTW.

It can be observed that Focal Loss is an effective loss function for improving frame classification performance: "biCLSTM+FL" model achieves the highest macro-averaged precision, recall and F1 scores in all variants compared. Its performance also exceeds those of other variants in terms of F1 scores in all classes except for non-standard Head, on which "biCLSTM" achieves the highest score.

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

All TSEN models achieved higher scores in standard biometry planes comparing to baseline "MSEN", which achieved F1 scores



Static Saliency Metrics

	AUC	CC	IG	KLD	NSS	SIM
Abdomen	🟩	🟧	🟧	🟧	🟧	🟧
Head	🟩	🟧	🟩	🟧	🟩	🟧
Femur	🟩	🟩	🟧	🟧	🟧	🟩

Scanpath Similarity Metrics

	DirSim	LenSim	PosSim	VecSim
Abdomen	🟧	🟩	🟧	🟧
Head	🟧	🟧	🟧	🟧
Femur	🟧	🟩	🟧	🟩

Score is higher on snippets that contain standard planes	🟧
Score is higher on snippets that does not contain standard planes	🟩

Fig. 16. Static saliency scores and scanpath similarity scores comparison on snippets that contain standard biometry planes (orange blocks) *vs.* non-standard clips (green blocks) on all anatomy (abdomen, head and femur) based on the result of "biCLSTM+sDTW" model. In general, the model performs better on snippets that contain standard biometry planes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

