A Dual Adversarial Calibration Framework for Automatic Fetal Brain Biometry

Yuan Gao[1*†] Lokhin Lee[1*] Richard Droste[1,2*] Rachel Craik[1,3] Sridevi Beriwal[1]
Aris Papageorghiou[1] Alison Noble[1]
[1] University of Oxford [2] Amazon [3] Kings College London

Abstract

This paper presents a novel approach to automatic fetal brain biometry motivated by needs in low- and medium- income countries. Specifically, we leverage high-end (HE) ultrasound images to build a biometry solution for low-cost (LC) point-of-care ultrasound images. We propose a novel unsupervised domain adaptation approach to train deep models to be invariant to significant image distribution shift between the image types. Our proposed method, which employs a ***Dual Adversarial Calibration*** *(DAC) framework, consists of adversarial pathways which enforce model invariance to; i) adversarial perturbations in the feature space derived from LC images, and ii) appearance domain discrepancy. Our* ***Dual Adversarial Calibration*** *method estimates transcerebellar diameter and head circumference on images from low-cost ultrasound devices with a mean absolute error (MAE) of 2.43mm and 1.65mm, compared with 7.28 mm and 5.65 mm respectively for SOTA.*

1. Introduction

Pregnancy dating is a crucial part of obstetric care because antenatal care and interventions aimed at improving pregnancy outcome rely on knowledge of the gestational age (GA). Measurements of the size of specific fetal head anatomies is routinely performed to estimate and validate GA. Biometries used include the fetal skull Head Circumference (HC) and the Transcerebellar Diameter (TCD) [1], easily measured on ultrasound images from high-end (HE) machines. US images from HE imaging machines have high imaging contrast, high imaging definition and low speckle noise compared to low-cost (LC) ultrasound images which are acquired with point-of-care (POC) ultrasound probes with greater varied image appearance and hence quality [5]. However, HE imaging may not be available in resource-constrained areas.

Previous literature has considered automated approaches for fetal brain biometry on HE 2D ultrasound (US) images for GA estimation. [20] use contour detection and graph cuts for HC estimation. [14] propose a regional convlutional neural network for detection of key anatomical structures. [21] use U-Nets for HC segmentation and measurement. More recently, [24] directly regresses HC measurements from ultrasound images without segmentation and [13] directly regresses GA from fetal head images using a Bayesian neural network. However, these methods estimate on mid-end US images, which may not be available in resource-constrained settings.

In this paper, we consider jointly learning HC and TCD automated fetal biometry from partially labelled US images acquired with a HE ultrasound machine (GE Voluson E8) combined with unlabelled data from a LC POC ultrasound probe (Konted C10R) for biometry on LC images. This is clinically relevant in a case where interobserver variation on ground truth labelling on LC images is high, due to the reduced image quality and fuzzy edges, and can be useful where a central corpus of well-labelled HE images are available for use for validation and inference on LC images. The core assumption is that domain invariant representations for feature extraction can be jointly learned from LC and HE ultrasound images, and unsupervised learning can calibrate the model in the LC domain so to produce consistent predictions. To this end, we propose a **Dual Adversarial Calibration** (DAC) approach exploiting two adversarial pathways. One pathway forces predictions from LC and HE US images to lie on the same output manifold by training a segmentation network with a discriminator which learns to classify between them. The other pathway forces the LC output to be invariant to self-paced adversarial noise perturbations. Additionally, we propose a novel asymmetric domain augmentation technique specifically designed to cope with the appearance discrepancy between HE and LC images. Experimental results presented show that our proposed approach significantly improves the performance of HC and TCD biometry on LC US images compared to a neural network trained on HE images alone. Ablation experiments also reveal that our dual adversarial pathways lead to a network that is able to learn useful feature representations that are domain invariant.

