2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761585

FIRST TRIMESTER VIDEO SALIENCY PREDICTION USING CLSTMU-NET WITH STOCHASTIC AUGMENTATION

Elizaveta Savochkina[1] *Lok Hin Lee*[1] *He Zhao*[1] *Lior Drukker*[2,3]
Aris T. Papageorghiou[2] *J. Alison Noble*[1]

[1]Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2]Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK
[3]Rabin Medical Center, Sackler Faculty of Medicine, Tel-Aviv University, Israel

ABSTRACT

In this paper we develop a multi-modal video analysis algorithm to predict where a sonographer should look next. Our approach uses video and expert knowledge, defined by gaze tracking data, which is acquired during routine first-trimester fetal ultrasound scanning. Specifically, we propose a spatio-temporal convolutional LSTMU-Net neural network (cLSTMU-Net) for video saliency prediction with stochastic augmentation. The architecture design consists of a U-Net based encoder-decoder network and a cLSTM to take into account temporal information. We compare the performance of the cLSTMU-Net alongside spatial-only architectures for the task of predicting gaze in first trimester ultrasound videos. Our study dataset consists of 115 clinically acquired first trimester US videos and a total of $45,666$ video frames. We adopt a Random Augmentation strategy (RA) from a stochastic augmentation policy search to improve model performance and reduce over-fitting. The proposed cLSTMU-Net using a video clip of 6 frames outperforms the baseline approach on all saliency metrics: KLD, SIM, NSS and CC (2.08, 0.28, 4.53 and 0.42 versus 2.16, 0.27, 4.34 and 0.39).

***Index Terms*—** Fetal ultrasound, first trimester, gaze tracking, video saliency prediction, U-Net, convolutional LSTM, stochastic augmentation.

1. INTRODUCTION

Our interest is in automating the steps in first-trimester fetal ultrasound (US) scanning. In this paper we focus on the imaging guidance task and specifically on automating the task of predicting where a sonographer should look next. Human visual attention is typically quantified via the distribution of gaze points, hereafter referred to as a *saliency map*. Our assumption is that automatic prediction of saliency maps can assist in the guidance to imaging planes, and hence potentially help a non-expert with abnormality finding.

This work is supported by the ERC (ERC-ADG-2015694581, project PULSE) and the EPSRC (EP/R013853/1 and EP/T028572/1). AP is funded by the NIHR Oxford Biomedical Research Centre.

Related work: Salvador et al. [1] designed an encoder-decoder network for semantic instance segmentation where an encoder is used for classification and a decoder is composed of a series of cLSTM layers merged with the encoder outputs in the form of skip connections. Xu et al. [2] proposed an LSTM multi-modal U-Net for brain tumor segmentation using hyper-dense connectivity to leverage different MRI modalities and temporal information. The authors first use a multi-modal U-Net to produce a pixel-wise segmentation mask which is then fed into the cLSTM. Unlike [1][2], we use gaze-tracking data as a strong prior to guide the model towards important US structures. Wu et al. [3] constructed a SalSAC network for video saliency prediction which follows a CNN-shuffle attention module-cLSTM pipeline. Similarly, we use encoder-decoder with cLSTM in the middle. Yet, we process the temporal input outside the spatial U-Net, pass it through the cLSTM and feed into the bottom of the decoder.

Previous work on gaze prediction for fetal ultrasound has been reported for the second trimester and first trimester [4, 5]. Savochkina et al. [5] investigated the prediction of spatial gaze distribution for the first trimester ultrasound. However, that approach could not differentiate between fast and slow-moving video segments due to lack of knowledge of the previous frames. In addition to the trimester of application, our approach is different to [4] in the use of spatio-temporal gaze patterns together with US video in training. Specifically, we utilise an encoder-decoder network with skip connections and add temporal information to improve the saliency prediction through cLSTM, exploiting the relationship between consecutive US video frames. Different from the above mentioned works that achieve performance gains due to pre-training on large image datasets, our model is trained from scratch.

