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

Richard Droste[1], Yifan Cai[1], Harshita Sharma[1], Pierre Chatelain[1], Aris T. Papageorghiou[2], J. Alison Noble[1]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
richard.droste@eng.ox.ac.uk
[2] Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK

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

Keywords: Fetal ultrasound · Video saliency prediction · Gaze tracking· Convolutional neural networks

1 Introduction

Recently, it has been demonstrated that sonographer gaze tracking can aid standard plane detection in fetal ultrasound (US) imaging. Cai et al. [8] proposed the SonoEyeNet model for abdominal circumference plain (ACP) detection. Recorded gaze tracking heat maps—hereafter referred to as *saliency maps*—are used as attention on feature maps which are extracted with a fine-tuned SonoNet model [4]. Next, Cai et al. [9] proposed the Multi-task SonoEyeNet. Instead of relying on gaze data as an input, an attention module learns to predict saliency maps so that no gaze tracking data is required for inference. Recently,

Droste et al. [14] demonstrated that a saliency predictor trained entirely without manual annotations can be transferred to perform standard plane detection in routine clinical videos. These models perform standard plane detection and saliency prediction on single-frames only. However, ultrasound data and human eye movements are inherently spatio-temporal signals.

In this work we aim at improving ultrasound saliency prediction through spatio-temporal modeling, i.e. video saliency prediction. Therefore, we aim to bridge the gap between existing spatio-temporal models which do not leverage gaze information, e.g. for fetal cardiology [18,16] or US video partitioning [22], and models like SonoEyeNet that do not utilize temporal information. Chaabouni et al. [10] present an early convolutional neural network (CNN) based approach for video saliency prediction, adding optical flow as an additional input channel to a single-frame CNN. Bak et al. [2] propose to include optical flow via a two-stream architecture [23]. Bazzani et al. [5] achieve much larger temporal depth with a recurrent mixture density network by aggregating feature vectors with a long short-term memory (LSTM) model. Wang et al. [27] recently proposed a large video saliency benchmark (DHF1K) and show that existing video saliency predictors do not outperform the best single-frame saliency predictors. In contrast, Wang et al. achieve state-of-the-art on their benchmark with an architecture consisting of a CNN encoder and a convolutional LSTM decoder.

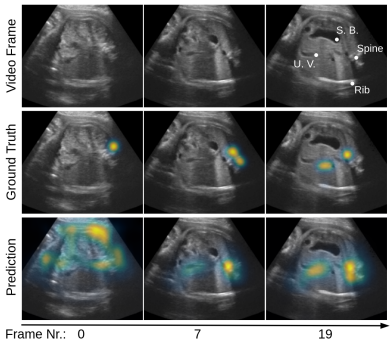

Fig. 1: BDS-Net predicted saliency maps. The key structures are marked. U. V. denotes umbilical vein and S. B. stomach bubble.

The above mentioned spatio-temporal models predict the saliency map of each video frame based on aggregated information of the previous frames. However, research in cognitive science suggests that human perception is not just reacting to past and present stimuli. Clark [13] argues that the brain is a 'prediction machine' that strives to minimize the prediction error between expectations versus sensory inputs. We transfer this insight to the problem of predicting sonographer visual saliency on US videos. Since the sonographer's expectations about future visual stimuli are unknown, we use future video frames as a proxy thereof, and ask the question: *To what extent are future video frames predictive for visual saliency?* Song et al. [25] recently proposed a bidirectional model for video salient object detection, which is a related application but aims at detecting and segmenting the most salient object in a scene rather than predicting the actual distribution of gaze points. Here, we propose an architecture combining a CNN and a temporally bidirectional recurrent neural network, BDS-Net, that predicts the visual saliency of each frame based on information of the entire video sequence, and compare the performance of the BDS-Net to an equivalent one-directional and a purely spatial model.