*Equal contribution.
†Corresponding author: Yuan.Gao2@eng.ox.ac.uk

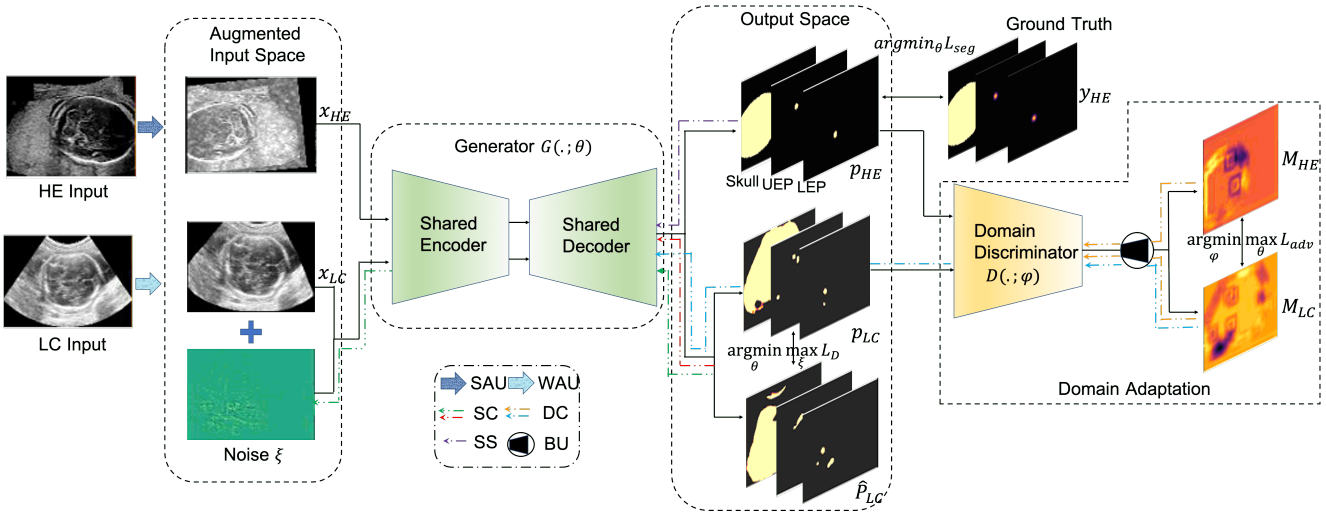

Figure 1: Overview of the training procedure for our proposed **dual adversarial calibration** network, which is trained with two adversarial signals simultaneously. 1) a Domain Calibration (DC) pathway designed to learn domain-invariant segmentation, and 2) a Segmentation Calibration (SC) pathway designed to learn robust segmentation on the unlabelled LC images. SAU: Strong Augmentation; WAU: Weak Augmentation; SS: Supervised Segmentation (on HE images); BU: Bilinear Upsampling. UEP, LEP explained in section 3.2.

1.1. Contributions.

Our contributions are three-fold: 1) we leverage the supervised learned knowledge from HE ultrasound images to significantly improve biometry estimation on LC ultrasound images; 2) we introduce a novel dual adversarial unsupervised intra-modality domain and semantic transfer for this purpose; 3) our results are shown to be competitive compared to SOTA for both automated HC and TCD estimation on LC US images.

2. Related Work

2.1. Medical Cross-Modality Domain Adaptation.

Medical cross-modality domain adaptation aims to retain network performance from the distribution change from an image resulting from one imaging modality to another. Examples of imaging modalities include magnetic resonance imaging (MRI), computed tomography (CT) and ultrasound (US) imaging. Prior literature focused on cross-modality domain adaptation between CT/MR [7, 11, 3, 6]. In detail, [7] uses an adversarial domain adaptation module to map target input features to the output domain space. [11] use a tumour-preserving cycle-consistency loss to map CT and MRI images before training a U-Net to segment lung MRI scans. [3] use a shared encoder space with adversarial based domain adaptations for the segmentation of cardiac structures. [6] share convolutional kernels between MRI and CT images, but use modality specific normalization layers to improve cross-modality performance. Comparatively, literature on cross-modality domain adaptation involving US is limited. [12] generate synthetic MR fetal head images from US scans, but validation is limited as only appearance is evaluated without segmentation results.

2.2. Intra-Modality US Domain Adaptation

Previous literature on intra-modality US domain adaptation focused on adaptation between different HE imaging devices [4, 15, 18]. [4] use a u-net for left atrium segmentation in 3D ultrasound, and incorporate the imaging device used as prior knowledge during inference. [15] use a hierarchical style transfer network to modify image appearances from a target domain to the source domain for fetal head and abdomen segmentation. [18] considered intra-domain ultrasound adaptation for image classification using mutual information minimization. However, in all the examples above, the domain gap in considered was limited as both source and target domain images were acquired with HE ultrasound machines with similar imaging capabilities.

2.3. Low-Cost Ultrasound Probe Image Analysis

Three recent papers investigate learning from low-cost POC probes [8, 23, 17]. Gao et al.[8] developed an image quality assessment framework to identify frames from LC US that can be used for manual TCD measurement by a sonographer. Van den Heuvel et al.[23] investigate gestational age estimation on downsampled HE images to evaluate GA estimation on degraded images, but do not directly evaluate from LC POC probes. Maraci et al.[17] segments the cerebellum for TCD measurement, but the LC US image led to poor segmentation performance.

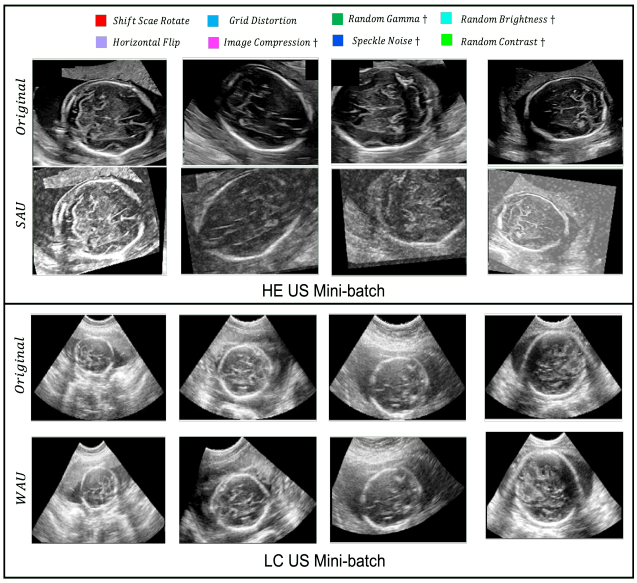

Figure 2: Asymmetrical Domain Augmentation. SAU: Strong Augmentation; WAU: Weak Augmentation. †: Extra augmentations included in SAU.