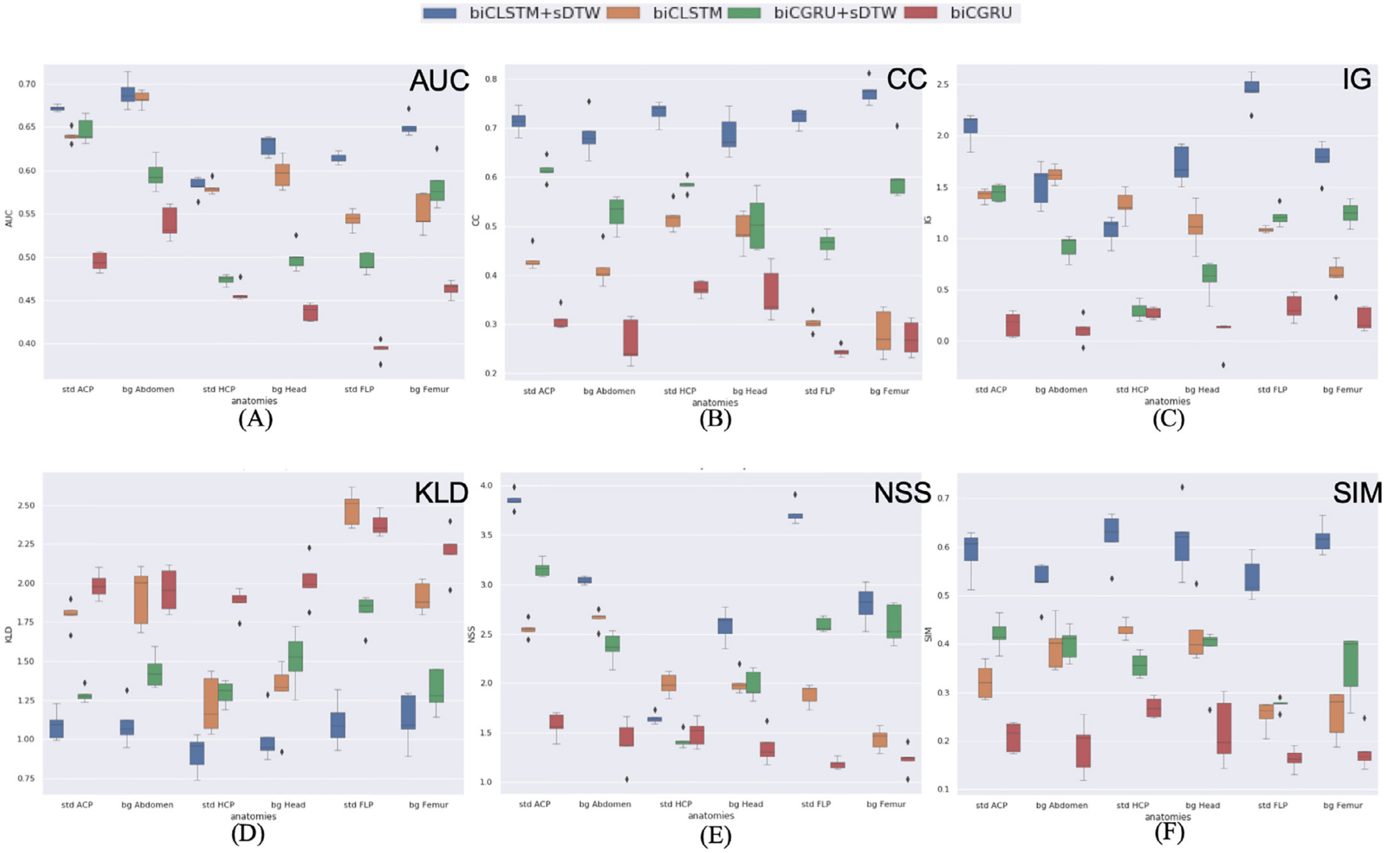


Fig. 17. Boxplots demonstrating the anatomy-specific static saliency scores (A-F) of four selected variants of TSEN models on standard and non-standard snippets. The four variants are, by order, "biCLSTM+sDTW" (blue), "biCLSTM" (orange), "biCGRU+sDTW" (green), and "biCGRU" (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of 68.3% for standard ACP, 68.1% for standard HCP and 60% for standard FLP. The best performing TSEN model, "biCLSTM+FL", improved F1 scores on these standard biometry planes to 83.7%, 89.9% and 81.1%, respectively. These results are also superior to the values achieved by variants of the SonoNet architecture trained from random initialisation for this frame-classification task with best F1 scores of 40.4%, 49.7% and 44.8% on corresponding biometry planes.

4.2.1. t-SNE Feature visualisation

A feature dimensionality reduction method t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008)