Contributions. The contributions of this study are three-fold: (1) To the best of our knowledge, this is the first study to propose a temporally bidirectional model for video saliency prediction, both in medical imaging and in computer vision more generally. Since the model considers the entire video sequence for saliency prediction of each frame, this approach is fundamentally different from previous models that only consider past and present frames. (2) We demonstrate that it is possible to train an effective video saliency predictor with few more than one hundred sequences, despite high inter-sequence variance. We achieve high data-efficiency by employing effective transfer learning and regularization techniques, and by reducing model complexity where possible, e.g. using a gated-recurrent-unit recurrent convolutional network (GRU-RCN) instead of a convolutional LSTM. (3) We demonstrate that the trained US video saliency predictor has learned meaningful aspects of sonographers' cognition in selecting the ACP. Therefore, we expect the model to be beneficial as part of architectures such as the Multi-task SonoEyeNet [9].

2 BDS-Net

The BDS-Net architecture consists of a *truncated SonoNet-64* model as frame-wise encoder, *adaptation I* that extracts the task-relevant features, a *bidirectional GRU-RCN* to aggregate the features temporally, and *adaptation II* to assemble the saliency map, followed by a *softmax* function (Fig. 2). In the following, we will use the vector notation $\mathbf{v}_t = [v_0^t, v_1^t, ..., v_n^t]^\top$.

Truncated SonoNet-64. The SonoNet model was recently proposed for US standard plane detection [4]. It is derived from the VGG-16 architecture [24], removing the final max-pooling and replacing the fully-connected layers with adaptation layers of 1×1-convolutions followed by global average pooling. Also, batch normalization [19] is added to each convolutional layer. The authors present three model variants with different numbers of convolutional kernels and train them on over 27 thousand US standard plane images. For this work, we use the largest variant, SonoNet-64, which was shown to achieve the highest overall precision. To use the model as a feature extractor, we remove the adaptation layers since they are classification-task specific. Further, we truncate the model by discarding the final three 3×3-convolutional layers to obtain lower-level features. We use the remaining 10 convolutional layers as frame-wise encoder of the BDS-Net. Since the SonoNet training data is substantially larger than the data available for this work, we use the model with fixed pre-trained weights.

Bidirectional GRU-RCN. We propose a bidirectional gated-recurrent-unit recurrent convolutional network (GRU-RCN) as the spatio-temporal decoder of the network. GRU networks [11] mitigate the exploding/vanishing gradient problem of a regular RNN similarly to LSTM networks [17] by updating the hidden state through element-wise additive and multiplicative gates instead of matrix multiplications. Compared to LSTM networks, however, they yield faster training

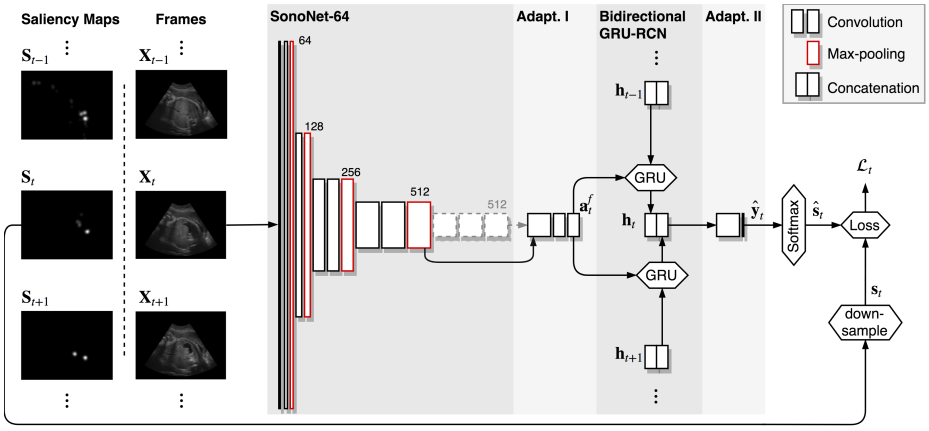

Fig. 2: Schema of the BDS-Net architecture and training procedure. For better readability, activation and normalization layers are not explicitly shown. The dashed part of the SonoNet-64 is only used for an ablation study.