Contribution: Our contribution is two-fold. First, we consider video saliency prediction (VSP) from a first trimester multi-modal US dataset. The model learns a mapping between the US and ground truth (GT) saliency maps in routine first trimester scans, predicting gaze for all structures and planes that come into sonographer view. Second, we propose a new variation of a U-Net [6] with feature sharing between 2 in-

puts where an additional cLSTM module incorporates temporal information, learns an intra-dependence of frames within a sequence, and enforces a better data representation.

2. METHODS

2.1. Data and Data Preparation

For our experiments, we use 115 first trimester US videos and a total of $45,666$ video frames with a $60/20/20$ training/validation/test split. We use the same data and preparation steps that are detailed in [5].

The input to the spatial U-Net is a single frame which is used to predict a saliency map, whilst the input for the cLSTM is the combination of temporal frames and the same single frame. We performed an ablation study, reported in Table 2, to evaluate which placement of additional temporal frames with regards to a single frame (before or after) adds the most value to the model prediction performance.

To accommodate the cLSTM module, the data is sampled as shown on the left of Fig. 1, where a fixed video segment is an input to the cLSTM. We take training samples from the original video dataset using a shifting window that is the width of the desired frame length. This way, we capture temporal variation without loss of temporal resolution, i.e we sample with an interval of 1 frame, therefore, consider all the temporal change. Each video contains 90 frames (3 sec), therefore, the number of video segments that can fit in 3 seconds is: # Video segments = 90 frames - Video segment. In addition, we investigate different video clip lengths and select one by its performance, as summarized in Table 1.

2.2. Gaze and Image Augmentation

To reduce over-fitting, we employ a Random Augmentation (RA) strategy using stochastic augmentation policy search for segmentation purposes [5][7].

We adopt a grid search with fixed magnitude schedule and a total of K=16 transformations, as in [5]. Each augmentation policy is defined by n, which is the number of transformations from the list of K an image undergoes, and m, which is the magnitude distortion of each transformation. These transformations are applied to the mixed-example images (Mix.RA), with which we share the m hyperparameter.

We empirically compared two random augmentation (RA) strategies with tuple values *n, m* = $\{5, 5\}$ and *n, m* = $\{7, 9\}$, the two best results in [7] and [5], respectively. We found RA with values *n, m* = $\{7, 9\}$ gave superior segmentation results and have used these values in the results reported subsequently; denoted as $RA(7, 9)$. Out of all transformations in [5], we removed non-linear transformations which are already stochastic in nature; these are elastic and grid distortions as they produce random magnitude transformations for both, US images and their corresponding GT saliency maps. We retain speckle as it only affects the quality of US frames keeping the GT saliency maps unchanged.

The sequence data samples are augmented at random whereas the type of augmentation transformation is shared between each image in a sequence. Such a procedure is crucial to preserve the temporal information of US frames and the sonographer gaze pattern.

2.3. VSP Network Architecture

We experimented with a VSP architecture that takes two inputs. An overview of the network is shown in Fig. 1. The cLSTMU-Net has two modules. U-Net is an encoder-decoder network with skip connections, and cLSTM is a recurrent network that manages a series of data that are chronologically ordered. Our input consists of two parts, one is a single US frame and the other is a sequence of frames preceding and including the US frame in question. The first input is fed into a spatial U-Net, and the second input becomes part of the cLSTM module. US video frames and their corresponding GT saliency maps are sampled from a video clip as described in Section 2.1. The encoder layer structure is mimicked by the temporal input. We adopt a time distributed layer as it shares the weights between the images in a sequence, training them in parallel. This way, we avoid excessive training time and different detection of features that are not linked between the images in an input sequence. We use cLSTM to extract different image features and preserve their chronological order. The output of the cLSTM is fed into a decoder where after a transpose convolution, it is concatenated with skip connection from the corresponding encoder layer. The rest of the network follows the structure of the decoder with a softmax activation function applied to the final output layer.