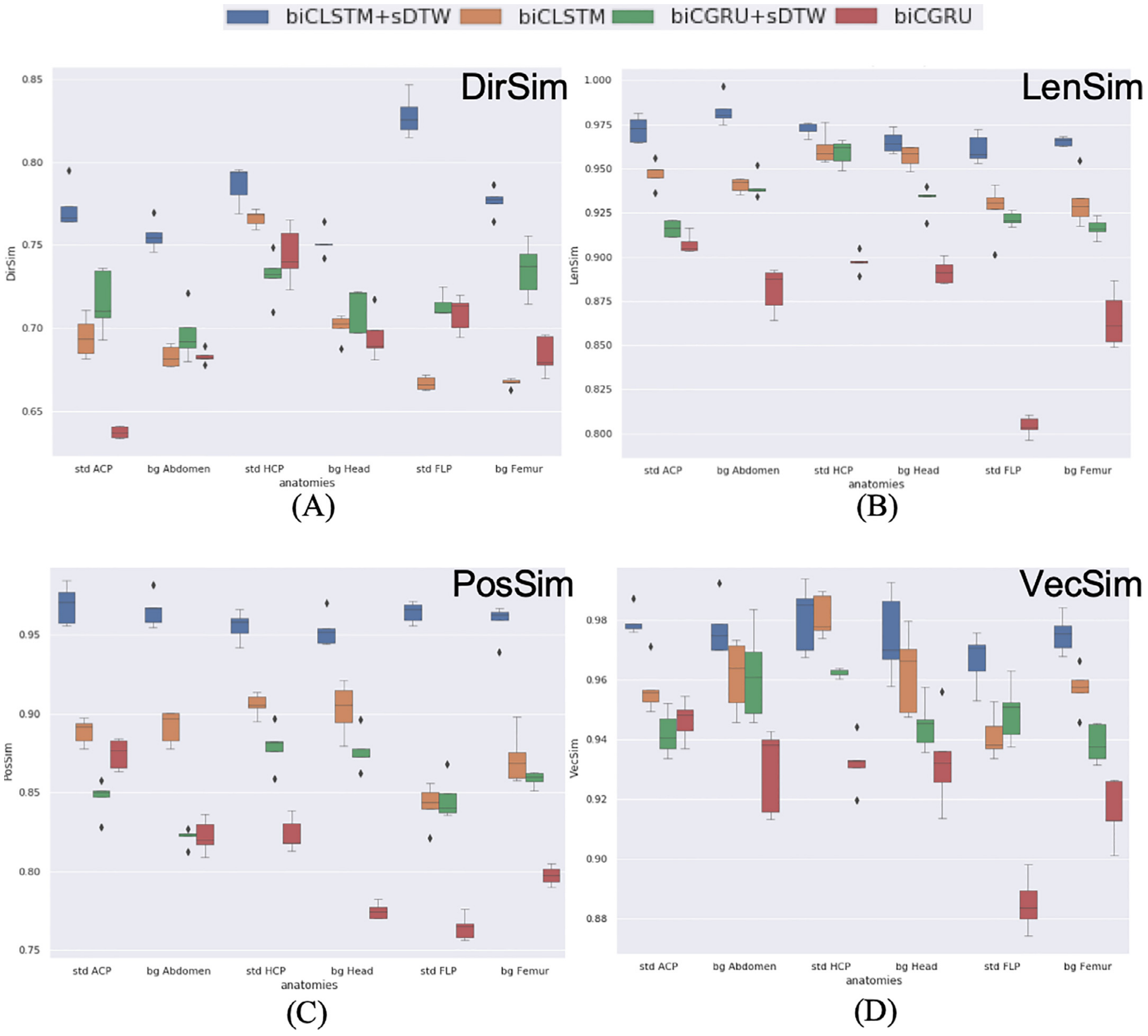


Fig. 18. Boxplots demonstrating the anatomy-specific scanpath similarity scores (A-D) of four selected variants of TSEN models on standard and non-standard snippets. The four variants are, by order, "biCLSTM+sDTW" (blue), "biCLSTM" (orange), "biCGRU+sDTW" (green), and "biCGRU" (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

was used to visualise the feature embedding generated at O^t in VCM and $\tilde{h}^t$ in the TAM of selected variants of TSEN models. Also, raw pixel values of the input video frames as well as the feature embedding of the last layer before adaptation layers in the classification branch and the last layer of the visual attention branch in MSEN are visualised for comparison. Compared to the t-SNE representation of the original images (Fig. 19(I)), feature embedding of O^t (the output features of the VCM as seen in Fig. 9) in the "biCLSTM+FL" model shows maximum separation between different classes. O^t of "biCLSTM" and "biCGRU" demonstrated lesser level of separation. However, overlaps still exist between the standard biometry planes and their corresponding non-standard planes due to anatomical similarities.

It is interesting to notice that $\tilde{h}^t$ (the last features of TAM as seen in Fig. 8) of "biCLSTM+FL" (Fig. 19(B)), though not trained for frame classification, demonstrated a good separation among different classes. For example, the embedding of original images (Fig. 19(I)) shows that the "Other" class is overlaps with the Abdomen ("std HCP" and "bg Abdomen"), while in Fig. 19(B) the "Other" class is clearly separated from Abdomen. Such observation can also be made for the $\tilde{h}^t$ of "biCLSTM" and "biCGRU", as well as "MSEN_att", indicating by learning to model human visual attention, the learnt feature embedding contain spatio-temporal information specific to different classes.

