Cross-Device Cross-Anatomy Adaptation Network for Ultrasound Video Analysis

Qingchao Chen[1], Yang Liu[1], Yipeng Hu[3], Alice Self[2], Aris Papageorghiou[2], and J.Alison Noble[1]

[1] Department of Engineering Science, University of Oxford, U.K.
{qingchao.chen, yang.liu, alison.noble}@eng.ox.ac.uk
[2] Nuffield Department of Women's and Reproductive Health,University of Oxford, U.K.
{alice.self, aris.papageorghiou}@wrh.ox.ac.uk
[3] Wellcome/EPSRC Centre for Interventional Surgical Sciences, University College London
{yipeng.hu}@ucl.ac.uk

Abstract. Domain adaptation is an active area of current medical image analysis research. In this paper, we present a cross-device and cross-anatomy adaptation network (CCAN) for automatically annotating fetal anomaly ultrasound video. In our approach, deep learning models trained on more widely available expert-acquired and manually-labeled free-hand ultrasound video from a high-end ultrasound machine are adapted to a particular scenario where limited and unlabeled ultrasound videos are collected using a simplified sweep protocol suitable for less-experienced users with a low-cost probe. This unsupervised domain adaptation problem is interesting as there are two domain variations between the datasets: (1) cross-device image appearance variation due to using different transducers; and (2) cross-anatomy variation because the simplified scanning protocol does not necessarily contain standard views seen in typical free-hand scanning video. By introducing a novel structure-aware adversarial training module to learn the cross-device variation, together with a novel selective adaptation module to accommodate cross-anatomy variation domain transfer is achieved. Learning from a dataset of high-end machine clinical video and expert labels, we demonstrate the efficacy of the proposed method in anatomy classification on the unlabeled sweep data acquired using the non-expert and low-cost ultrasound probe protocol. Experimental results show that, when cross-device variations are learned and reduced only, CCAN significantly improves the mean recognition accuracy by 20.8% and 10.0%, compared to a method without domain adaptation and a state-of-the-art adaptation method, respectively. When both the cross-device and cross-anatomy variations are reduced, CCAN improves the mean recognition accuracy by a statistically significant 20% compared with these other state-of-the-art adaptation methods.

1 Introduction

Although ultrasound (US) imaging is recognized as an inexpensive and portable imaging means for prenatal care, training skilled sonographers is time-consuming and costly, resulting in a well-documented shortage of sonographers in many countries including the UK and the US. 99% of world-wide maternal deaths occur in low-and-middle-income (LMIC) countries where the access to ultrasound imaging is even more limited [7]. To address this challenge, recent academic research on solutions for LMIC setting has proposed a three-component approach: i) the adoption of inexpensive and portable US equipment, ii) designing simple-to-use US scanning protocols, e.g. the obstetric sweep protocol (OSP) [1] and iii) innovating intelligent image analysis algorithms [6,10,9]. Arguably, the third innovation plays a bridging role in this approach, enabling the other two cost-effective components of the solution.

Whilst simplified scanning protocols are less-dependent on user skills, they generate diagnostic images that deviate in appearance to those acquired using the standardized protocols used in fetal assessment. Even experienced sonographers can struggle to interpret and analyze data obtained using simple protocols on inexpensive machines, which are often equipped with older generation transducers and processing units. Furthermore this data degrades modern machine learning models compared to those developed with well-curated datasets [5,4,9]. Refining these existing models trained with site-specific data is an option, but requires additional data to be acquired and manually annotated, which may present substantial logistic challenges in expertise and cost.

As an alternative, in this work, we propose an unsupervised domain adaptation approach to train and adapt deep neural networks, learning from a ***source domain*** of high-end ultrasound machine images and expert annotations to classify a ***target domain*** of unpaired unlabeled images. As illustrated in Fig.1a, in the fetal anomaly examination application of interest, as an "instructor" dataset, the source-domain images are acquired by experienced sonographers following an established fetal anomaly screening free-hand ultra-sound protocol [8], hereafter referred to as the ***free-hand dataset***. The target-domain dataset is ultrasound video acquired using a simplified single-sweep protocol [10], also illustrated in Fig. 1b, hereafter referred to as the ***single-sweep dataset***. Classifying video frames in such single-sweep data into multiple anatomical classes is useful for assisting a range of clinical applications, including anomaly detection, gestational age estimation and pregnancy risk assessment [9].

