TOWARDS STANDARD PLANE PREDICTION OF FETAL HEAD ULTRASOUND WITH DOMAIN ADAPTION

Qianhui Men[1] *He Zhao*[1] *Lior Drukker*[2,3] *Aris T. Papageorghiou*[2] *J. Alison Noble*[1]

[1] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
[2] Nuffield Department of Women's & Reproductive Health, University of Oxford, UK
[3] Department of Obstetrics and Gynecology, Tel-Aviv University, Israel

2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230542

ABSTRACT

Fetal Standard Plane (SP) acquisition is a key step in ultrasound based assessment of fetal health. The task detects an ultrasound (US) image with predefined anatomy. However, it requires skill to acquire a good SP in practice, and trainees and occasional users of ultrasound devices can find this challenging. In this work, we consider the task of automatically predicting the fetal head SP from the video approaching the SP. We adopt a domain transfer learning approach that maps the encoded spatial and temporal features of video in the source domain to the spatial representations of the desired SP image in the target domain, together with adversarial training to preserve the quality of the resulting image. Experimental results show that the predicted head plane is plausible and consistent with the anatomical features expected in a real SP. The proposed approach is motivated to support non-experts to find and analyse a trans-ventricular (TV) plane but could also be generalized to other planes, trimesters, and ultrasound imaging tasks for which standard planes are defined.

***Index Terms*—** Fetal ultrasound, Adversarial learning, Domain adaption, Image synthesis

1. INTRODUCTION

Standard plane (SP) acquisition is a routine clinical examination task during obstetric ultrasound (US) scanning. In this study, we consider the trans-ventricular (TV) plane which is the plane used for head circumference (HC) biometry and to assess fetal brain development. The TV standard plane is obtained by live B-mode scanning stopped at a *cine-buffer* frame with particular anatomical structures that are clearly presented. However, the quality of the retrieved US image is highly dependent on the experience of the sonographer. An inadequate visualisation of any key anatomical structures of the desired plane will typically require a second image capture that causes burden for both the pregnant mother and sonographer, and failure to detect any abnormal structures may lead to misdiagnosis.

To support medical education of trainee sonographers, previous related work has mainly focused on simulating random anatomical views of US from scratch. One branch of work in this field [1, 2] target at searching numerical solutions of the biological tissues in wave space or using physic ray-tracing approaches, which are computationally expensive to use. Another branch has proposed deep learning-based frameworks to simulate ultrasound images from phantoms [3] or to enhance synthetic images [4]. Whilst promising results have been achieved, such models are suitable for training simulation but challenging to deploy in real-world clinical scanning with large anatomical variations. With this observation, [5] proposed to generate anatomy based on human annotations of clinical US and control over variability of the generated fetal head. Liang et al. [6] simulated US images from manually-labeled segmentation maps. There are also a few attempts to predict or synthesize medical images from videos of other modalities. For example, [7] used a spatio-temporal prediction network to outline the lesion area in 4D CT brain perfusion imaging. However, that method is not directly applicable to US prediction due to the large potential changes in the anatomical planes.

In this paper, we propose a Domain-Adaptive SP Generator (DASPG) that learns to predict the SP image with the underlying anatomical structures inferred from the spatial and temporal features of the video searching for the SP (defined as SP search video). The challenges faced in designing a solution for this task are: 1) compared to a SP image, SP search video usually contains noticeable artefacts such as motion blur and distortion; 2) many frames in the video are not related to the final SP. To address these obstacles, the key step of DASPG is to translate the raw video features to a clean, standard biometric plane with a domain adaption (DA) module. Domain adaption has been broadly exploited to transform representations across different modalities, such as image-to-video of daily activities [8] and magnetic resonance (MR) imaging to computed tomography (CT) in medical resources [9]. Here, we pose the problem of predicting the SP image from video as a transfer learning task. Specifically, we regard the video during scanning as the source domain and its resulting SP image as the target domain, and apply domain adaption to align between these two imaging representation domains. To avoid an averaged solution, we separately model the spatial and temporal features of the video input with a U-Net [10] and a temporal convolutional network (TCN) [11], respectively, that keeps the spatial properties of the US image appearance while propagating the temporal dynamics. Furthermore, we add a Generative Adversarial Network (GAN) to our model to create realistic ultrasound textures in the predicted image.