5. Discussion and conclusion

TSEN potentially would fit nicely into clinical sonography workflow as an automated algorithm for standard fetal biometry plane detection, after which biometric measurements (manual or automatic) and clinical decisions would be made. It could operate in

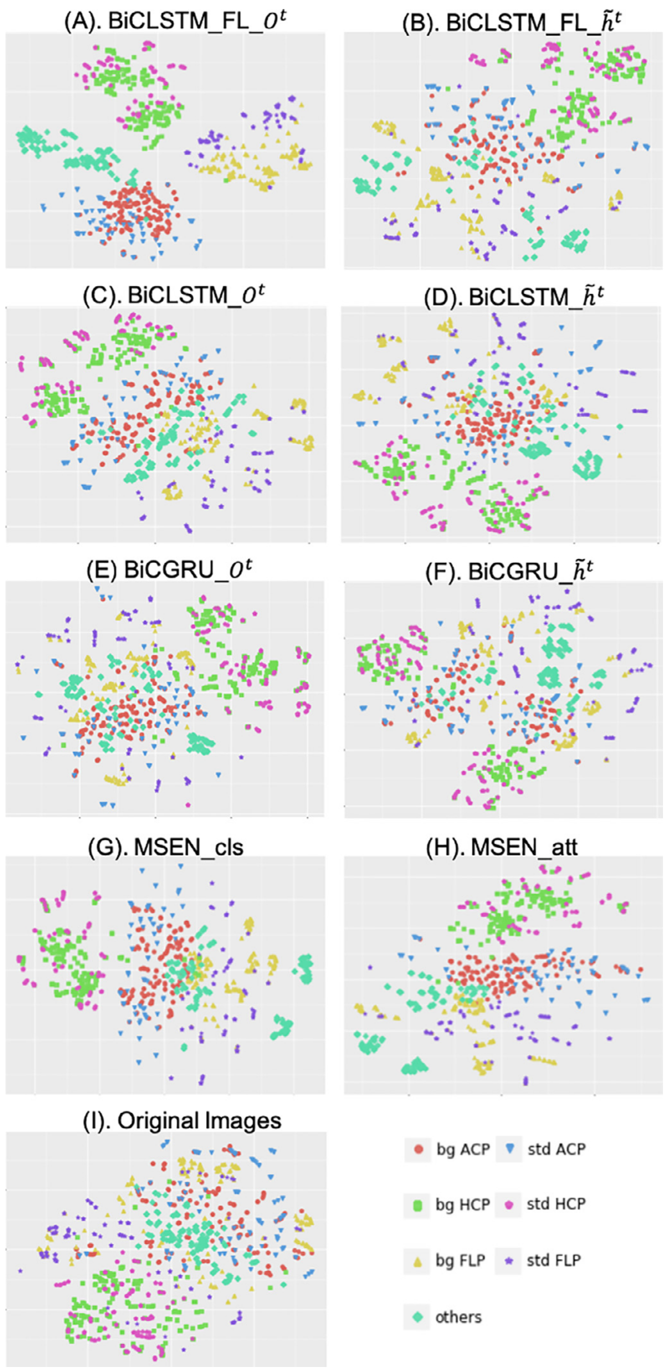


Fig. 19. t-SNE visualisation of the feature embedding of selected TSEN variants as well as base line model MSEN. The suffix “O^t” indicates the output features of the VCM (Fig. 9), while the suffix “$\tilde{h}^t$” indicates the last features of TAM (Fig. 8). “MSEN_cls” and “MSEN_att” indicates the last features from the classification and saliency prediction branches, respectively (Cai et al., 2018b).

one of two modes: assistive, or fully-automated. When in an assistive mode, the selection of standard biometry planes, and thus the final clinical decision-making, is still fully controlled by the sonographer; in this case the accuracy of TSEN overall minimally affects the clinical decision but will affect workflow efficiency. On the other hand, if TSEN was used in a fully-automated mode, inaccuracies in automated standard plane detection would in turn directly affect accuracy of biometry estimation and the usability of the combined solution in practice. Usability also depends on the intended end-user (experienced sonographer *versus* occasional user for instance). Future work would need to study this.