The proposed unsupervised domain adaptation approach in this paper addresses two unique and specific dataset variations: 1) the variations of anatomical appearance between the source and target training images attributed to acquiring data with two different ultrasound devices (the ***cross-device variation***); and 2) the variations between anatomical class labels where the target domain label set is a smaller subset of the source domain label set (the ***cross-anatomy variation***). We argue that the cross-anatomy variation is common, since there exists richer anatomical structures and a larger anatomical label set in the free-hand dataset (source domain) compared with the single-sweep dataset (target domain). More specifically, we propose a novel structure-aware adversarial domain adaptation network with a selective adaptation module to reduce two types of variations, such that, the model trained using the free-hand dataset with four anatomical-class-labels can be used to effectively classify a single-sweep dataset with two classes.

The contributions of this paper are summarized as follows: 1) for the first time, we propose a Cross-device and Cross-anatomy Adaptation Network (CCAN) to reduce cross-device and cross-anatomy variations between two datasets, as described in section

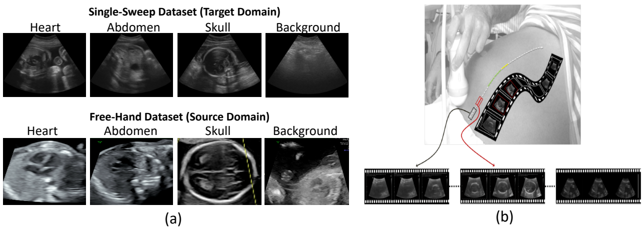
Single-Sweep Dataset (Target Domain)
Heart
Abdomen
Skull
Background
Free-Hand Dataset (Source Domain)
Heart
Abdomen
Skull
Background
(a)
(b)

Fig. 1: (a) Example frames illustrating the cross-device variations between free-hand and single-sweep datasets. (b) Illustration of the single-sweep protocol.

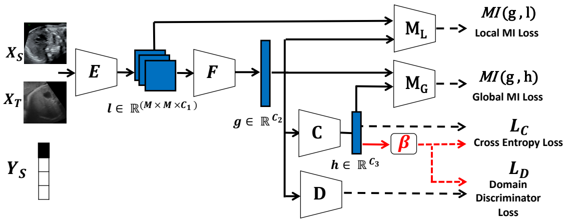

Fig. 2: Architecture of a Cross-Device Cross-Anatomy Adaptation Network.

2; 2) we propose a novel structure-aware adversarial training strategy based on multi-scale deep features to effectively reduce cross-device variations; 3) we propose a novel anatomy selector module to reduce cross-anatomy variations; 4) we demonstrate the efficacy of the proposed approaches with experiment results on two sets of clinical data.

2 Methods

2.1 Cross-Device and Cross-Anatomy Adaptation Network

We assume that the source image X_S with the discrete anatomy label Y_S are drawn from a source domain distribution $P_S(X,Y)$, and that the target images X_T are drawn from the target domain distribution $P_T(X,Y)$ without observed labels Y_T. In our application, the source and target distributions are represented by the free-hand and single-sweep datasets, respectively. Since direct supervised learning using the target labels is not possible, CCAN instead learns an anatomy classifier driven by source labels only, and then adapts the model for use in the target domain.

As illustrated in the modules connected by black lines in Fig.2, the proposed CCAN includes an encoder E, projection layer F, the anatomy classifier C, the domain classifier D and two Mutual Information (MI) discriminators M_L and M_G. Specifically, the source images are first mapped by the encoder E to the convolutional feature $E(X_S)$, and then projected to a latent global feature representation $F(E(X_S))$. Then

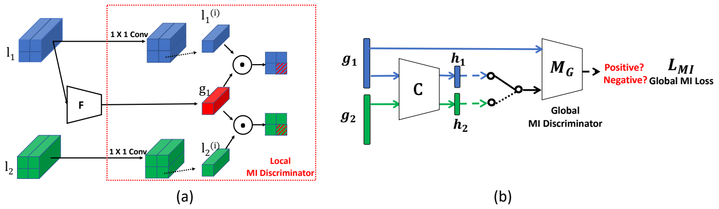

Fig. 3: Architecture and Implementation of (a) local and (b)global MI Estimation.

the anatomy classifier C minimizes a cross-entropy loss L_C, between the ground-truth Y_S and the predicted source labels $C(F(E(X_S)))$, i.e., $\min\limits_{E,F,C} \mathcal{L}_C$. We adopt the adversarial training loss L_D [5] to learn domain invariant features, where the domain classifier D tries to discriminate between features from the source and target domain, while E and F tries to "confuse" D, i.e. $\max\limits_{E,F} \min\limits_{D} \mathcal{L}_D$.