convergence and higher accuracy on tasks like video captioning, despite reduced complexity and fewer learned parameters [12]. While the standard GRU operates on 1D feature vectors, the GRU-RCN [3] is a straightforward extension for stacked 2D feature maps, replacing matrix products with convolutions. This modification vastly reduces the number of parameters compared to a fully-connected GRU and preserves the spatial feature topology. The bidirectional GRU-RCN is constructed from two separate GRU-RCN instances that propagate their hidden states forwards and backwards through time, respectively.

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

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

$$\widetilde{\mathbf{h}}_t^{\rightarrow} = \tanh\left(\mathbf{W}_h^{\rightarrow} * [\mathbf{r}_t^{\rightarrow} \circ \mathbf{h}_{t-1}^{\rightarrow} \, | \, \mathbf{a}_t^f] + \mathbf{b}_h^{\rightarrow}\right) \, , \tag{2}$$

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

$$\mathbf{z}_t^{\rightarrow} = \sigma(\mathbf{W}_z^{\rightarrow} * [\mathbf{h}_{t-1}^{\rightarrow} | \, \mathbf{a}_t^f] + \mathbf{b}_z^{\rightarrow}) \tag{3}$$

$$\mathbf{h}_t^{\rightarrow} = (1 - \mathbf{z}_t^{\rightarrow}) \circ \mathbf{h}_{t-1}^{\rightarrow} + \mathbf{z}_t^{\rightarrow} \circ \widetilde{\mathbf{h}}_t^{\rightarrow} \, . \tag{4}$$

The backward GRU-RCN activation $\mathbf{h}_t^{\leftarrow}$ is computed equivalently, using the activation $\mathbf{h}_{t+1}^{\leftarrow}$ of time $t+1$ as the previous activation. Finally, the activations of the forward and backward RCNs are concatenated as $\mathbf{h}_t = [\mathbf{h}_t^{\rightarrow} | \, \mathbf{h}_t^{\leftarrow}]$.

We normalize the activations and gates throughout the GRU with layer normalization [1]. We found no increase in performance by replacing the layer normalization with instance normalization [26] in the GRU. Batch normalization is not compatible since we set the batch size to one due to the high variability of the sequence lengths. Further, we observed no improvement through dropout on the GRU inputs [29] or variational recurrent dropout [15]. To avoid over-fitting, the kernel size of $\mathbf{W}_r$ and $\mathbf{W}_z$ are set to size 1×1. Only the kernels of $\mathbf{W}_h$ for computing the candidate activation are set to size 3×3.

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

Loss function. At each time step $t \in \{0, 1, ..., T\}$ and output pixel $i \in \{0, 1, ..., P\}$, we obtain the predicted saliency $\hat{s}_i^t$ from the activations $\hat{y}_i^t$ of the final adaptation layer via the softmax function $\hat{s}_i^t = e^{\hat{y}_i^t} \, (\sum_{j=0}^{P} e^{\hat{y}_j^t})^{-1}$. Consequently, we implicitly treat the saliency maps as generalized Bernoulli distributions of fixations over the pixels of each frame [20]. We compute the loss as the sum of the Kullback-Leibler Divergences between the predicted distributions[3] and the downscaled ground truth distributions $\mathbf{s}_t$ as $\mathcal{L} = \sum_{t=0}^{T} \sum_{i=0}^{P} s_i^t \cdot (log(s_i^t) - log(\hat{s}_i^t))$

3 Experiments

3.1 Data

The gaze data for this study had previously been recorded based on 33 fetal US videos, which were acquired according to a manual US sweep protocol, moving the probe from the bottom to the top of pregnant women's abdomen. Table 1 summarizes the data preparation procedure. *(1) Discarding of irrelevant frames:* From each of the 33 full sweeps an *ACP sweep* was extracted by discarding the frames which do not show the fetal abdomen. *(2) ACP selection by sonographers:* Each ACP sweep was presented to eight sonographers independently with the task of selecting the abdominal circumference plane (ACP). The sonographers were able to scroll through the frames using a keyboard until deciding on one ACP frame. The gaze of the sonographers was recorded at 30 Hz using an eye

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

Table 1: Data preparation procedure.

