Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps

Yifan Cai (✉), Harshita Sharma, Pierre Chatelain, and J. Alison Noble

Institute of Biomedical Engineering, University of Oxford,
Old Road Campus Research Building, Oxford OX3 7DQ, UK
yifan.cai@eng.ox.ac.uk

Abstract. We present a novel multi-task convolutional neural network called Multi-task SonoEyeNet (*M-SEN*) that learns to generate clinically relevant visual attention maps using sonographer gaze tracking data on input ultrasound (US) video frames so as to assist standardized abdominal circumference (AC) plane detection. Our architecture consists of a generator and a discriminator, which are trained in an adversarial scheme. The generator learns sonographer attention on a given US video frame to predict the frame label (standardized AC plane / background). The discriminator further fine-tunes the predicted attention map by encouraging it to mimick the ground-truth sonographer attention map. The novel model expands the potential clinical usefulness of a previous model by eliminating the requirement of input gaze tracking data during inference without compromising its plane detection performance (Precision: 96.8, Recall: 96.2, F-1 score: 96.5).

Keywords: multi-task learning, generative adversarial network, gaze tracking, fetal ultrasound, saliency prediction, standardized plane detection.

1 Introduction

The detection of fetal Intra-Uterine Growth Restriction using ultrasound-based diagnostic methods relies on the detection of standardized 2D ultrasound (US) planes for several biometric measurements, such as the abdominal circumference (AC), the head circumference (HC), the bi-parietal diameter (BPD), and the femur length (FL) [1]. Increasing demand for sonographers [2] has encouraged attempts to automate standardized plane detection using Random Forests [3] and, more recently, convolutional neural networks (CNNs) [4]. However, the large amount of labeled data required to train traditional classification CNNs is normally not available in medical image analysis, which requires more efficient use of the limited time of clinical experts available for labeling data. In the clinical domain of interest here, sonographer eye movements are a strong prior for human interpretation of US video frames. Recently, the SonoEyeNet [5] architecture was proposed that used sonographer gaze tracking data in tandem with US

video frames as two inputs to detect standard AC planes. However, SonoEyeNet requires sonographer gaze tracking data for inference, which significantly limits its usefulness in a clinical setting. In this paper, we address this limitation by proposing a novel framework that learns sonographer gaze tracking data to predict visual attention maps which assist in standardized plane detection without the requirement of gaze tracking data for inference.

Contributions. This paper proposes an end-to-end multi-task CNN called Multi-task SonoEyeNet (*M-SEN*) with a primary task to classify abdominal fetal ultrasound video frames into standard AC planes or background, and an auxiliary task to predict sonographer visual attention on those frames to assist the primary task. This novel architecture (I) substantially expands the potential usefulness of SonoEyeNet [5] in a clinical setting as a biometry assistance tool by removing the requirement of input gaze tracking data for inference without compromising frame classification performance; (II) adopts a novel adversarial regulariser to improve the quality of the generated attention map, which subsequently improves frame classification results; (III) demonstrates that the novel gaze tracking data collected during clinical experts' labeling time can also be learnt to improve model performance on a moderately-sized dataset, which makes the approach attractive for modeling other medical imaging problems.

Related work. Two themes related to this work are saliency prediction and US image classification. Many attempts have been made to model human visual attention: from models built purely from hand-crafted features [6] to the state-of-the-art that learns image features using deep neural networks [7]. Generative Adversarial Networks (GAN) were recently used to model human attention on natural images [8]. For US image classification using images alone, Yaqub *et al.* [3] automated standardized plane detection using Random Forests; Baumgartner *et al.* used transfer learning and a FCN to detect 12 standard planes [4]. The first attempt to use sonographer gaze tracking data to assist standard AC plane detection was [9], which mimicked sonographer's visual behavior using a pictorial structures model inspired by observing human eye movement patterns. Recently, SonoEyeNet [5] was the first to implement gaze tracking data assisted standardized plane detection within a deep learning framework.