However, facing the large anatomical variations, it is still an open question as to which levels of deep features to align and which should be domain-invariant. To answer this question, and referring to the notation in Fig.2, we propose to align the *distribution of multi-scale deep features* in adversarial training by compressing information from local convolutional feature maps l, and the classifier prediction h into a unified global semantic feature g and to reduce cross-domain variations of g. More specifically, we maximize the *local* and *global* MI losses, $MI(g,l)$ and $MI(g,h)$, between two feature pairs, (g,l) and (g,h) respectively. The MI estimation [2] is achieved by two binary classification losses, distinguishing whether two features are a positive or negative pair from the same image, as shown in Fig.3. Taking $MI(g,h)$ as an example, it relies on a sampling strategy that draws positive and negative samples from the joint distribution $P(g,h)$ and from the marginal product $P(g)P(h)$ respectively. In our case, the positive samples (g_1, h_1) are features of the same input, while the negative samples (g_1, h_2) are obtained from different inputs. That is, given an input g_1 and a set of positive and negative pairs from a minibatch, the global MI discriminator M_G aims to distinguish whether the other input h_1 or h_2 from the same input image as g_1 or not, as shown in Fig.3.

The overall objective can be summarized by the following minimax optimization:

$$\min_{E,F,C} \max_{D} \; -\mathcal{L}_C + \alpha\mathcal{L}_D + \gamma(MI(g,l) + MI(g,h)) \tag{1}$$

where α and γ are the weights of $\mathcal{L}_D$ and MI losses respectively. In this work, the hyper-parameters are set empirically (via grid-searching from the evaluation set) to weight between the classification loss $\mathcal{L}_C$, the domain classification loss $\mathcal{L}_D$ and MI loss $MI(g,l) + MI(g,h)$. The detailed domain discriminator loss $\mathcal{L}_D$ and MI losses therefore, are given by the Eqns (2)-(5). Note that before the inner-product operation in Eqn.(3), we used two projection layers W_h and W_l for classifier prediction h and local feature map l respectively.

$$\mathcal{L}_D = -\frac{1}{N_S}\sum_{j=1}^{N_S} log(D(F(E(x_S^j)))) - \frac{1}{N_T}\sum_{i=1}^{N_T} log(1 - D(F(E(x_T^i)))), \quad (2)$$

$$M_G(g,h) = g^T W_h h,\ M_L(g,l) = \frac{1}{M^2}\sum_{i}^{M^2} g^T W_l l^{(i)}, \quad (3)$$

$$MI(g,h) = \mathbb{E}_{X_P}[log\sigma(M_G(g_1,h_1))] + \mathbb{E}_{X_N}[log(1-\sigma(M_G(g_1,h_2)))] \quad (4)$$

$$MI(g,l) = \mathbb{E}_{X_P}[log\sigma(M_L(g_1,l_1))] + \mathbb{E}_{X_N}[log(1-\sigma(M_L(g_1,l_2)))] \quad (5)$$

2.2 Selective Adaptation Module for Cross-Anatomy Variations

Due to different scanning protocols used to acquire the free-hand and single-sweep datasets, the available anatomical categories of the source domain Y_S often do not correspond to those of the target domain label Y_T. Often, the anatomical class set of source domain C_S may contain classes outside of the one in target domain C_T which henceforth we refer to as the outlier. When this large cross-anatomy variation exists, the network training described in section 2.1 still aims to match *identical* class categories between source and target domain, leading to a trained network prone to cross-anatomy misalignment of the label space. For example, if we directly adapt a source domain model trained using four-class data X_S to the three-class (shared anatomy) target domain data X_T, mismatch may occur between the three-class features and the four-class ones, due to the lack of anatomically paired features. As a result, some features in the target domain may be randomly aligned with the features from the outlier anatomy class, possibly due to the indiscriminative marginal feature distribution.