The paper contributions are three-fold: 1) We propose the first model to predict an ultrasound SP from a search video starting at a random position. 2) We show that domain adaption can effectively convert a raw search video to the SP image with clear anatomy. 3) The feasibility of our approach is demonstrated on real-world fetal brain ultrasound of both *Anomaly* and *Growth* scans. The predicted plane is realistic with required anatomical structures that can be used as a target image to guide the SP detection.

2. METHODOLOGY

The proposed DASPG model predicts the corresponding standard anatomical plane image when given a search video of TV. Different from identifying an observed SP [12, 13], our aim is to generate a standard view to assist inexperienced sonographers. The overall architecture is given in Fig. 1. Let $X = \{x_t\}_{t=t_i}^{t_n}$ denotes the input US video sequence (the search video), and y its SP image frame at time T. The proposed generator G learns a mapping $X \to y$ that consists of a stepwise spatial encoder E_S and a temporal extractor E_T to separately model the spatial and temporal feature of the search video, a domain adaption module DA to transfer the observed scanning knowledge to the targeted SP representation, and a spatial decoder D_S to synthesize the SP image.

2.1. Domain-Adaptive Standard Plane Generator

Since most of the interpretable structures are within the fetal skull, the US image is initially transformed by a pretrained Spatial-Temporal Network (STN) [14] to discard the surrounding structures before feature encoding. This preprocessing step ensures the area of interest in fetal head structure is located in the center of the image.

The obtained US sequence X is convoluted separately in the space and time dimensions, which is more efficient in preserving image properties than a joint 3D convolution [15]. Specifically, we leverage the contracting path of U-Net to form the 2D spatial encoder E_S. The output $E_S(X)$ aggregated for input length $|t_n - t_i|$ is then fed into a TCN E_T to capture the dynamics of the anatomical context from adjacent time slices. E_T consists of four residual convolution blocks operated along the time channel. Within each block, there are two layers of dilated causal convolution, weight normalization, and nonlinear activation (ReLU) followed by a residual connection at the input and output feature representations. Compared to a recurrent convolutional network (RCN), TCN reduces the computational complexity in sequential modeling with lower memory cost, which is an important design consideration for prediction in real-world clinical scanning.

Table 1. The output feature shape of each module component in DASPG. B denotes the mini-batch size.

Module	Output Size
E_S	$B \times |t_n - t_i| \times 256 \times 14 \times 14$
E_T	$B \times 256 \times 14 \times 14$
DA	$B \times 256 \times 14 \times 14$
D_S	$B \times 224 \times 224$

Domain adaption By definition, an SP contains defined structures with anatomical meanings which may not exist or be clearly visible in other frames of a search video that are "off plane". To investigate the implicit correlation between the search video and the searched SP, we exploit transfer learning to adapt the knowledge between them. We first define the US video-based high-level representation $E_T(E_S(X))$ is from the source domain, and the image-based representation $E_S(y)$ embedded from the shared spatial encoder E_S is from the target domain. As shown in Fig. 1, the domain adaption (DA) module has a residual block [16] with two fully-connected (FC) layers on the flattened feature map to extract the domain expertise from video representation.

The output representation of DA is fed into a spatial decoder D_S to translate the SP image. D_S is formed by the expansive path (decoder) of U-Net without skip connections. Table 1 demonstrates the detailed feature sizes of the above-mentioned modules in generating a 224×224 SP image.

2.2. Objective Function and Adversarial Training

The DA module is optimized through a domain transfer loss $\mathcal{L}_{DA}$ to maximize the cosine similarity between the flattened high-level representations of the US video DA$(E_T(E_S(X)))$ and image $E_S(y)$:

$$\mathcal{L}_{DA} = 1 - < \text{DA}(E_T(E_S(X))), E_S(y) > \quad (1)$$

To constrain the spatial encoder-decoder (*i.e.*, E_S and D_S), we reconstruct the search video X and SP image y with $\mathcal{L}_1$ loss, respectively:

$$\mathcal{L}_{rec} = \frac{1}{|t_n - t_i|} \sum_{t=t_i}^{t_n} ||D_S(E_S(x_t)) - x_t|| + ||D_S(E_S(y)) - y|| \quad (2)$$

The autoencoder structure for reconstruction is shared between video frames and the SP image to recognize the ultrasound-specific spatial semantics. We also regularize