It has been demonstrated that TSEN models successfully learn temporal visual attention and perform better than the MSEN model, both qualitatively and quantitatively, even though the latter was demonstrated to produce good quality visual attention maps in Cai et al. (2018b). The disparity is attributed to the difference in the nature of data used, and equally importantly, the tasks that sonographers were performing when gaze-tracking data were recorded. In Cai et al. (2018b), sonographers had the freedom to view each single frame for as long as they wanted, allowing full inspection of the contents in that particular frame. Thus, gaze information recorded on one frame were less dependent on the other frames, which is different from the gaze-tracking data recorded in the PULSE dataset. The gaze-tracking data used here were recorded in real-time during anomaly scan sessions. Therefore, without RNNs to encode spatio-temporal information from consecutive frames with dependent gaze information, MSEN cannot model the behavior of sonographers sampling the visual field both spatially and temporally.

The question of which module is better, LSTM or GRU, for modelling temporal information has long been discussed and the results are not conclusive. The consensus was GRUs are at least comparable to LSTM, and performance depends on specific tasks involved (Chung et al., 2014). Our observation is that TSEN variants trained with CLSTM consistently outperform those with CGRU (*e.g.* in Table 1, “biCLSTM” v.s. “biGRU” and “uniCLSTM” v.s. “uniCGRU”). This is most probably due to the additional gating mechanism that CLSTM employs (thus additional parameters) that allows such TSEN variants to better model complex temporal variation of sonographer visual attention though with additional computational expenses. In addition, our observation is consistent with literature that bi-directional RNNs are better at modelling temporal information than uni-directinal RNNs (Schuster and Paliwal, 1997). Giving TSEN models the ability to examine video frames from both directions allows them to mimic sonographer behavior of comparing several candidate frames iteratively before making final decisions.

It is interesting to notice that TSEN models generally perform better on standard snippets than on non-standard snippets of each biometry, as demonstrated in Fig. 16. Abdomen and head are the two anatomies showing higher static saliency scores and scanpath similarity scores in standard snippets than non-standard snippets. This can be attributed to the fact that abdomen and head are less intuitive to interpret, and sonographers follow the protocol guidelines when potential candidate standard planes appear. For example, sonographers consistently search for the stomach bubble and umbilical veins, according to the study (Ahmed and Noble, 2016), with constant reference to the spine when determining the standard ACP. This consistency makes it easier to learn spatio-temporal transitions of visual attention. In non-standard snippets, sonographer gaze data is more unstructured, resulting in less well localised visual attention predictions with occasional failure, as can be seen in Fig. 20, where a non-standard abdomen snippet is presented. In the five consecutive video frames (top to bottom on in the left column), key anatomical structures such as the umbilical vein and stomach bubble disappear in the third frame and re-appear in the next two frames, possibly due to out-of-plane rotation of the probe. In this case, TSEN did not replicate sonographer visual attention; rather, it predicted visual attention with less certainty by spreading predictions in large areas. The reason for such occasional failure is a subject of further analysis.

As demonstrated in Tables 3 and 4, the Dynamic Attention Maps predicted by TAM of TSEN models assist frame classification tasks, supported by the fact that F1 scores of ACP, HCP and FLP all demonstrated significant improvement compared to those of baseline models. All TSEN models are built upon the architecture used