3. Method

In this section, we present the **Dual Adversarial Calibration** framework in detail, which is illustrated in Fig.1. The framework consists of four main parts: 1) a domain dependent asymmetrical augmentation module of the input space; 2) a segmentation network consisting of a shared encoder-decoder framework; 3) a domain calibration (DC) adversarial pathway for semantic transfer on the predicted output space; 4) a segmentation calibration (SC) cycle pathway for unsupervised semantic transfer.

3.1. Asymmetrical Domain Augmentation.

We observe that in practice, HE data are fairly consistent in imaging quality and appearance, whereas LC data can vary quite substantially. Key anatomies, such as the cerebellum and the thalamus are always clearly visible on HE data, but may not be so on LC data, as imaged structures do not have clear edges and acoustic artefacts such as shadows and speckle can lead to further image degradation. We therefore augment each input domain asymmetrically. Specifically, weak augmentation is applied on LC data to maximize network generalization whilst preserving the spatial visibility of key anatomies by the inclusion of linear and non-linear grid distortions, and horizontal flipping. Strong augmentation is applied on HE data to simulate noisy images including random gamma, random brightness and contrast adjustment, image compression and artificial speckle noise. We qualitatively observe these data augmentations act to decrease domain gap between HE and LC data, as shown in Figure 2, which leads to a stronger response to domain calibration.

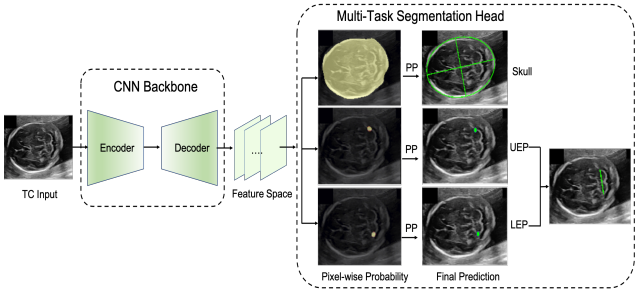

Figure 3: The pipeline from segmentation to the final biometry.

3.2. Core Segmentation Network.

We formulate fetal biometry via segmentation tasks. The HC is derived from segmentation of the fetal skull. However in LC US images, segmentation of the cerebellum is challenging due to low image contrast and noise from signal attenuation. We therefore adopt a strategy inspired by TCD measurement in real-clinical practice, where two points are placed on the cerebellum boundary edge for TC diameter estimation. Thus, the network predicts the upper extreme point (UEP) and lower extreme point (LEP) of the cerebellum, from which the TCD is estimated.

We therefore target two image analysis tasks - fetal skull segmentation for the HC, and UEP and LEP detection for TCD estimation. We employ a U-Net based segmentation network, represented by $\{G(x;\theta), x \in (x_{LC}, x_{HE})\}$ where x_{LC} and x_{HE} are examples of HE and LC input US images respectively and θ represents model parameters. We use ResNet18 [9] as an encoder-decoder backbone with residual blocks (fine to coarse: 64, 64, 128, 256 ,512 feature channels) for the encoder and residual blocks with 2D bilinear upsampling (coarse to fine: 256, 128, 64, 32, 16 feature channels) for the decoder with skip connections between the encoder-decoder. Our segmentation head consists of a single 3×3 2D convolutional layer to map the decoded feature maps (16 channels) to 3 channels, corresponding to the skull, UEP and LEP predictions. We then apply a pixel-wise sigmoid function to the segmentation output to obtain a pixel-wise probability map for each anatomy. To address class imbalance, we use a DICE loss to train the segmentation model defined as:

$$L_{seg} = 1 - \frac{2\sum_{h,w,c} y_{HE} p_{HE}}{\sum_{h,w,c} y_{HE}^2 + \sum_{h,w,c} p_{HE}^2} \quad (1)$$

where $y_{HE} \epsilon \mathbb{R}^{h \times w \times c}$ and $p_{HE} \epsilon \mathbb{R}^{h \times w \times c}$ are the ground truth annotations and the pixel-wise probability maps for HE images, h, w and c are height, width and number of classes respectively (c=3 in our experiments).

The whole pipeline from input to final biometric estimation is depicted in Figure 3. We got the segmentation probability maps for each structure i.e. Skull, UEP and LEP. For computing HC, we segment the entire fetal head, retrieve a skull contour from the probability map, then fit an ellipse to the contour, obtain the center, major and minor axis, and rotation of the fitted ellipse. For computing TCD, we perform non-maximum suppression to find the pixel with greatest probability for UEP and LEP, then draw a line between the two estimated point locations.