2 Methods

***M-SEN* architecture.** The *M-SEN* architecture consists of two CNN modules: the generator (G) and the discriminator (D). It is summarized in Fig. 1.
Generator Architecture. G is a multi-task module that can be trained independently without D to generate both a predicted visual attention map $\hat{A}$ and a classification score vector of the input frame $\hat{y}$: $(\hat{A}, \hat{y}) = G(I; \theta_G)$, where I represents the US video frame and θ_G represents weights of the generator network. First, image features are extracted using the first three convolutional blocks of a

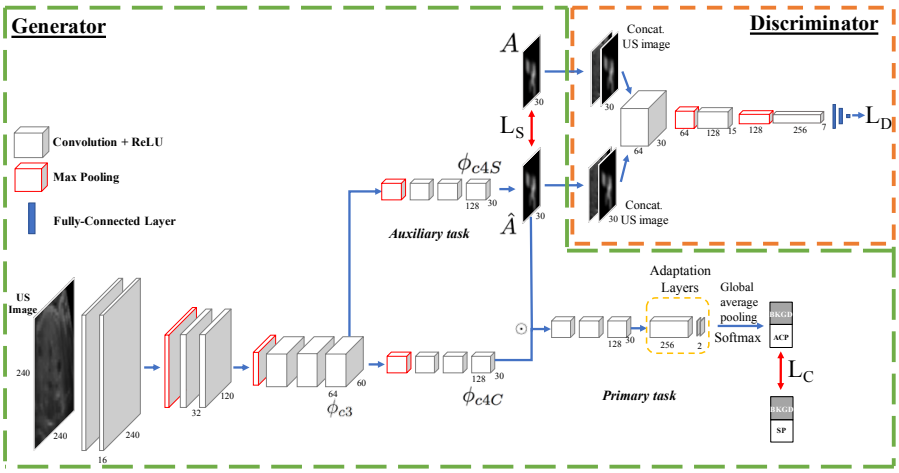

Fig. 1. Architecture of the multi-task SonoEyeNet (*M-SEN*). It has two modules: the generator (in Green-dashed polygon) and the discriminator (Orange-dashed box). The generator has two tasks: a primary task to classify frames (bottom) and an auxiliary task to predict visual attention map ($\hat{A}$). The discriminator differentiates between real (A) and predicted ($\hat{A}$) attention maps. The dotted circle $\odot$ indicates element-wise multiplication. L_S, L_C and L_D represent the losses of saliency prediction, frame classfication, and the discriminator respectively.

pre-trained SonoNet [4]. All layers use 3×3 convolutional kernels and the number of kernels used can be seen in Fig. 1. Feature maps are down-sampled by a factor of 2 after each block using max pooling. The network separates into two branches after the third convolutional block: one branch for the auxiliary task of saliency prediction so as to mimic sonographer visual attention on US video frames; the other for frame classification. Feature maps ϕ_{c3} are first spatially down-sampled by 4 and then passed through 3 convolutional layers with 3×3 kernels in each branch. This produces ϕ_{c4S} and ϕ_{c4C} for saliency prediction and classification respectively. Convolution with 1×1 kernels is performed on ϕ_{c4S} to generate $\hat{A}$. The attention map $\hat{A}$ is then fused with ϕ_{c4C} through element-wise multiplication. The resultant feature maps are passed through another convolutional block, and then through two adaptation layers [4] which used 256 and two 1×1 kernels respectively. Global average pooling on the two resultant feature maps is performed before softmax so as to predict class scores for the standard AC plane and background. Classification loss L_C is defined between the predicted class scores and actual label, and saliency loss L_S between the predicted and the actual sonographer visual attention maps.

Discriminator architecture. The discriminator module is a CNN with three convolutional layers with 64, 128 and 256 3×3 kernels respectively, each with max pooling and leaky rectified linear unit (leaky ReLU) activation, followed

by three fully-connected (FC) layers. Hyperbolic tangent (tanh) activation was used for the first two FC layers and sigmoid activation for the last FC layer. As sonographer attention is conditional on the US video frame, I is concatenated to A or $\hat{A}$ as inputs into D.