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

tracker (The EyeTribe) placed beneath the screen. In addition, the current sweep frame was registered for each gaze point. This yielded $33 \times 8 = 264$ sequences of gaze point-sweep frame pairs, which we will refer to as *ACP search sequences*. The ACP search sequences represent the way the sonographers moved through the frames to find the ACP. Therefore, they are a potentially useful approximation of freehand US video sequences where the sonographers find the ACP by moving the probe. The sequences were manually inspected and recordings with low-quality gaze data or miscalibration were discarded, leaving 116 sequences for further processing. *(3) Gaze point aggregation:* Since we want our model to learn the search strategy of sonographers from the first glance until the final ACP plane decision, we want to train the model on entire ACP search sequences. At 30 Hz, however, the sequences are too long (at least several hundred frames) and contain high redundancy among consecutive frames. Therefore, the gaze points were aggregated over intervals of 1000 ms at every 8^{th} gaze sampling time, reducing the sampling rate from 30 Hz to 3.75 Hz. Gaze points outside the US image fan were discarded. A gaze map was computed for each aggregated set of gaze points by setting each pixel value to its corresponding number of gaze points. The frames were re-sampled at the same sampling times. The resulting sequences of gaze map-sweep frame pairs are of length 13 to 147 (avg. 33.6). *(4) Gaze maps filtering:* The saliency maps were computed by smoothing the gaze maps with a Gaussian kernel. The resulting final sequences of saliency map-sweep frame pairs are henceforth referred to as *saliency sequences*. The ACP sweeps were divided into 30 sweeps for training and 3 sweeps for validation with five-fold cross-validation.

3.2 Implementation Details

Preprocessing. The frames are preprocessed following Baumgartner et al. [4]. Data augmentation is performed by random rotation with an angle uniformly sampled from $[-25, 25]$ degrees and random horizontal flipping. Scale augmentation is omitted since it results in cropping of parts of the fetal abdomen. Next, the frames are normalized to zero-mean and unit-variance, multiplied by 255 and resized to 288×224 px. For the calculation of the loss, the ground truth saliency

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

Training. The model is trained over 140 epochs via stochastic gradient descent (SGD) with Nesterov momentum of 0.9 and initial learning rate 0.004. In accordance with Keskar et al. [21], we find that SGD yields better generalization to the validation set compared to ADAM. The learning rate is decayed by a factor of 0.5 each time the validation loss stagnates. The batch size is one to allow for varying sequence lengths. Sequences longer than 60 frames are truncated. The model is regularized via weight decay of $1 \cdot 10^{-6}$, dropout with rate 0.2 before the second and the last adaptation layers, as well as clipping the gradients outside the interval $[-5, 5]$. The z-gate bias is initialized to 1 to stabilize training by learning the spatial features first.

Comparative models. Two comparative models are implemented: A one-directional GRU-RCN model and a purely spatial, single-frame model. The one-directional model is constructed by removing the backward GRU-RCN module from the BDS-Net. All other architectural and training parameters are identical. For the spatial model, the bidirectional GRU-RCN is simply replaced by an additional convolutional layer of dimension 3×3×128×128. Moreover, the layer normalization modules are removed and batch normalization is added to each layer. Training is performed on batches of 16 randomly selected frames and dropout with rate 0.5 is added to all layers. The initial learning rate is increased to 0.01 and no weight-decay is performed. Furthermore, we perform an ablation study to examine the effect of using the full SonoNet-64 (only adaptation layers removed) or the truncated SonoNet-32 models instead of the truncated SonoNet-64. The results are presented in subsection 4.3.