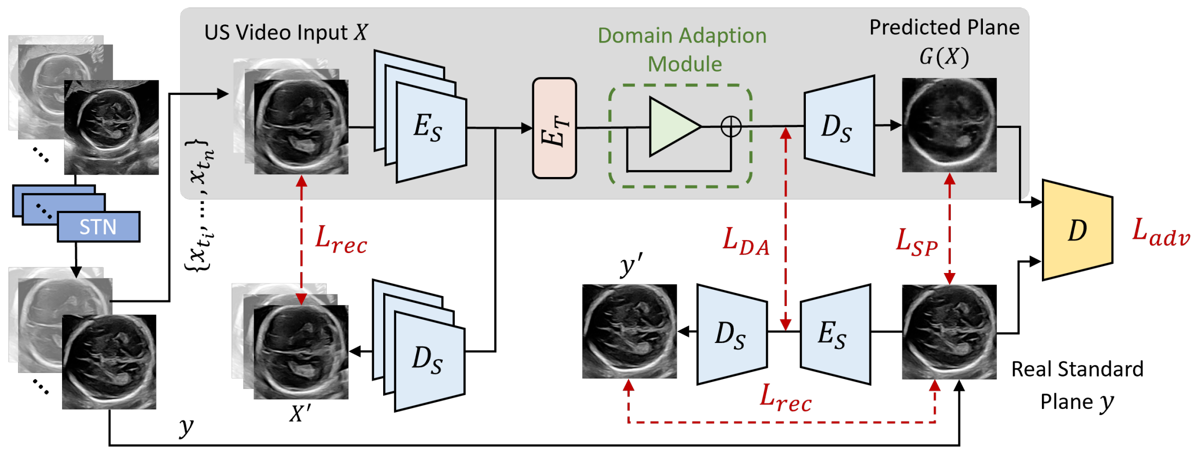

Fig. 1. The architecture of DASPG training and inference. The inference process is shaded in gray. X' and y' represent the reconstructions from the input video and SP image, respectively.

the predicted image $G(X)$ with the intensity loss $\mathcal{L}_{SP} = ||G(X) - y||$ supervised by the content of the real image y.

As an auxiliary loss, adversarial learning is also employed in the proposed DASPG to increase the realism of the synthesized SP image. We use MobileNetV2 [17] with lightweight depthwise-separable convolutions as the discriminator D to classify between a real SP image y and the predicted SP image $G(X)$, and the adversarial loss $\mathcal{L}_{adv}$ is given by:

$$\mathcal{L}_{adv} = \mathbb{E}_X \log(1 - D(G(X))) + \mathbb{E}_y \log D(y) \quad (3)$$

The overall objective combines all four losses $\mathcal{L} = \mathcal{L}_{DA} + \mathcal{L}_{SP} + \mathcal{L}_{adv} + \mathcal{L}_{rec}$ with equal weights.

3. EXPERIMENTS

3.1. Dataset and Implementation Details

The experimental dataset contains 103 routine obstetric videos of the fetal head from the *Anomaly* scan (within the second trimester) and the *Growth* scan (within the third trimester). An US video clip is selected within 10s before the cine-buffer-corrected SP and downsampled to 6Hz. The training/test scan split is 74/29. The training scans are augmented by 1) random flipping horizontally or vertically, and 2) randomly selecting 12 consecutive frames (*i.e.*, $|t_n - t_i| = 12$) as training input. Each test scan is split into four non-overlap clips of 12 frames in length to form the test clips for evaluation. For model implementation, the kernel size of E_T is set to 2, and the output channels of its 4 layers are 8, 6, 3, and 1. In DA, there are 1,024 hidden units and a ReLU activation between the two FC layers. The whole network is trained for 300 epochs with an AdamW optimizer. The initial learning rate is 1e-3 decayed by 1e-2 every 100 epochs.

Table 2. Quantitative results of different temporal architectures and losses in terms of KLD (↓) and FSD (↓).