In this work, we investigate the case where categories of Y_T are a subset of the class categories of Y_S, as the single-sweep dataset contains a smaller number of anatomical classes than the free-hand dataset. We propose CCAN-β, a variant of CCAN with a small modification, shown as the highlighted red β module in Fig.2 to selectively adapt the model training, focusing on the shared anatomical categories $C_S \cap C_T$ while defocusing from the outlier class $C_S \setminus C_T$.

As shown in Fig.2, β is an anatomy-wise weighting vector with the length of the source domain class categories $|C_S|$, with its k^{th} element indicating the contribution of the k^{th} source domain class. Ideally, β functions to down weight the classes from the outlier anatomy class $C_S \setminus C_T$ and promoting the shared anatomy classes in the set $C_S \cap C_T$. Based on this principle, we calculate β simply using the average of target classification predictions $C(F(E(X_T)))$, $\hat{\beta} = \frac{1}{N_T}\sum_{i=1}^{N_T} C(F(E(x_T^i)))$, where β is the normalized vector of $\hat{\beta}$, $\beta = \frac{\hat{\beta}}{\max(\hat{\beta})}$. N_T and x_T^i are the total number of target domain samples and individual target samples, respectively. For the shared class k_s in $C_S \cap C_T$, its weight $\beta[k_s]$ should be relatively larger than the $\beta[k_o]$ of the outlier anatomy class k_o, where k_o belongs to the set $C_S \setminus C_T$. The reason is due to the fact that $C(F(E(x_T^i)))$ are the class predictions of the target domain from the shared class categories, which should have higher probabilities and higher values in relevant positions of β.

Using β as the weighting vector with the previous loss functions $\mathcal{L}_C$ and $\mathcal{L}_D$ leads to two new loss functions, selectively focusing on the anatomy classifier C and domain classifier D on the samples from $C_S \cap C_T$, as follows:

$$\mathcal{L}_C = \frac{1}{N_S} \sum_{j=1}^{N_S} \beta_{y_s^j} \mathcal{L}_C(C(F(E(x_s^j))), y_s^j) \tag{6}$$

$$\mathcal{L}_D = -\frac{1}{N_S} \sum_{j=1}^{N_S} \beta_{y_s^j} log(D(F(E(x_s^j)))) - \frac{1}{N_T} \sum_{i=1}^{N_T} log(1 - D(F(E(x_T^i)))). \tag{7}$$

3 Experiment and Results

3.1 Implementations

We used the ResNet50 as the encoder design and the F is a Fully-Connect (FC) layer with the output dimension of 1024. The domain discriminator consists of three FC layers, with the hidden layer sizes of 512 and 512 respectively. The global MI discriminator consists of two FC layers and the local MI discriminator uses a two single 1x1 convolutional layer. When updating the classifier weight β, it is performed after each epoch.

Two datasets were used for this study to evaluate the CCAN. The first free-hand dataset (50-subjects) was acquired during a fetal anomaly examination using a GE Voluson E8 with a convex C2-9-D abdominal probe, carried out at the maternity ultrasound unit, Oxford University Hospitals NHS Foundation Trust, Oxfordshire, United Kingdom. The four class labels “Heart”, “Skull”, “Abdomen” and “Background”, with 78685, 23999, 51257 and 71884 video frames, respectively, were obtained and annotated by experienced senior reporting sonographers. The single-sweep POCUS dataset were acquired using a Philips HD9 with a V7-3 abdominal probe, which follows the single-sweep scanning protocol [1] also shown in Fig.1(b). The single-sweep dataset were also labeled with the same four classes, with 4136, 5632, 10399 and 17393 video frames respectively. It is important to note that the background class is clinically defined as image frames obtained during the scouting at the beginning of the procedure and transition periods between localizing other anatomical regions of interest (i.e. the foreground classes). Therefore, the background classes in the free-hand dataset may contain different anatomical contents. The target domain single-sweep dataset was split into 70% training, 10% evaluation and 20% unseen test datasets, whilst the 80% and 20% of the source domain free-hand dataset were used for training and evaluation, respectively. The mean recognition accuracy is reported on the test target domain data. In addition, the A-distance is also reported to measure the distribution discrepancy (d_A)[3]. The smaller the A-distance, the more domain-invariant the features are, in general, which suggests a better adaptation method to reduce the cross-domain divergence.