Saliency Map Prediction: The dataset $D = \{(\mathbf{X}^{(t)}, G^{(t)})\}_{t=1}^{N^x}$ consists of N_x pairs of video frames and gaze point sets. Given an image and a gaze point set $(\mathbf{X}, G)\epsilon D$, we generate a visual saliency map $\mathbf{S}\epsilon[0,1]^{H_D \times W_D}$, where $S_{i,j}$ is the probability that pixel $X_{i,j}$ is fixated upon. The saliency map is then used as the target for the predicted probability map $\hat{\mathbf{S}}$. Around the gaze points in G, $\mathbf{S}$ is a sum of Gaussians normalized such that $\sum_{i,j} S_{i,j} = 1$. The saliency map yields the training target $\mathbf{S}^*\epsilon[0,1]^{H_D \times W_D}$. Finally, the training loss is computed via the Kullback-Leibler divergence (KLD) between the predicted and true distribution:

$$\begin{aligned} \mathbf{L}_s(\mathbf{S}^*, \mathbf{S}) &= D_{KL}(\mathbf{S}^* \parallel \mathbf{S}) \\ &= \sum_{i,j} S^*_{i,j} \cdot (log(S^*_{i,j}) - (log(S_{i,j})) \end{aligned} \quad (1)$$

3. EXPERIMENTS AND RESULTS

3.1. Network Implementation Details

A VSP architecture was trained from scratch via Adam optimization with a momentum of 0.01 and a learning rate of

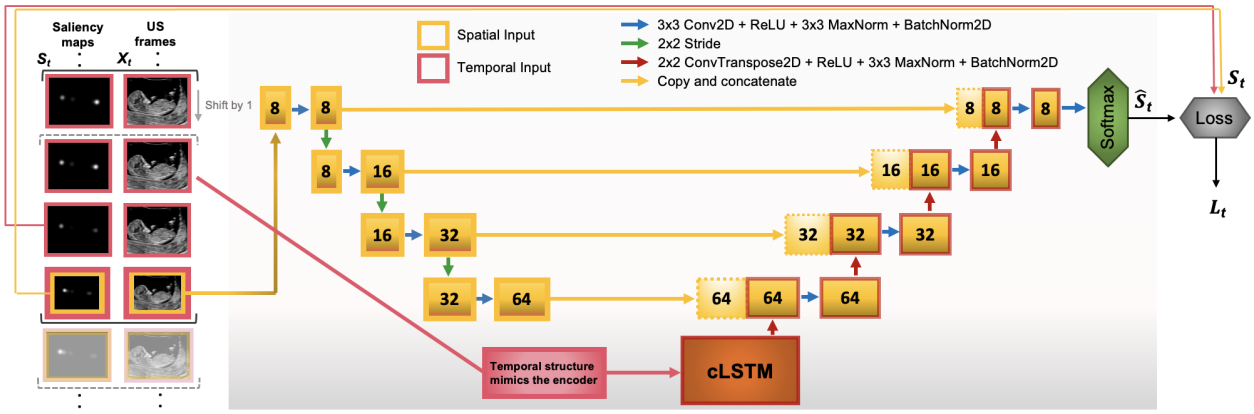

Fig. 1: Overview of the proposed architecture for video saliency prediction.

Table 1: Quantitative results of visual saliency prediction. Sequence length is displayed at the top. The best performing model is marked in **bold**.

	RA(7,9)	RA(3,9)	Mix.RA(3,7)	cLSTMU-Net(7,9)			
Seq. length	1	1	1	3	**6**	9	10
KLD	2.16	2.27	2.28	2.22	**2.08**	2.11	2.22
SIM	0.25	0.27	0.26	0.23	**0.28**	0.26	0.25
NSS	4.19	4.21	4.34	4.16	**4.53**	4.41	4.06
CC	0.39	0.38	0.38	0.38	**0.42**	0.40	0.37

0.0001 with early stopping. The batch size was set to 16 across all models. The models were implemented in Tensorflow 2.1 on a Nvidia GTX 2060 Ti. Image manipulations were performed with Pillow 7.1.2 and OpenCV 3.4.9 libraries.