Architecture			Loss				KLD	FSD
R3D	2D+RCN	2D+TCN	$\mathcal{L}_{DA}$	$\mathcal{L}_{SP}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{rec}$		
		✓			✓	✓	1.217±0.070	182.24
		✓		✓	✓	✓	0.513±0.086	97.46
		✓	✓	✓		✓	0.265±0.100	83.41
✓			✓	✓	✓	✓	0.368±0.228	93.44
	✓		✓	✓	✓	✓	0.476±0.107	129.96
		✓	✓	✓	✓	✓	**0.251±0.101**	**80.41**

3.2. Quantitative Evaluation

Numerically we compare the quality of a generated image with a real SP using Kullback-Leibler Divergence (KLD) and Fréchet SonoNet Distance (FSD) [5]. KLD characterizes tissue-specific speckle differences based on the histogram statistics of two US images, and FSD tests the overall quality of US image appearance. Different from Fréchet Inception Distance (FID) [18] designed for natural images, FSD is more effective in measuring the ultrasound-specific image quality by using SonoNet-64 [19] as the image feature extractor [5].

We first evaluate each component of DASPG in Table 2. The baseline model given in the first row is when only considering the GAN-based generator [20] along with the reconstruction loss to constrain the autoencoder. When comparing the KLD in the first and second rows, we observe that regularization of image intensity substantially benefits the generator. Furthermore, by adapting the predicted distribution to the target SP, the involvement of the DA module consistently increases the performance (see the results in the bottom row). The reduction in FSD also indicates that DA narrows the image quality gap between the SP searching and capture stages. In terms of architecture, integrally modeling the spatio-temporal characteristics in US video using 3D residual convolutions (R3D) [21] is less stable with a large standard

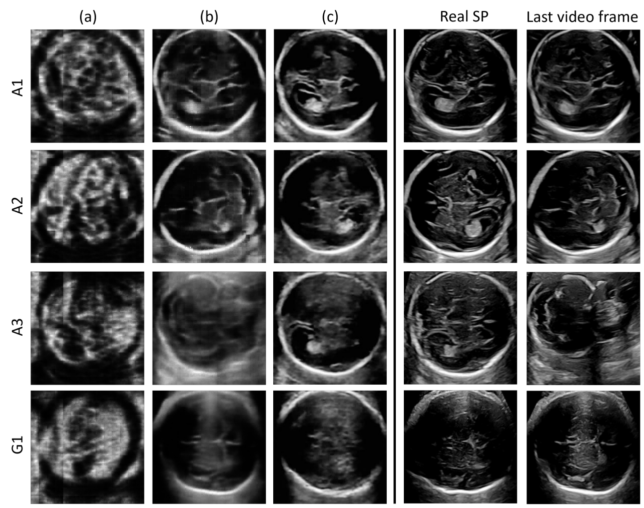

Fig. 2. Qualitative performance of the predicted standard plane images. Here, (a) baseline with only $\mathcal{L}_{adv}$ and $\mathcal{L}_{rec}$ (b) the proposed generator without DA module (c) the proposed generator with DA module. Note that A1-3 are three examples of *Anomaly* scans, and G1 is an example of a *Growth* scan.

deviation, where spatial features might mix up with temporal features in the SP regression. When comparing the bottom two rows, the dilated convolution in TCN is superior to RCN in modeling the temporal patterns in ultrasound SP searching.

3.3. Qualitative Evaluation

Example SP predictions for different models and scans are presented in Fig. 2. The visual appearance of the main anatomical structures, such as the skull, the choroid plexus (CP), midline, and cavum septum pellucidum (CSP) are recognized in predicted TV planes with the individual visibility varying between examples, as shown in Fig. 2(c). Comparing Fig. 2(a) and (b) shows that the intensity loss improves the baseline by preserving the grayscale map and speckle texture of an US image. However, compared to Fig. 2(c), the predicted plane without DA in Fig. 2(b) is blurred and closer in appearance to the last plane in the input scanned video (the rightmost column in Fig. 2). A shift to the target plane using DA (in Fig. 2(c)) helps create a realistic SP with clearer boundaries of the anatomical structures.

In terms of data diversity, while image appearance in the search video is far from the SP (shown in the challenging case of A3), the prediction still correctly estimates the outline of head and the direction of internal anatomical structures. Further, the prediction of G1 shows that the model can generate a realistic result for the third-trimester scan which has higher variability in fetal head appearance. This shows the model has generalizability toward different phases of obstetric scan.