3.2 Cross-Device Adaptation Results

To evaluate the ability of CCAN to reduce cross-device variation, we compare CCAN with the method trained using source-domain data only and the benchmark domain adversarial neural network (DANN) [5]. In this work, the ResNet-50 network was pre-trained using ImageNet and fine-tuned using the source domain free-hand dataset only, which was used as the *No-Adaptation* model. For comparison, the no-adaptation recognition accuracy was obtained by directly classifying the single-sweep dataset using this model. For the three compared models, we perform two adaptation tasks: ***Exp1)*** using all samples from three anatomical classes, excluding the back-ground classes in both datasets (as they may contain different anatomical features described in section 1 and 3); ***Exp2)*** using all samples from the four classes in both source and target datasets.

As shown in Table 1, compared to the no-adaptation model, statistically significant improvement (on an average of 23%) in recognition accuracy was observed by using the adaptation techniques (DANN and CCAN), with both p-values<0.001 based on a pairwise Wilcoxon signed-rank test at a significance level of $\alpha = 0.05$ (used for all the p-values reported in this study unless otherwise indicated). Compared with the DANN, our CCAN increased the mean recognition accuracy by 9.3% in Exp1 and in particular, when considering the challenging 4-class adaptation, CCAN outperforms DANN by 19.0%. These two results are both statistically significant, with the p-values<0.001.

The confusion matrices of CCAN using four-class and three-class adaptation are shown in Table 2, summarizing the recognition accuracies on a per-class basis. In ***Exp2***'s confusion matrix, recognition accuracies of 26.9% and 57.6% were obtained for the fore-ground heart and abdominal classes, respectively. We hypothesized that it is challenging for both DANN and CCAN that the images from free-hand dataset back-ground class may include more diverse and unknown anatomical structures compared to the background class for the single-sweep dataset, as shown in Fig.1(a) and section 3.1. This also motivated the selective adaptation module proposed in section 2.2, with its results presented in section 3.3.

3.3 Cross-Anatomy Adaptation Results

As described in section 2.2, the proposed CCAN-β reduces cross-anatomy variations, such that a network trained with a source domain dataset can be adapted to a target domain dataset with only a subset of the classes defined in the source domain. In our application, the free-hand dataset with the four classes were used in the training, while two experiments in which the subset classes from the single-sweep dataset were tested. The first experiment (***Exp3***) uses three foreground classes of heart, abdominal and skull and the second experiment (***Exp4***) uses two classes of heart and background in the single-sweep dataset. The recognition accuracies are compared in the two experiments, between the four models, the no-adaptation, the DANN, the CCAN and the CCAN-β with the selective adaptation module.

Tables 1 (c),(d) summarises the results from Exp3 and Exp4. Without the selective adaptation module, CCAN significantly outperforms the no-adaptation and the DANN results by 11.7% and 6.2%, respectively in Exp3 (p-values<0.001). The outperformance of CCAN-β was further boosted to 27.2% using the proposed selective adaptation module, achieving a mean recognition accuracy of 73.5%. The results from Exp4 are summarized in Table 1 (d), which indicate statistically significant outperformances of 6.9%, 9.7% and 12.5% using CCAN-β in mean recognition accuracy (p-value<0.001), compared to CCAN, DANN and no-adaptation respectively.

Table 1: Recognition rates (%), statistics and A-distance (d_A) of cross-device adaptations in (a) and (b), cross-anatomy adaptations in (c) and (d).

(a) Exp1

Methods	Acc.	Median,[1^{st},3^{rd}] Quartile	d_A
No-Adaptation	60.9	73.7,[25.1,100.0]	1.96
DANN [5]	79.2	88.9,[25.1,100.0]	1.85
CCAN(Ours)	**88.5**	95.2,[73.4, 100.0]	1.64

(b) Exp2

Methods	Acc.	Median,[1^{st},3^{rd}] Quartile	d_A
No-Adaptation	55.0	0.0,[0.0,61.5]	1.98
DANN [5]	62.6	14.0,[4.0,79.7]	1.94
CCAN(Ours)	**81.6**	76.3,[39.6, 97.8]	1.77

(c) Exp3

Methods	Acc.	Median,[1^{st},3^{rd}] Quartile	d_A
No-Adaptation	34.6	6.4,[0.0,94.8]	1.99
DANN [5]	40.1	0.0,[0.0,94.8]	2.00
CCAN(Ours)	46.3	26.2,[0.0,72.5]	1.95
CCAN-β(Ours)	**73.5**	96.1,[23.0, 100.0]	1.88