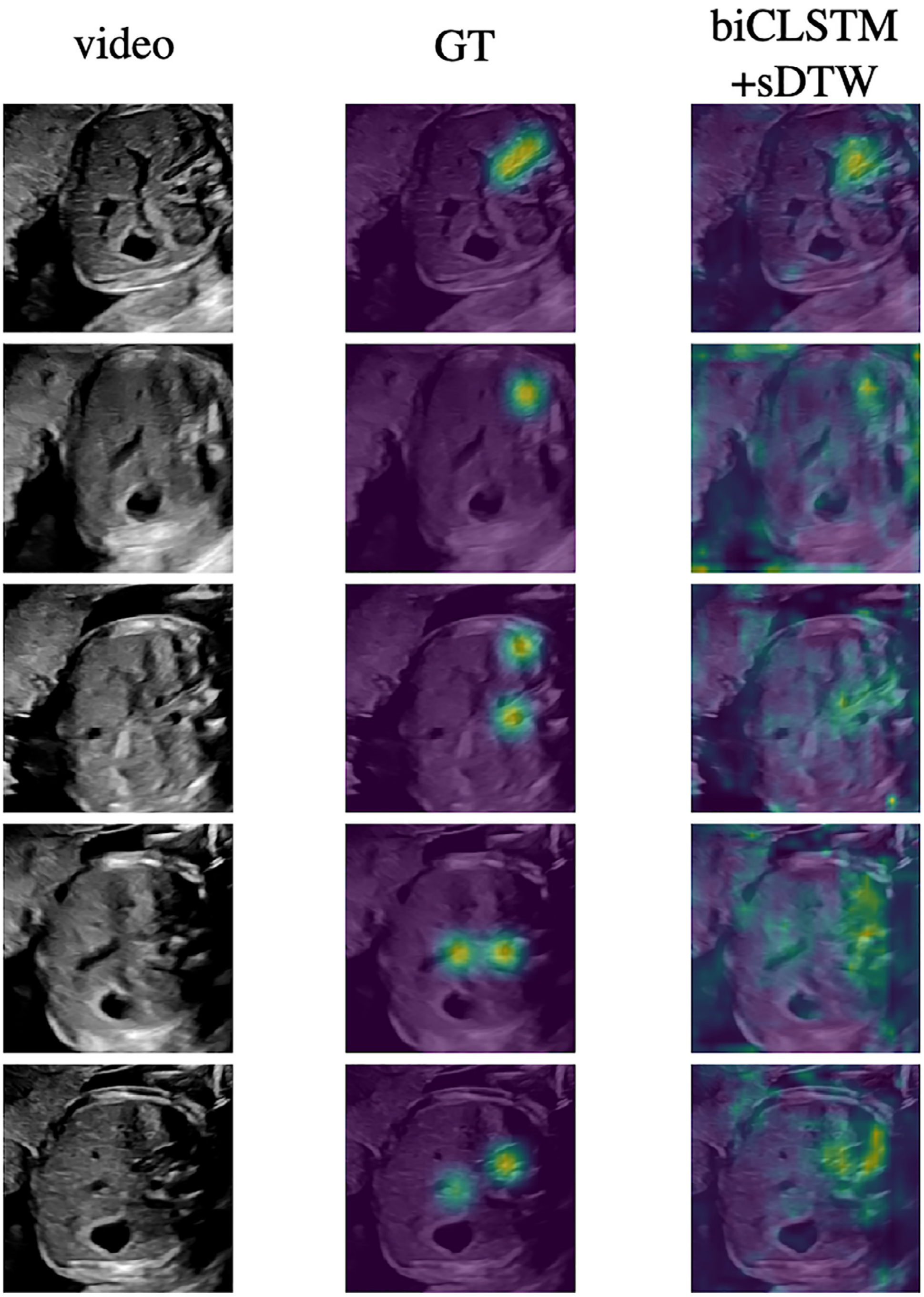


Fig. 20. A failure case where visual attention prediction becomes less localised in a non-standard abdominal video snippet.

in SonoNet-16 with slightly more computational overhead, and the frame classification performance surpasses those of heavier models of SonoNet-64 and SonoNet-32 trained on PULSE dataset. This result further supports the idea that sonographer gaze information assists frame classification task, and learning temporal information in US video snippet and gaze data can further improve frame classification performance.