To test how the choice of input video affects prediction

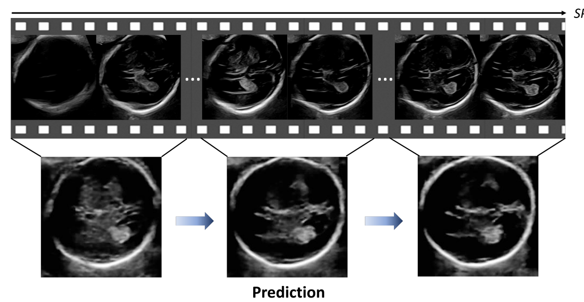

Fig. 3. The predictions at different observation levels.

quality, in Fig. 3, we compare predicted planes generated by input video clips at different temporal distances from the real SP. Predictably, the quality of the generated plane increases when the input clip is closer to the real SP. The anatomical structures predicted from the scan approaching SP are more clear and more recognizable since the video dynamic becomes stable with the visual content more similar to the standard view.

4. CONCLUSION

We have proposed a fetal US SP image predictor (DASPG) with anatomical structures inferred from the input SP search video. Apart from a video-based encoder and an image-based decoder, the main architecture of DASPG is a domain adaption module that translates the encoded features of an SP search video to the SP image representation in target domain. The predicted SP image on TV scan is shown to be anatomically consistent with a real SP image, which can be used as the target image with the expected structures to guide the clinical SP acquisition. As future work, we will apply DASPG on more standard views with clinical evaluations such as anatomical landmark detection and usability study.

5. COMPLIANCE WITH ETHICAL STANDARDS

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

6. ACKNOWLEDGMENTS

This work was supported by the ERC (ERC-ADG-2015 694581, project PULSE), the EPSRC (EP/MO13774/1, EP/R013853/1), and the NIHR Biomedical Research Centre.

7. REFERENCES

[1] Benny Burger, Sascha Bettinghausen, Matthias Radle, and Jürgen Hesser, "Real-time gpu-based ultrasound

simulation using deformable mesh models," *IEEE Transactions on Medical Imaging*, vol. 32, no. 3, pp. 609–618, 2012.

[2] Oliver Mattausch, Maxim Makhinya, and Orcun Goksel, "Realistic ultrasound simulation of complex surface models using interactive monte-carlo path tracing," in *Computer Graphics Forum*, 2018, vol. 37, pp. 202–213.

[3] Yipeng Hu, Eli Gibson, Li-Lin Lee, Weidi Xie, Dean C Barratt, Tom Vercauteren, and J Alison Noble, "Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks," in *MICCAI RAMBO*, pp. 105–115. 2017.

[4] Lin Zhang, Tiziano Portenier, Christoph Paulus, and Orcun Goksel, "Deep image translation for enhancing simulated ultrasound images," in *MICCAI ASMUS*, pp. 85–94. 2020.

[5] Lok Hin Lee and J Alison Noble, "Generating controllable ultrasound images of the fetal head," in *ISBI*, 2020, pp. 1761–1764.

[6] Jiamin Liang, Xin Yang, Yuhao Huang, Haoming Li, Shuangchi He, Xindi Hu, Zejian Chen, Wufeng Xue, Jun Cheng, and Dong Ni, "Sketch guided and progressive growing gan for realistic and editable ultrasound image synthesis," *Medical Image Analysis*, vol. 79, pp. 102461, 2022.

[7] Kimberly Amador, Matthias Wilms, Anthony Winder, Jens Fiehler, and Nils Forkert, "Stroke lesion outcome prediction based on 4d ct perfusion data using temporal convolutional networks," in *MIDL*, 2021, pp. 22–33.

[8] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao, "Deep image-to-video adaptation and fusion networks for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3168–3182, 2019.

[9] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *AAAI*, 2019, vol. 33, pp. 865–872.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[11] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[12] Xiaoli Wang, Zhonghua Liu, Yongzhao Du, Yong Diao, Peizhong Liu, Guorong Lv, and Haojun Zhang, "Recognition of fetal facial ultrasound standard plane based on texture feature fusion," *Computational and Mathematical Methods in Medicine*, 2021.

[13] Ruowei Qu, Guizhi Xu, Chunxia Ding, Wenyan Jia, and Mingui Sun, "Standard plane identification in fetal brain ultrasound scans using a differential convolutional neural network," *IEEE Access*, vol. 8, pp. 83821–83830, 2020.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," *NeurIPS*, vol. 28, 2015.

[15] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018, pp. 6450–6459.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, vol. 30, 2017.

[19] Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert, "Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2204–2215, 2017.

[20] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[21] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *CVPR*, 2018, pp. 6546–6555.