3.3. Domain Calibration Pathway.

We observe that x_{HE} and x_{LC} have very different intensity distributions. This reflects in imaging quality discrepancy, and networks trained on one do not perform well on the other. To address this, we introduce an adversarial pathway to calibrate the underlying output space to be invariant to the input domain [22]. Specifically, the output space can be modelled as a low-dimensional manifold that contains simple representations and rich semantic information about target anatomies. By minimizing the distance between HE US predictions $p_{HE} = G(x_{HE}; \theta)$ and LC US predictions $p_{LC} = G(x_{LC}; \theta)$, the model can learn specific target anatomical regions of interest from a low dimensional representational space. We adopt a Least-Square GAN [16] loss for domain adaptation. The domain adaptation is modelled with the objective:

$$\begin{aligned}\underset{\varphi}{\operatorname{argmin}} L_{adv}(D) := \frac{1}{2}\mathbb{E}_{p\sim p_{out}(HE)}[(D(p_{HE};\varphi)-1)^2] \\ + \frac{1}{2}\mathbb{E}_{p\sim p_{out}(LC)}[(D(p_{LC};\varphi))^2]\end{aligned} \tag{2}$$

$$\underset{\theta}{\operatorname{argmin}} L_{adv}(G) := \frac{1}{2}\mathbb{E}_{p\sim p_{out}(LC)}[(D(p_{LC};\varphi)-1)^2] \tag{3}$$

Here $D(p;\phi)$ represents the input probality map p to our discriminator parameterised by ϕ. The discriminator consists of five convolutional layers (fine-to-coarse: 64, 128, 256, 512, 1 feature channels) with a kernel size of 3 and stride of 1 and each followed by a leaky ReLU parameterised by 0.2 and a max pooling layer.

3.4. Segmentation Calibration Pathway.

The generator $G(x, \theta)$ is trained on supervised segmentation with the labelled source input x_{HE} with available ground truth y_{HE}. We further regularize the model using unlabelled x_{LC} images by using adaptive perturbations by noting that the predicted segmentation should be locally smooth to adversarial perturbations in the input x_{LC} images [19]. As shown in Fig.1, we generate an augmented LC input tuple $X = \{(x_{LC}, x_{LC} + \xi), \xi \sim N(0, 1)\}$ and prediction tuple $P = (p_{LC}, \hat{p}_{LC})$ from the segmentation network $G(X; \theta)$. ξ is self-paced and updated with a cycle pathway by maximizing the distance between $(p_{LC}, \hat{p}_{LC})$, generating ξ_{adv}, an adversarial perturbation, then minimising the Kullbeck-Leibler loss $KL\left[G(x;\theta) \parallel G(x+\xi_{adv})\right]$ to reduce network sensitivity to noise perturbations. To do this we compute the derivative of L_D w.r.t. ξ_{adv} defined as $g_\xi = \bigtriangledown_{\xi_{adv}} L_D$ evaluated with backpropagation, and the perturbation as $\xi_{adv} = \epsilon \frac{g_\xi}{\|g_\xi\|_2}$, where ϵ is a hyper-parameter that controls the strength of perturbation. The goal is therefore for the network to be robust to the perturbations so as to produce consistent outputs from the perturbed inputs. The distance loss L_D is therefore defined as:

$$\begin{aligned}&\underset{\theta}{\operatorname{argmin}} L_D(x, \xi, \theta) := \\ &\mathbb{E}_{x\sim p_{data}(LC)} KL[[G(x_{LC};\theta)] \parallel [G(x_{LC}+\xi_{adv};\theta)]] \\ &\text{s.t. } \xi_{adv} := \underset{\xi}{\operatorname{argmax}} \{L_D(x_{LC}, \xi, \theta); \|\xi\|_2 \leq \epsilon\}\end{aligned} \tag{4}$$

3.5. Optimisation.

Our model optimization is performed in a two-step process during model training. The domain discriminator is first optimized by minimizing the loss $L_{adv}(D)$ while the generator is frozen. The generator is then subsequently optimized by optimizing the joint loss L_{joint}, defined as:

$$\begin{aligned}&L_{joint}(x_{HE}, x_{LC}, \xi, \theta) = \\ &L_{seg}(X_{HE}, \theta) + \alpha L_D(X_{LC}, \xi, \theta) + \beta L_{adv}(G)\end{aligned} \tag{5}$$

where the hyperparameters α and β determine the strength of segmentation and domain calibration respectively.