Feature embedding visualisation using t-SNE, as presented in Fig. 19, demonstrated that features learnt for visual attention prediction separates samples of different classes even though it did not receive any information regarding sample classes. As demonstrated by Droste et al. (2019a), feature representations learnt for visual attention modelling on 2D US video frames are predictive for fetal anomaly standard plane detection. It will be of research interest to see if such spatio-temporal features learnt for visual attention modelling on US video snippets are more discriminative for frame prediction.

All TSEN models were built on the gaze information of a single sonographer. This was determined by the nature of simultaneous gaze-tracking experiment where each video was viewed only by one sonographer, unlike in retrospective gaze-tracking where a US scan video could be presented to multiple sonographers for viewing. However, our method can be generalized to potentially incorporate gaze information from multiple sonographers. The design of such a gaze-tracking experiment, and the mode of combining multiple sources of gaze information (with different viewing behaviour) to train a single visual attention predictor, are worth further explorations.

In conclusion, we have proposed TSEN, a deep-learning based architecture that effectively learns the temporal visual attention transitions of a sonographer from 2D US video snippets, and utilizes the predicted sonographer visual attention maps for finding three standard fetal biometry planes. The proposed TSEN model achieves better performance in visual attention prediction and frame classification tasks compared to models that learn only spatial information. It is found that the best architecture employed a bi-directional CLSTM to model spatio-temporal information, and sDTW as an effective and novel loss function for visual attention regularisation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yifan Cai: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Richard Droste:** Conceptualization, Methodology, Software, Investigation, Data curation, Writing - review & editing. **Harshita Sharma:** Conceptualization, Methodology, Investigation, Data curation, Writing - original draft, Writing - review & editing, Project administration. **Pierre Chatelain:** Conceptualization, Methodology, Investigation, Data curation, Writing - original draft, Writing - review & editing, Project administration. **Lior Drukker:** Conceptualization, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing. **Aris T. Papageorghiou:** Conceptualization, Investigation, Resources, Data curation, Writing - review & editing, Supervision. **J. Alison Noble:** Conceptualization, Methodology, Investigation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

We acknowledge the ERC (ERC-ADG-2015 694581, project PULSE), the EPSRC (EP/GO36861/1, EP/MO13774/1, EP/R013853/1), and the NIHR Biomedical Research Centre funding scheme.

References

Abramowicz, J.S., 2013. Benefits and risks of ultrasound in pregnancy. In: Seminars in Perinatology, 37. Elsevier, pp. 295–300.

Ahmed, M., 2014. The fusion of eye tracking and machine learning to boost detection of anatomical features in fetal abdominal ultrasound.

Ahmed, M., Noble, J.A., 2016. An eye-tracking inspired method for standardised plane extraction from fetal abdominal ultrasound volumes. In: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on. IEEE, pp. 1084–1087.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging 36 (11), 2204–2215.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. In: European Conference on Computer Vision. Springer, pp. 404–417.

Bellman, R., 1952. On the theory of dynamic programming. Proc. Natl. Acad. Sci. U.S.A. 38 (8), 716.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F., 2018. What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell..

Cai, Y., Sharma, H., Chatelain, P., Noble, J., 2018. Sonoeyenet: standardized fetal ultrasound plane detection informed by eye tracking. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 1475–1478.

Cai, Y., Sharma, H., Chatelain, P., Noble, J.A., 2018. 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. Springer, pp. 871–879.

Chatelain, P., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2018. Evaluation of gaze tracking calibration for longitudinal biomedical imaging studies. IEEE Trans. Cybern. (99) 1–11.

Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A., 2015. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE J. Biomed. Health Inform. 19 (5), 1627–1636.

Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cuturi, M., Blondel, M., 2017. Soft-dtw: a differentiable loss function for time-series. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 894–903.

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., Holmqvist, K., 2012. It depends on how you look at it: scanpath comparison in multiple dimensions with multimatch, a vector-based approach. Behav. Res. Methods 44 (4), 1079–1100.