3.2. Quantitative Results

Table 1 reports the average test scores for 3 best performed spatial models from [5] and the spatio-temporal cLSTMU-Net across different video clip lengths. Models are evaluated using Kullback-Leibler divergence (KL), normalized scanpath saliency (NSS), Pearson's correlation coefficient (CC) and Similarity metric (SIM) [8]. Particularly, the 3 models include $RA(7,9)$, $RA(3,9)$ and $Mix.RA(3,7)$. The results show that the spatio-temporal cLSTMU-Net $RA(7,9)$ using a video clip of 6 frames outperforms all models on all metrics.

3.3. Representative Examples

Fig. 2 shows exemplary test results of the VSP model and the comparative spatial-only models. The spatio-temporal cLSTMU-Net network with $RA(7,9)$ using a video clip of 6 frames better localizes the nasal bone and rump than all the other spatial-only models. Models are compared to the GT gaze distribution (yellow). Since the training and validation data were divided scan-wise fulfilling the case for 90 consecutive frames, the frames are unseen by the network.

From the GT frames, the sonographer primarily focuses on the nasal bone, nasal tip and checks the rump for guidance during scanning. In the first three exemplary frames, 3 spatial-only models fail to predict the sonographer gaze. Our

Table 2: Ablation study on the placement of temporal information in regard to the single frame. The best placement of the temporal information is marked in **bold**.

	Temporal Information		
	Before	Before & After	After
KLD	**2.08**	2.14	2.15
SIM	**0.28**	0.23	0.26
NSS	**4.53**	4.05	4.33
CC	**0.42**	0.38	0.39

cLSTMU-Net model shows an almost identical saliency map prediction of the nasal bone and gives low probability values to rump (white). The GT fixations are on the nasal bone with extremely low probability assigned to the rump at frame zero.

The latter 3 frames show cLSTMU-Net steady adjustment of the saliency prediction from the maxima around the palate to the nasal bone, which is the correct saliency map location. The less salient rump is correctly predicted in the last 2 frames, with slight misalignment towards the buttocks. The alternative models focus on the bottom end of the palate. Only $RA(3,9)$ over-estimates the gaze of the sonographer looking at rump; the other models fail to localize the structure.

3.4. Ablation Study

We evaluated the impact that placement of temporal information with respect to a single frame has on saliency prediction. We performed an ablation study with results reported in Table 2. A video clip of length 6 is used as an example. We observe that the addition of temporal information before the predicted frame adds the most value to saliency prediction.

4. DISCUSSION AND CONCLUSION

We presented a VSP network for first trimester US images. The results show that the spatio-temporal cLSTMU-Net network architecture with $RA(7,9)$ using a video clip of 6 frames outperformed all other spatial-only models. This can be credited to the training of cLSTMU-Net in the spatio-temporal domain which allows gradients to back-propagate