4. Experiments and Results

4.1. Datasets.

We have two different datasets acquired from two different clinical studies, examples of which can be seen in Fig. 4. HE and LC images were acquired from a clinical US scanner and a POC probe respectively described in Table 1, along with the distribution of scans used for training and testing. The difference in image quality is apparent in Fig. 4. This includes reduced imaging contrast between tissues of high echogenicity (fetal skull) and low echogenicity (internal brain tissue), reduced clarity of internal brain structures such as the cerebellum and ventricles, reduced edge sharpness as can be seen from the boundary of the skull. All of these features are clinically used for determination of TCD and HC. Furthermore, image resolution is reduced and there is increased noise in LC images. We used TC

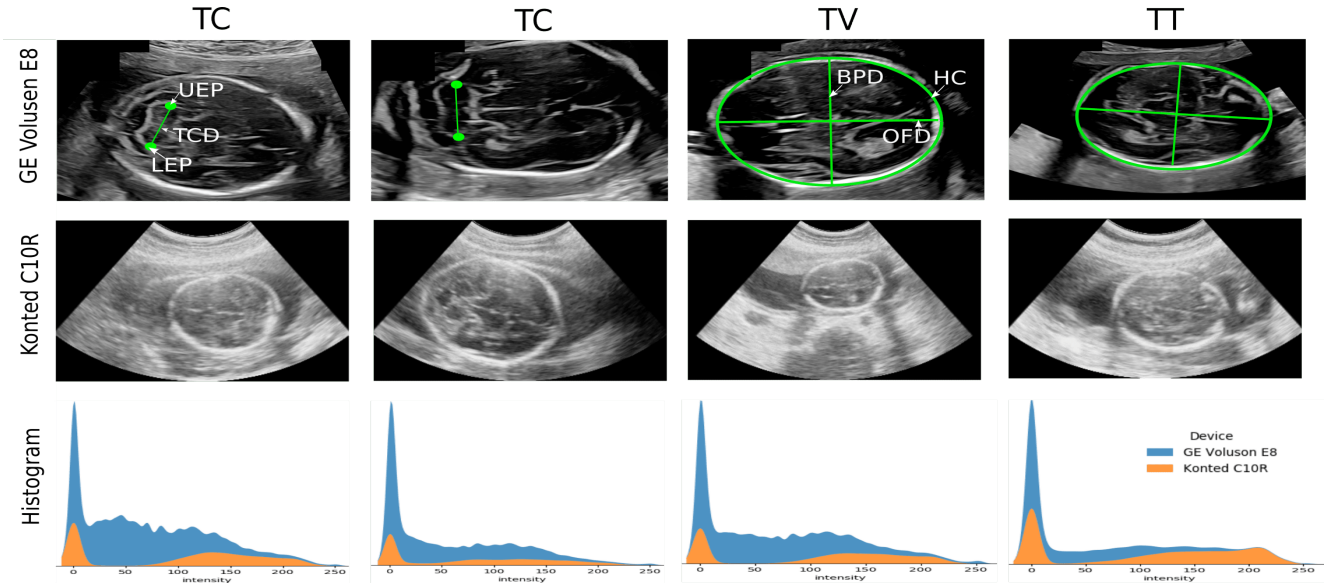

Figure 4: Examples of fetal brain biometry planes acquired by different devices and pixel intensity distributions. HE scanner: GE Voluson E8, LC probe: Konted C10R. TC: Transcerebellar planes, TV: Transventricular planes, TT: Transthalamic planes, UEP: Upper Extreme Point, LEP: Lower Extreme Point, TCD: Transcerebellar Diameter, HC: Head Circumference, BPD: Biparietal Diameter, and OFD: Occipito-Frontal Diameter.

Table 1: Dataset description for training and evaluation of our model. Planes are split between test and training on a subject basis to prevent data leakage.

	No. of Subjects	Acquisition Device	Training (Seg. Labels)	Testing (Biometry [mm])	
			TC Planes	TC Planes (HC)	TT/TV Planes (TCD)
HE Dataset	540	GE Volusen E8	519 (Labelled HC & TCD)	-	-
LE Dataset	560	Konted GEN1 C10R	418 (Unlabelled)	387 (Labelled)	526 (Labelled)

planes from the HE dataset segmented with HC and TCD measurement points along with unlabelled TC planes from the LC dataset to train our model. The performance of our model was then evaluated by comparison between biometry from inferred segmentations and the biometry extracted by an expert sonographer.

Images from each source differed in input resolution (784×1008 px HE, 228×378 px LC). All images were resized to 448×576px. Ground truth pixel maps for the fetal skull were labelled $x_i \in \{0, 1\}$ for pixels inside and outside the skull respectively. For TCD estimation, instead of labelling the entire cerebellum, a Gaussian kernel was centered on the sonographer annotated key points i.e. UEP and LEP used as the ground truth for training. As seen in table 2, we achieve mean TCD error of 2.43 mm and mean HC error of 1.65mm on the best performing model.

4.2. Network Training.

Our model was implemented with pytorch 1.4.0 and trained on a single Quadro RTX 5000 GPU. $G(\theta)$ was optimized with Nesterov accelerated SGD with an initial learning rate of 0.1, momentum of 0.9 and a weight decay of 10^{-3}. $D(\phi)$ was optimized with Adam with an initial learning rate of 10^{-4} and weight decay of 10^{-4}. Both $D(\phi)$ and $G(\theta)$ were optimized for 70 epochs with a minibatch size of 4. Both learning rates were multiplied by a factor of 0.1 after epochs 40 and 60. Grid search was performed on a logarithmic scale for α and β; $\alpha = 10^{-1}$ and $\beta = 10^{-3}$ gave the best performance.