Loss functions. The discriminator loss L_D and the generator loss L_G for each mini-batch with m samples are defined [10] as:

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

$$L_G = \lambda_1 L_S + \lambda_2 L_C - \frac{\lambda_3}{m}\sum_{i=1}^{m} \log(D(I_i, \hat{A}_i; \theta_D)) \tag{2}$$

where $D(I, A; \theta_D)$ is the probability that the discriminator successfully recognizes the real attention map, while $D(I, \hat{A}; \theta_D)$ is the probability that the discriminator is fooled. θ_D represents the weights of D. The generator loss is designed to include both classification and saliency losses L_C and L_S as well as an adversarial regulariser by using the discriminator loss on $\hat{A}$; this regulariser was not used during generator pre-training. Hyper-parameters λ_1, λ_2, and λ_3 determine the relative contributions of the three losses. The saliency loss L_S is defined as the pixel-level content loss between $\hat{A}$ and A, which is used to train the generator of the attention maps. Two loss functions were experimented with in the models shown in Table 1: *M-SEN MSE* uses the mean squared error (MSE) loss, a base-line loss as it has been used in many visual saliency prediction works [11]; *M-SEN BCE* uses binary cross-entropy (BCE) loss, which is mathematically equivalent to Kullback-Leibler divergence, arguably the best metric to measure saliency prediction performance [12]. For the classification task, cross-entropy loss was used as L_C, the same as in [5].

Training details. The generator was independently pre-trained for 30 epochs before adding the discriminator as the adversarial regulariser. The network was initialized using the first three convolutional blocks of SonoNet [4]; all other layers were initialized using a zero-mean Gaussian distribution with standard deviation 0.01. Batch normalization and dropout (rate = 0.2) were used for each convolutional layer before the adaptation layers. The weight λ_1 was dynamically changed from 2 to 1 over epochs so as to allow the generator to focus on learning attention maps first and then frame classification. The weight λ_2 was set to 1, as classification was the primary task of the network. After 30 epochs, the network was further fine-tuned for 2000 steps using an adversarial training scheme by training the discriminator and the generator once per step in an alternating manner. When training the discriminator, one-sided label smoothing [13] was used; when training the generator, the weights of discriminator were not updated. The network was trained using adaptive moment estimation (Adam) with an initial learning rate 2×10^{-4}. Batch size was set to 64. Five-fold cross-validation was used.

US dataset and preprocessing. The dataset was acquired following a free-hand US sweep protocol by moving the probe from the bottom to the top of pregnant women's abdomens. The dataset consists of 1616 frames from 33 fetal abdominal US video clips, each belonging to a unique patient. Each frame was assigned a class either the standard AC plane (ACP) or background (BG). Data augmentation by horizontal flipping and rotation was performed; equal number of ACP and BG frames were sampled for each batch during training. All frames were cropped at the positions of the abdominal wall and resized to 240×240 pixels. The dataset was separated video-wise into training and testing sets: frames from 25 videos clip (80% of all clips) were used for training, and those from remaining clips for testing.

Eye movement acquisition and filtering. Gaze tracking data (x-y coordinates and time stamps) were acquired and filtered following the protocol in [5]. An eye tracker (The EyeTribe) recorded sonographer gaze tracking data at 30 Hz when they identified standard ACPs in each US video clips. A temporal moving average filter of window size 3 (100 ms) was applied to remove high frequency noise caused by eye tremor. Angular velocity threshold of $30°/s$ [14] was applied to separate fixations from saccades. Fixations $0.5°$ visual angle apart were merged, and those shorter than 80 ms in durations were discarded. A binary map of fixation points (1) and saccades or background (0) of 240×240 pixels was thus created. Since the human field of view typically extends to $1.5°$ visual angle [15] around a fixation point, the binary map was convolved with a 2-D Gaussian kernel with $\sigma = 30$ pixels, given an observer-to-screen distance of 0.5 m and screen dimensions of 20.7 cm $\times$ 33.2 cm. Attention maps corresponding to each frame, as detailed in the next section, were processed the same way by cropping, rotation and flipping.