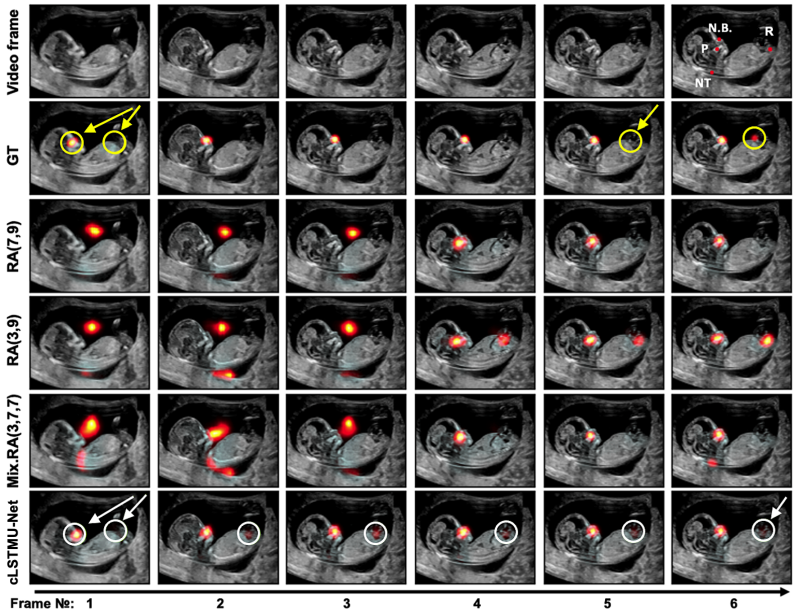

Fig. 2: Six frames from an exemplary search sequence. The rows show the input frames, the ground truth saliency annotations, 3 spatial-only saliency models with the best metric results from [5] against video saliency predictions of cLSTMU-Net, respectively. The relevant anatomical structures denoted in the last input frame (top right) include palate (P), nasal bone (N.B.), rump (R) and nuchal translucency (NT). The ground truth is circled in yellow and cLSTMU-Net predictions are circled in white.

with respect to time and space which aids training. In contrast, the gradients of spatial-only models are solely back-propagated with respect to each frame (i.e. only space).

The available computer memory could handle a video clip of 10 frames. The best performing spatio-temporal model used a video clip of 6 frames. For our dataset, we discovered that 6 consecutive frames would account for good gaze variation with more frames adding little useful information. After performing an ablation study we found that adding temporal information before the frame that the saliency map is predicted for gave the best model performance.

Quantitatively, the KLD metric is highly penalized if any GT fixation locations are missed, for instance the nasal bone or rump in Fig. 2. In comparison to the three best performed models from [5], the cLSTMU-Net led to a decrease in KLD score of 0.08. SIM and CC metrics improved by 0.01 and 0.03, respectively. CC penalizes false negatives and SIM penalizes predictions that fail to account for all the GT density. NSS is the only location-based metric, it benefits from the temporal information with score increase of 0.19. The NSS metric is sensitive to false positives which is seen in Fig. 2 showing no false saliency prediction on the NT.

In conclusion, the proposed cLSTMU-Net model is able to better track changes in sonographer gaze compared to previous methods [5]. This may form the basis of a useful automatic guidance mechanism for real-time first trimester US scanning where the saliency predictions direct sonographer gaze to important anatomy. This may be investigated in our future work.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051) and the ERC ethics committee.

6. REFERENCES

[1] A Salvadoret al., "Recurrent neural networks for semantic instance segmentation," *arXiv:1712.00617*, 2017.

[2] F Xu et al., "LSTM multi-modal U-Net for brain tumor segmentation," in *ICIVC*. IEEE, 2019, pp. 236–240.

[3] X Wu et al., "SalSAC: A video saliency prediction model with shuffled attentions and correlation-based ConvLSTM," in *AAAI*, 2020, vol. 34, pp. 12410–12417.

[4] Cai et al., "Spatio-temporal visual attention modelling of standard biometry plane-finding navigation," *MIA*, vol. 65, pp. 101762, 2020.

[5] E Savochkina et al., "First trimester gaze pattern estimation using stochastic augmentation policy search for single frame saliency prediction," in *MIUA*. Springer, LNCS, 2021, pp. 361–374.

[6] Ronneberger et al., "U-net: convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[7] Lee et al., "Principled ultrasound data augmentation for classification of standard planes," in *IPMI*. Springer, 2021, pp. 729–741.

[8] Z Bylinskii et al., "What do different evaluation metrics tell us about saliency models?," *T-PAMI*, vol. 41, no. 3, pp. 740–757, 2018.