4.3. Results and Ablation Study

To better understand the individual contributions from each component of our model, we perform an ablation study which composed of three settings: 1) we train the segmentation network with labelled HE images and the directly apply to LC images at test time; 2) we investigate domain adaptation from either feature or output space by applying adversarial training; 3) in addition to domain adaptation, we incorporate the self-paced unsupervised branch for segmentation calibration. We show in Table 2 (①, ②) which are

Table 2: Ablation study of our method and comparison to SOTA for HC and TCD estimation on LC test dataset. Mean Absolute Error (MAE) ± std is reported.

Method	Aug.	DC, Adpt. Loss, Adpt. Space	SC	TCD (mean ± SD [mm])	HC (mean ± SD [mm])
① W/o	✓, Weak	-	-	46.63±8.64	36.27±7.81
② W/o	✓, Strong	-	-	30.96±7.46	22.52±7.54
③ DC	✓, Asymmetrical	✓, V-GAN Loss, out. space	-	13.80±3.91	10.21±3.63
④ DC	✓, Asymmetrical	✓, LS-GAN Loss, feat. space	-	8.64±1.42	6.53±1.11
⑤ DC	✓, Asymmetrical	✓, LS-GAN Loss, out. space	-	7.93±1.64	4.31±1.25
⑥ DAC	✓, Asymmetrical	✓, LS-GAN Loss, feat. space	✓	4.62±0.46	3.26±0.43
⑦ **DAC**	**✓, Asymmetrical**	**✓, LS-GAN Loss, out. space**	✓	**2.43±0.37**	**1.65±0.31**
CycleGAN [25]	✓, Asymmetrical	✓, Cycle-GAN Loss, in. space	-	14.79±3.20	9.69±3.19
CyCADA [10]	✓, Asymmetrical	✓, CyCADA Loss, in.&feat. space	-	**7.28±2.72**	**5.65±2.25**
UCMDA [7]	✓, Asymmetrical	✓, Adversarial loss, in feat. space	-	8.13±2.30	6.74±2.19
SIFA [2]	✓, Asymmetrical	✓, SIFA Loss, in.&feat.&out. space	-	8.69±1.21	6.32±1.04

* Note: W/o: Without Adaptation; DC: Domain Calibration; SC: Segmentation Calibration; DAC: Dual Adversarial Calibration.

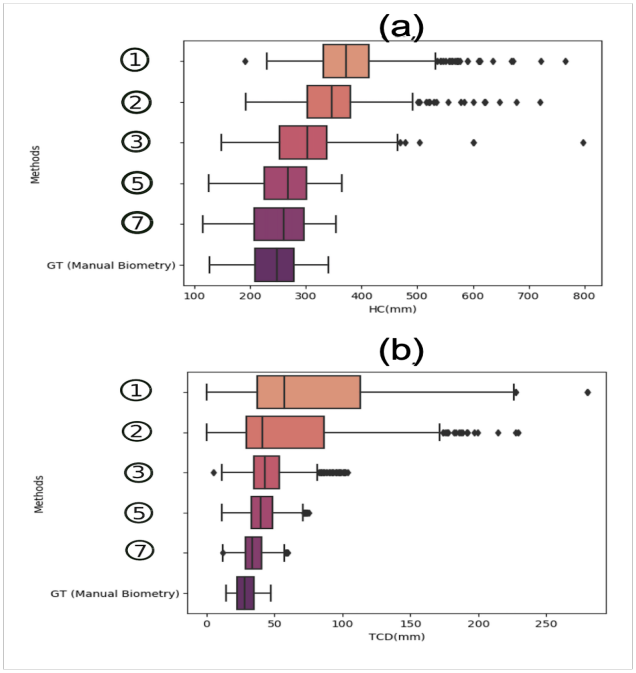

Figure 5: Biometry distributions on LC test dataset from different models numbered 1-7 as in Table 2. (a) predicted HCs (b) predicted TCDs for different models. GT: expert manual biometry.

models trained with HE images directly evaluated on LC images. Direct estimation of biometries from LC images from a model trained with HE images performs poorly as expected due to the substantial domain shift. Inclusion of adversarial calibration (③, ④, ⑤) significantly improves biometry estimation. We also find that training with a LS-GAN adversarial loss outperforms a Vanilla GAN loss, and adversarial calibration on the output space leads to better localization compare to feature space (⑤, ⑦ vs.④, ⑥), which suggests that the output manifold is a suitable rich environment for domain calibration. Further addition of adversarial segmentation calibration (⑥, ⑦) leads to the best performance with MAE of 1.65mm for HC and 2.43mm for TCD estimation.