3.3 Evaluation Metrics

We evaluate the models on the metrics of the MIT Saliency Benchmark [7]. For this, we ported the MATLAB code published by the authors[4] to Python. We consider the fixation point (gaze map) based metrics Normalized Scanpath Saliency (NSS) and Area Under the ROC Curve by Judd (AUC-J), as well as the distribution (saliency map) based metrics Kullback-Leibler divergence (KLD), Linear Correlation Coefficient (CC) and Similarity (SIM). AUC-J, KLD and SIM are more sensitive to false negatives than to false positives, while NSS and CC treat them symmetrically [6]. To compute the average scores on each validation set, the scores are first averaged across time for each sequence and then across sequences. Thus, shorter and longer sequences weigh equally in the average. The differences between the respective cross-validated model scores are tested for statistical significance with the Wilcoxon test.

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

Table 2: Cross-validation scores (mean ± standard deviation) of the BDS-Net and the spatial and one-directional models. The best scores are marked in bold. The superscripts * and † denote an improvement with $p < 0.05$ over the spatial and one-directional models, respectively.

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

4 Results

4.1 Quantitative Results

Table 2 shows the validation scores of the BDS-Net and comparative models. The BDS-Net receives the best scores on all metrics except SIM. Moreover, both spatio-temporal models perform better on average than the spatial model on all metrics except SIM. For SIM, the BDS-Net scores better than the one-directional model but is on par with the spatial model.

4.2 Representative Examples

Figures 1 and 3 show examples of the predictions of the BDS-Net model and the comparative models on validation data. Since training and validation data were divided scan-wise, the frames are entirely unseen by the networks. Moreover, the networks are agnostic as to which sonographer is observing the scan.

Fig. 1 shows input frames, ground truth saliency maps and BDS-Net predictions for three representative frames of one exemplary sequence. At frame zero, the prediction is highly uncertain. Areas throughout the middle and the upper boundary of the abdomen are predicted as fixation candidates. The highest probability is assigned to the area around the upper rib, which is not well visible. The ground truth fixation is between the spine and the upper rib. At frame seven, the BDS-Net assigns high saliency values to the spine and lower values to the umbilical vein. The ground truth fixations are at the spine. At frame nineteen, near the end of the sequence, the network assigns approximately equal probabilities to umbilical vein and spine. The ground truth fixations are indeed on umbilical vein and spine.

Fig. 3 shows a more detailed example of BDS-Net predictions for five representative frames of another exemplary sequence. Additionally, the predictions of the spatial and one-directional models are shown. At frame zero, both spatio-temporal models predict uncertain saliency maps with a spread-out peak between spine and upper rib as denoted with *a)* in the figure. The spatial model, which does not have information about the position of the frame in the sequence, predicts saliency at the spine with high certainty. The ground truth fixations are both at spine and the upper rib. Over the next frames, the recurrent models predict