3 Results and Discussion

Frame classification performance. Classification results of all models are presented in Table 1. Two observations can be made. First, all *M-SEN* models that use learned saliency maps to assist frame detection outperform the *SonoNet* models [4], which are supervised only by image-level labels. The classification precision of 79.3% for the *SonoNet-32* model was increased to 96.8% ($p < 0.05$) in *M-SEN BCE + GAN*, which uses BCE as saliency loss and is further fine-tuned using adversarial regulariser. Recall increased from 82.1% to 96.2% ($p < 0.05$). Second, models that adopted adversarial regularizers achieve better results: performances of both BCE and MSE models are improved by training with an adversarial discriminator. For example, introducing adversarial regularizer to *M-SEN MSE* increased its precision from 92.4% to 94.8% ($p < 0.05$), recall from 75.6% to 91.9% ($p < 0.05$), and F-1 score from 83.2% to 93.3% ($p < 0.05$). The best performing *M-SEN BCE + GAN* model achieves performance competitive to that of *SonoEyeNet-Late FT* [5] ($p = 0.692$), which uses both the US frame and sonographer visual attention map for inference. *SS-cls Net* that attempts to classify US video frames solely on sonographer visual attention map achieved a performance similar to that of *SonoEyeNet-Late FT*.

Table 1. *Comparative evaluation of classification performance. In column "Inputs", "I"and "A" refer to US images and attention maps, respectively. "SS-cls Net" refers to single-stream network trained only on attention maps to classify US video frames.*

Models	Inputs	Precision	Recall	F1-score
M-SEN BCE + GAN	I	96.8	96.2	96.5
M-SEN BCE	I	96.7	90.5	93.5
M-SEN MSE + GAN	I	94.8	91.9	93.3
M-SEN MSE	I	92.4	75.6	83.2
SonoNet-32 [4]	I	79.3	82.1	80.7
SonoNet-16 [4]	I	73.6	74.1	73.8
SS-cls Net	A	71.5	76.4	73.9
SonoEyeNet-Late FT [5]	I and A	96.5	99.0	97.8

Table 2. *Quantitative metrics of saliency prediction on the test set. "SS-att" indicates those single-stream models for saliency prediction without a classification branch. Saliency metrics include information gain (IG), pearson's cross-correlation (CC), normalized saliency scan path (NSS), similarity (SIM), and area under curve (AUC) [16].*

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

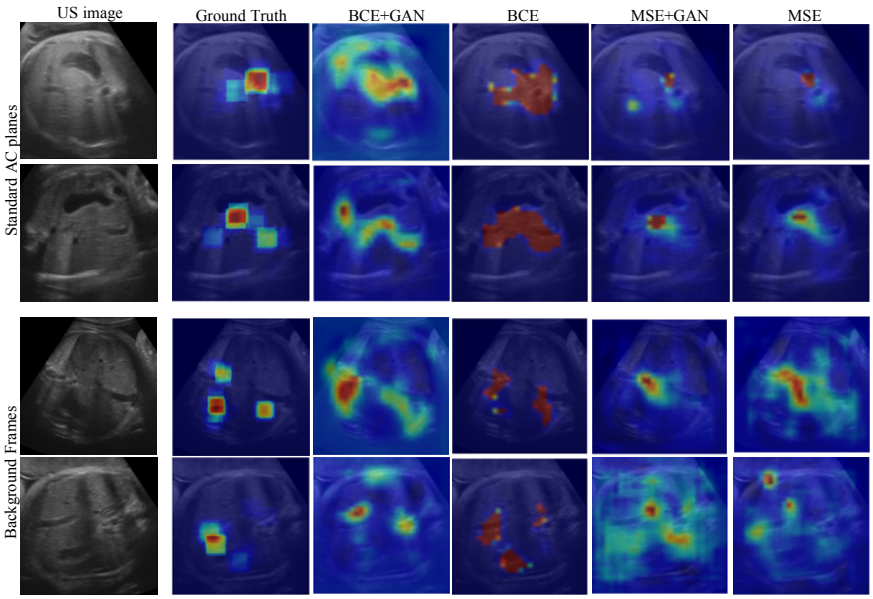

Fig. 2. Attention maps generated by different variations of the Generator. From left to right: US image, Gound-truth (Sonographer's actual attention map), *M-SEN BCE + GAN*, *M-SEN BCE*, *M-SEN MSE + GAN*, *M-SEN MSE*.