Also, as can be seen in Fig. 5 (a) and (b), without domain adaptation i.e. ①, ②, there are a number of outliers and the biomertry is significantly out of the distribution compared to the expert's manual biometry for both the HC and TCD. By incorporating the adversarial training ③, ⑤, the model's predicted distribution is shifted towards the ground truth. DAC ⑦ results in the most closed distribution to the expert biomerty. Additionally, we found HC predicted by DAC model tends to be better agreed with expert measurements, compared to TCD and the HC mean absolute error is smaller than TCD's. This may because the appearance of the TCD landmark is prone to be affected by image quality and artefacts, however, the skull signal is bright and consistent, which is less affected by the imaging quality changes.

4.4. Comparison with SOTA.

We compare our proposed model with current SOTA in the lower rows of Table 2. We found that domain adaptation methods that adapt from the input space (CycleGAN [25]) give a high estimation error, as only style is preserved by internal image content and anatomies are not well conserved, especially for TC images. Taking into account the feature space (CyCADA [10]) leads to better performance, but continues to misjudge points of the cerebellum. We also investigated the SOTA methods for unsupervised cross-

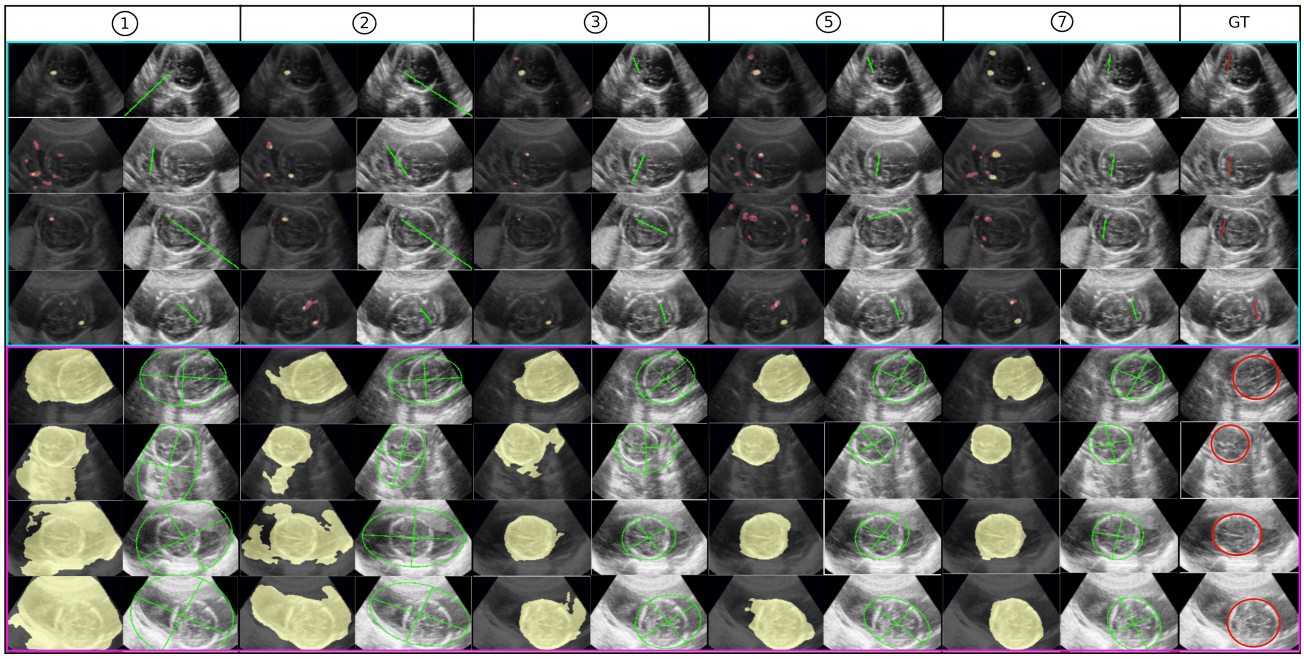

Figure 6: Example outputs from different models numbered 1-7 as in Table 2, evaluated on LC test images (Blue box: TCD, purple box: HC). The GT column represents the ground truth.

modality domain adaptation, UCMDA and SIFA [7, 2]. UCMDA[7] proposed a domain critic module (DCM) that minimizes difference between domains in the feature space. SIFA uses a shared encoder and performs domain adaptation in the output space, and use cycle-GAN based identity loss to learn domain invariant features. We find that these cross modality models not necessarily works on our intra-modality adaptation task and underperform compared to our best performing model ⑦. This may because intra-modality domain shift is relatively moderate compared to cross-modality and the cross-modality models are too complex to fit on the moderate domain shift. Our proposed DAC design focus on aligning the intra-modality domain shift by simply incorporating a self-paced distribution alignment process. The self-paced calibration helps to align the intra-domain feature space and make models more robust against artefacts and variation in imaging quality, introduced by LC probes.