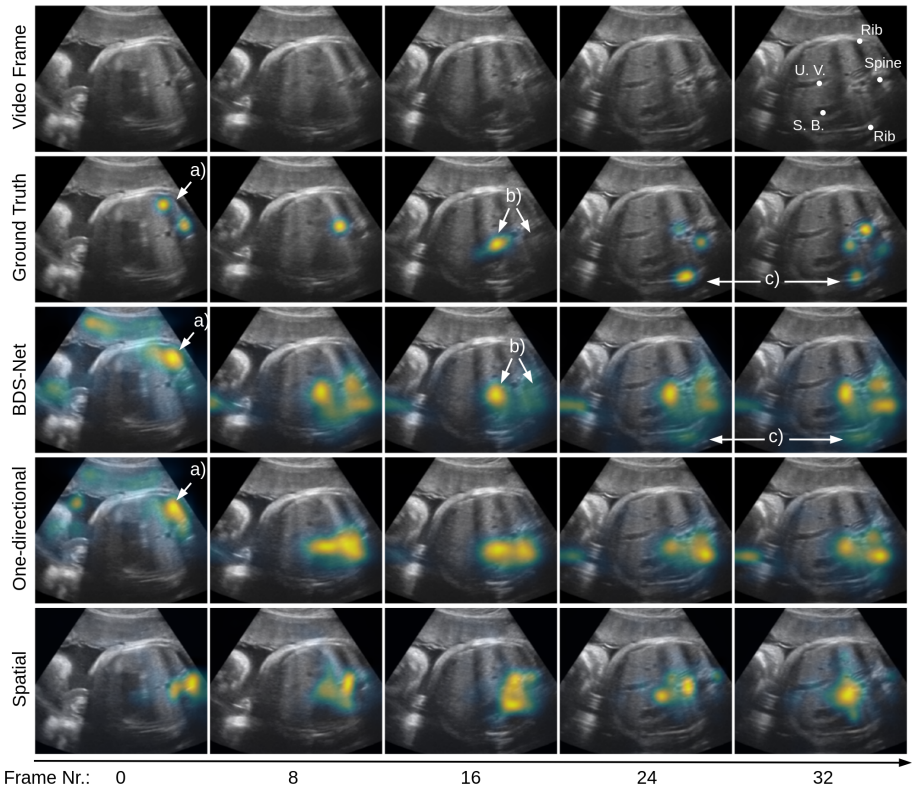

Fig. 3: Five frames from an exemplary ACP search sequence. The rows show the input frames, the ground truth saliency annotations, and the saliency predictions of the BDS-Net and the spatial and one-directional models, respectively. The relevant anatomical structures are denoted in the last input frame (top right).

temporally smooth saliency maps with slightly varying maxima around the center of the abdomen and the spine. The maxima of the spatial model fall into the same regions but the maps are less temporally smooth. Two key advantages of the BDS-Net predictions are denoted with *b)* and *c)*. At frame sixteen, in the middle of the sequence, the sonographer fixates the center of the abdomen. The spatial model predicts fixations around the spine and the one-directional model predicts fixations at either the spine or the center. Only the bidirectional model, which has information about both the previous and the subsequent frames, correctly predicts the fixation at the center and omits the spine, as denoted by *b)*. On frames 24 and 32, towards the end of the search sequence, the sonographer fixates on the spine, the center of the abdomen and the lower rib, which is not well visible in these frames. Only the BDS-Net correctly assigns probability of fixation to the lower rib, indicated by *c)*.

Table 3: Scores for the ablation study of the feature extractor on one randomly chosen validation set.

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

4.3 Ablation Study

The quantitative results for the ablation study of the feature extractor are shown in Table 3. The ranking of the models is consistent across all five metrics: The full SonoNet-64 with higher-level features performs least favorable, the smaller truncated SonoNet-32 ranks second and the truncated SonoNet-64 performs best.

5 Discussion

The BDS-Net outperforms both comparative models on the AUC-J and NSS metrics, which are the default metrics of the MIT Saliency Benchmark [7]. Moreover, the one-directional model outperforms or matches the score of the spatial model on those metrics, despite the fact that training the spatial model is arguably easier for several reasons. First, gradient steps can be computed for each of the 3901 saliency maps per epoch separately. For the recurrent models, in contrast, gradient steps can only be computed for the 104 sequences per epoch. Second, the gradients are noisier for recurrent models in general. We mitigate this problem through gradient clipping, but it is an ad-hoc solution, and it does not resolve vanishing gradients. Finally, batch normalization is applied in the spatial model. For the recurrent models, since we set the batch size to one to account for the varying sequence lengths, we revert to layer normalization, which is known to stabilize training less for most tasks [28]. The fact that the recurrent models perform better despite the more difficult training conditions is a strong indication that spatial information (the current frame) alone is not sufficient to predict US video saliency accurately. This is in accordance with the results of Wang et al. [27] who have shown for natural videos that a recurrent architecture can outperform sophisticated single-frame saliency predictors.