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

Droste, R., Cai, Y., Sharma, H., Chatelain, P., Papageorghiou, A., Noble, J., 2019b. Towards capturing sonographic experience: cognition-inspired ultrasound video saliency prediction.

Gao, Y., Maraci, M.A., Noble, J.A., 2016. Describing ultrasound video content using deep convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 787–790.

Giles, C.L., Kuhn, G.M., Williams, R.J., 1994. Dynamic recurrent neural networks: theory and applications. IEEE Trans. Neural Networks 5 (2), 153–156.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Huang, X., Shen, C., Boix, X., Zhao, Q., 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 262–270.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach.Intell. (11) 1254–1259.

James, A., Vieira, D., Lo, B., Darzi, A., Yang, G.-Z., 2007. Eye-Gaze Driven Surgical Workflow Segmentation. Springer, pp. 110–117.

Jarodzka, H., Holmqvist, K., Nyström, M., 2010. A vector-based, multidimensional scanpath similarity measure. In: Proceedings of the 2010 symposium on eye-tracking research & applications. ACM, pp. 211–218.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kirwan, D., 2010. Nhs fetal anomaly screening programme. 18+ 0 to 20+ 6 Weeks Fetal Anomaly Scan National Standards and Guidance for England.

Koch, C., Ullman, S., 1987. Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of Intelligence. Springer, pp. 115–141.

Lawn, J.E., Cousens, S., Zupan, J., Team, L.N.S.S., et al., 2005. 4 Million neonatal deaths: when? Where? Why? Lancet 365 (9462), 891–900.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Mongelli, M., Ek, S., Tambyrajia, R., 1998. Screening for fetal growth restriction: a mathematical model of the effect of time interval and ultrasound error. Obstetr. Gynecol. 92 (6), 908–912.

Nodine, C.F., Kundel, H.L., 1987. Using eye movements to study visual search and to improve tumor detection.. Radiographics 7 (6), 1241–1250.

Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., et al., 2014. International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. Lancet 384 (9946), 869–879.

Pearlmutter, B.A., 1989. Learning state space trajectories in recurrent neural networks. Neural. Comput. 1 (2), 263–269.

Ramanathan, S., Katti, H., Huang, R., Chua, T.-S., Kankanhalli, M., 2009. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In: Proceedings of the 17th ACM International Conference on Multimedia. ACM, pp. 729–732.

Sakoe, H., Chiba, S., Waibel, A., Lee, K., 1990. Dynamic programming algorithm optimization for spoken word recognition. Read. Speech Recognit. 159, 224.

Salomon, L., Bernard, J., Duyme, M., Doris, B., Mas, N., Ville, Y., 2006. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. Ultrasound Obstetr. Gynecol. 27 (1), 34–40.

Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D., Papageorghiou, A., Fetal, I., for the 21st Century (INTERGROWTH-21st), N.G.C., 2012. Intra-and interobserver variability in fetal ultrasound measurements. Ultrasound Obstetr. Gynecol. 39 (3), 266–273.

Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention-gated networks for improving ultrasound scan plane detection. In: International Conference on Medical Imaging with Deep Learning (MIDL) 2018.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45 (11), 2673–2681.

Shanmuga Vadivel, K., Ngo, T., Eckstein, M., Manjunath, B., 2015. Eye tracking assisted extraction of attentionally important objects from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3241–3250.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Treisman, A.M., Gelade, G., 1980. A feature-integration theory of attention. Cognit. Psychol. 12 (1), 97–136.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164.

Wang, W., Shen, J., Guo, F., Cheng, M.-M., Borji, A., 2018. Revisiting video saliency: a large-scale benchmark and a new model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4894–4903.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, pp. 802–810.

Xu, J., Collins, M.D., Singh, V., 2013. Incorporating user interaction and topological constraints within contour completion via discrete calculus. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1886–1893.

Yaqub, M., Kelly, B., Papageorghiou, A.T., Noble, J.A., 2015. Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 687–694.