Saliency prediction performance. Examples of predicted visual attention maps generated by variations of *M-SEN* models on the test set US video frames can be seen in Fig. 2. All independently trained *M-SEN* models generate visually good quality attention maps that capture the salient regions fixated by sonographers, e.g. edges of the stomach bubble. Interestingly, *M-SEN BCE* extends beyond the constraint of the ground-truth salient regions and explores other key anatomical structures, e.g. the umbilical veins, that the sonographers had not necessarily looked at during examination. In addition, we observe a similar trend previously observed in Table 1 that adding an adversarial regulariser improves model performance. *M-SEN BCE + GAN* is able to learn a more realistic visual attention map while retaining the ability to assign confidence to other key anatomical structures. Adding an adversarial discriminator regularises the predicted saliency map by reducing confidence in its false-positive points, as can be observed in Fig. 2; this can also be observed in Table 2, where saliency scores measured by all five metrics increase.

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

In general, MSE models generate attention maps with more false-negatives (predicted as non-salient while fixated by sonographer), while BCE models generate more false-positives (predicted as salient but not fixated by sonographer). A change of loss function from *M-SEN MSE* to *M-SEN BCE* improves saliency performance measured by four out of five metrics. The performance improvement is consistent with the way gaze tracking data was collected. Our experiments allow sonographers to inspect each frame as long as they want, so each pixel can be modelled as an independent binary variable (fixated or not), which is best modelled using BCE loss; MSE, on the other hand, does not assume pixel independence and models the probability distribution of fixation on a frame in a brief glimpse. As confirmed by other literature [8], BCE loss performs better than MSE loss. The only exception is in NSS, where *M-SEN MSE* outperforms *M-SEN BCE*. This can be attributed to the fact that NSS is extremely sensitive to false positives, which *M-SEN BCE* exploits to cover non-fixated anatomical structures to benefit classification task.

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

4 Conclusion

This paper presents a novel and effective algorithm that models sonographer visual attention on US video frames to assist frame classification in an end-to-end deep learning framework. The multi-task network surpasses a previously reported model in classification performance [4]; it has great potential to be used in a clinical setting as it doesn't require gaze tracking data during inference. Our result suggests that it is better to model saliency prediction as a binary classification problem (using BCE loss), rather than a simple regression (using MSE loss). Adopting an adversarial regulariser proves to be effective in fine-tuning the generated attention maps, which further assists the classification task. The presented approach is general; the pipeline of gaze-tracking data collection, multi-task network design and adversarial regulation could be generalized to a wide range of other medical imaging challenges.

Acknowledgments. We acknowledge the ERC (ERC-ADG-2015 694581 for project PULSE) and the EPSRC (EP/GO36861/1, and EP/MO13774/1).

References

1. Hack, M., et al.: Outcomes of extremely low birth weight infants. Pediatrics **98**(5) (1996) 931–937
2. Sarris, I., et al.: Intra- and inter-observer variability in fetal ultrasound measurements. Ultrasound in Obstetrics & Gynecology **39**(3) (2012) 266–273
3. Yaqub, M., et al.: Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans. In: MICCAI, Springer (2015) 687–694
4. Baumgartner, C.F., et al.: SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Transactions on Medical Imaging **36**(11) (2017) 2204–2215
5. Cai, Y., et al.: SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking. In: IEEE ISBI. (2018) 1475–1478
6. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of Intelligence. Springer (1987) 115–141
7. Kümmerer, M., et al.: DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563 (2016)
8. Pan, J., et al.: SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. In: arXiv:1701.0181v2 [cs.CV]. (January 2017)
9. Ahmed, M., et al.: An eye-tracking inspired method for standardised plane extraction from fetal abdominal ultrasound volumes. In: ISBI, IEEE (2016) 1084–1087
10. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS. (2014) 2672–2680
11. Kruthiventi, S.S., et al.: Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing **26**(9) (2017) 4446–4456
12. Huang, X., et al: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: ICCV. (2015) 262–270
13. Radford, A., et al.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
14. Fuchs, A.F.: The saccadic system. The control of eye movements (1971) 343–362
15. Mathe, S., Sminchisescu, C.: Dynamic eye movement datasets and learnt saliency models for visual action recognition. ECCV (2012) 842–856
16. Bylinskii, Z., et al.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)