(d) Exp4

Methods	Acc.	Median,[1^{st},3^{rd}] Quartile	d_A
No-Adaptation	77.1	100.0,[0.0,100.0]	1.84
DANN [5]	78.9	94.5,[42.6,100.0]	1.85
CCAN(Ours)	82.5	100.0,[46.0, 100.0]	1.79
CCAN-β(Ours)	**89.6**	98.7, [79.3, 100.0]	1.55

Table 2: Confusion matrix (%) of using CCAN for Exp1 (88.5%), Exp2 (81.6%) and Exp3 (73.5%). H, A, S, B stand for heart, abdomen, skull and background.

(a) Exp1

		Predicted Class		
		H	A	S
Actual	H	77.4	22.6	0.0
	A	12.2	84.5	3.3
	S	0.3	0.6	99.1

(b) Exp2

		Predicted Class			
		H	A	S	B
Actual	H	26.9	4.2	0.4	68.5
	A	0.3	57.6	4.7	47.4
	S	0.0	0.0	91.8	8.2
	B	0.0	0.0	0.7	99.3

(c) Exp3

		Predicted Class			
		H	A	S	B
Actual	H	95.6	0.0	0.0	4.4
	A	66.4	29.0	2.3	2.3
	S	5.0	0.0	92.3	2.7

Table 2 (c) reports the per-class accuracies of Exp3 and we can observe a considerable impact due to the anatomy variation being selectively adapted by the proposed β module. For example, the misclassifications of the three anatomical classes to the outlier anatomy class (the background class in free-hand dataset) are below 5%. Still, we can observe that 66.4% of abdomen class samples are misclassified to the heart class, potentially due to very similar image appearance between the abdomen in the single-sweep dataset and the heart in the free-hand dataset.

4 Conclusions

In this paper, we presented a cross-device and cross-anatomy adaptation network to classify an unlabelled single sweep video dataset guided by knowledge of a labelled freehand scanning protocol video dataset. The proposed novel CCAN approach significantly improved the automated image annotation accuracy on the single sweep video dataset, compared to the benchmark domain adaptation methods, by reducing both the cross-device and cross-anatomy variations between the clinical dataset domains.

Acknowledgement The authors gratefully acknowledge the support of EPSRC grants EP/R013853/1, EP/M013774/1 and the NIHR Oxford Biomedical Research Centre.

References

1. Abuhamad, A., Zhao, Y., Abuhamad, S., Sinkovskaya, E., Rao, R., Kanaan, C., Platt, L.: Standardized six-step approach to the performance of the focused basic obstetric ultrasound examination. American journal of perinatology **2**(01), 090–098 (2016) 2, 6
2. Belghazi, M.I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, R.D.: Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018) 4
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1), 151–175 (2010) 6
4. Chen, Q., Liu, Y., Wang, Z., Wassell, I., Chetty, K.: Re-weighted adversarial adaptation network for unsupervised domain adaptation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7976–7985 (2018) 2
5. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016) 2, 4, 7, 8
6. Gao, Y., Noble, J.A.: Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 305–313. Springer (2017) 2
7. van den Heuvel, T.L., Petros, H., Santini, S., de Korte, C.L., van Ginneken, B.: Combining automated image analysis with obstetric sweeps for prenatal ultrasound imaging in developing countries. In: Imaging for Patient-Customized Simulations and Systems for Point-of-Care Ultrasound, pp. 105–112. Springer (2017) 2
8. Kirwan, D.: NHS Fetal Anomaly Screening Programme: 180 to 20+ 6 Weeks Fetal Anomaly Screening Scan National Standards and Guidance for England. NHS Fetal Anomaly Screening Programme (2010) 2
9. Maraci, M.A., Yaqub, M., Craik, R., Beriwal, S., Self, A., von Dadelszen, P., Papageorghiou, A., Noble, J.A.: Toward point-of-care ultrasound estimation of fetal gestational age from the trans-cerebellar diameter using cnn-based ultrasound image analysis. Journal of Medical Imaging **7**(1), 014501 (2020) 2
10. Maraci, M.A., Bridge, C.P., Napolitano, R., Papageorghiou, A., Noble, J.A.: A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. Medical image analysis **37**, 22–36 (2017) 2