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

It is difficult to say how much room for improvement remains for the given dataset. Naturally, there is a certain inter-observer variability in the ACP search

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

Nonetheless, despite the missing reference values for the quantitative evaluations, the qualitative analyses have shown that the BDS-Net has learned meaningful spatio-temporal patterns in the sonographers' search strategies. The analyses of Cai et al. [9] have shown that similar learned experience can significantly improve US standard plane detection on single frames. We expect that our video saliency predictor can further improve the performance of such models.

6 Conclusion and Outlook

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

Acknowledgements. This work is supported by the ERC (ERC-ADG-2015 694581, project PULSE) and the EPSRC (EP/R013853/1 and EP/M013774/1). AP is funded by the NIHR Oxford Biomedical Research Centre.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. NIPS - Deep Learning Symposium (2016)
2. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction. IEEE Trans. Multimed. **20**(7), 1688–1698 (2018)
3. Ballas, N., Yao, L., Pal, C., Courville, A.: Delving Deeper into Convolutional Networks for Learning Video Representations. ICLR (2016)
4. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. IEEE Trans. Med. Imag. **36**(11), 2204–2215 (2017)
5. Bazzani, L., Larochelle, H., Torresani, L.: Recurrent Mixture Density Network for Spatiotemporal Visual Attention. ICLR (2017)
6. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What Do Different Evaluation Metrics Tell Us About Saliency Models? IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 740–757 (2019)

7. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: MIT Saliency Benchmark. http://saliency.mit.edu/
8. Cai, Y., Sharma, H., Chatelain, P., Noble, J.A.: SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking. ISBI (2018)
9. Cai, Y., Sharma, H., Chatelain, P., Noble, J.A.: Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps. In: MICCAI (2018)
10. Chaabouni, S., Benois-pineau, J., Hadar, O.: Deep Learning for Saliency Prediction in Natural Video. arXiv:1604.08010 (2016)
11. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP (2014)
12. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. NIPS (2014)
13. Clark, A.: Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences **36**(03), 181–204 (2013)
14. Droste, R., Cai, Y., Sharma, H., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention. Accepted at IPMI (2019)
15. Gal, Y., Ghahramani, Z.: A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. NIPS (2016)
16. Gao, Y., Noble, J.A.: Detection and Characterization of the Fetal Heartbeat in Free-hand Ultrasound Sweeps with Weakly-supervised Two-streams Convolutional Networks. MICCAI (2017)
17. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997)
18. Huang, W., Bridge, C.P., Noble, J.A., Zisserman, A.: Temporal HeartNet: Towards Human-Level Automatic Analysis of Fetal Cardiac Screening Video. MICCAI (2017)
19. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ICML (2015)
20. Jetley, S., Murray, N., Vig, E.: End-to-End Saliency Mapping via Probability Distribution Prediction. CVPR (2016)
21. Keskar, N.S., Socher, R.: Improving Generalization Performance by Switching from Adam to SGD. arXiv:1712.07628 (2017)
22. Sharma, H., Droste, R., Chatelain, P., Drukker, L., Papageorghiou, A., Noble, J.A.: Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans. Accepted at IEEE ISBI 2019 (2019)
23. Simonyan, K., Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. NIPS (2014)
24. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR (2015)
25. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection. ECCV (2018)
26. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance Normalization: The Missing Ingredient for Fast Stylization. arxiv:1607.08022 (2016)
27. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting Video Saliency: A Large-scale Benchmark and a New Model. CVPR (2018)
28. Wu, Y., He, K.: Group Normalization. ECCV (2018)
29. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent Neural Network Regularization. arXiv:1409.2329 (2014)