4.5. Qualitative Results.

As shown in Fig. 6, without domain or segmentation adversarial calibration, models (①, ②) fail to localize at least one of the TCD measurement points and fail to segment the skull for HC. We find that including domain calibration (③, ⑤) leads to noisy probability maps for the TCD measurement points, but the final prediction (after non-maximum suppression) become more accurate. Using a LS-GAN loss leads to smoother adversarial calibration for domain adaptation and which leads to more localized prediction of the TCD measurement points (⑤ vs. ③). In addition to domain calibration, we find that segmentation calibration (reflected in Table2 ⑤ vs. ⑦) increases the performance of TCD estimation more than HC estimation. This suggests that adversarial segmentation perturbation helps the network to localize of small targets in a noisy environment. Our complete model ⑦ outperforms all of the above. It correctly identifies the TCD measurement points and segments the fetal skull, even in challenging examples.

5. Conclusion

This paper addresses the problem of domain adaptation from clinical high-end ultrasound images to low cost point-of-care ultrasound images with greater varied imaging quality and increased noise. We proposed a novel dual adversarial calibration framework which enables the network to learn invariant features to both image types, and leverages high-end ultrasound images to enable a solution for accurate automatic biometry on LC US images. Our approach outperforms current SOTA for domain calibration and semi-supervised learning methods applied to this task.

Acknowledgements

We acknowledge the ERC (ERC-ADG-2015 694 project PULSE), the EPSRC (EP/R013853/1, EP/T028572/1) and the MRC (MR/P027938/1).

References

[1] Martin R Chavez et al. Fetal transcerebellar diameter measurement for prediction of gestational age at the extremes of fetal growth. *JUM*, 2007. 1

[2] Cheng Chen et al. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *MLMI*, 2018. 6, 7

[3] Cheng Chen et al. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *AAAI*, 2019. 2

[4] Markus A Degel, Nassir Navab, and Shadi Albarqouni. Domain and geometry agnostic cnns for left atrium segmentation in 3d ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 630–637. Springer, 2018. 2

[5] Christoph F Dietrich et al. Point of care ultrasound: a wfumb position paper. *Ultrasound in medicine & biology*, 2017. 1

[6] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020. 2

[7] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*, 2018. 2, 6, 7

[8] Yuan Gao, Sridevi Beriwal, Rachel Craik, Aris T Papageorghiou, and J Alison Noble. Label efficient localization of fetal brain biometry planes in ultrasound through metric learning. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 126–135. Springer, 2020. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 6

[11] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S Mageras, Joseph O Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 777–785. Springer, 2018. 2

[12] Jianbo Jiao, Ana IL Namburete, Aris T Papageorghiou, and J Alison Noble. Self-supervised ultrasound to mri fetal brain image synthesis. *IEEE Transactions on Medical Imaging*, 39(12):4413–4424, 2020. 2

[13] Lok Hin Lee, Elizabeth Bradburn, Aris T Papageorghiou, and J Alison Noble. Calibrated bayesian neural networks to estimate gestational age and its uncertainty on fetal brain ultrasound images. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 13–22. Springer, 2020. 1

[14] Zehui Lin, Shengli Li, Dong Ni, Yimei Liao, Huaxuan Wen, Jie Du, Siping Chen, Tianfu Wang, and Baiying Lei. Multi-task learning for quality assessment of fetal head ultrasound images. *Medical image analysis*, 58:101548, 2019. 1

[15] Zhendong Liu, Xiaoqiong Huang, Xin Yang, Rui Gao, Rui Li, Yuanji Zhang, Yankai Huang, Guangquan Zhou, Yi Xiong, Alejandro F Frangi, et al. Generalize ultrasound image segmentation via instant and plug & play style transfer. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 419–423. IEEE, 2021. 2

[16] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 4

[17] Mohammed A Maraci, Mohammad Yaqub, Rachel Craik, Sridevi Beriwal, Alice Self, Peter von Dadelszen, Aris Papageorghiou, and J Alison Noble. Toward point-of-care ultrasound estimation of fetal gestational age from the transcerebellar diameter using cnn-based ultrasound image analysis. *Journal of Medical Imaging*, 7(1):014501, 2020. 2

[18] Qingjie Meng et al. Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging. *TMI*, 2020. 2

[19] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*, 2015. 4

[20] Sylvia Rueda, , et al. Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge. *TMI*, 2013. 1

[21] Zahra Sobhaninia, Shima Rafiei, Ali Emami, Nader Karimi, Kayvan Najarian, Shadrokh Samavi, and SM Reza Soroushmehr. Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6545–6548. IEEE, 2019. 1

[22] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 4

[23] Thomas LA van den Heuvel, Hezkiel Petros, Stefano Santini, Chris L de Korte, and Bram van Ginneken. Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries. *Ultrasound in medicine & biology*, 45(3):773–785, 2019. 2

[24] Jing Zhang, Caroline Petitjean, Pierre Lopez, and Samia Ainouz. Direct estimation of fetal head circumference from

ultrasound images based on regression cnn. In *Medical Imaging with Deep Learning*, pages 914–922. PMLR, 2020. 1

[25] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 6