


BiometryNet: Landmark-based Fetal Biometry Estimation from Standard Ultrasound Planes

Netanell Avisdris ^{1,2}[0000-0002-0719-1327], Leo Joskowicz¹, Brian Dromey^{4,6}, Anna L. David⁶, Donald M. Peebles⁶, Danail Stoyanov^{4,5}, Dafna Ben Bashat^{2,3}, and Sophia Bano^{4,5}

- ¹ School of Computer Science and Engineering, The Hebrew U. of Jerusalem, Israel
`{netana03,josko}@cs.huji.ac.il`
- ² Sagol Brain Institute, Tel Aviv Sourasky Medical Center, Israel
- ³ Sagol School of Neuroscience and Sackler Faculty of Medicine, Tel Aviv U., Israel
- ⁴ Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, UK
- ⁵ Department of Computer Science, University College London, UK
- ⁶ Elizabeth Garrett Anderson Institute for Women's Health, U. College London, UK

Abstract. Fetal growth assessment from ultrasound is based on a few biometric measurements that are performed manually and assessed relative to the expected gestational age. Reliable biometry estimation depends on the precise detection of landmarks in standard ultrasound planes. Manual annotation can be time-consuming and operator dependent task, and may result in high measurement variability. Existing methods for automatic fetal biometry rely on initial automatic fetal structure segmentation followed by geometric landmark detection. However, segmentation annotations are time-consuming and may be inaccurate, and landmark detection requires developing measurement-specific geometric methods. This paper describes BiometryNet, an end-to-end landmark regression framework for fetal biometry estimation that overcomes these limitations. It includes a novel Dynamic Orientation Determination (DOD) method for enforcing measurement-specific orientation consistency during network training. DOD reduces variabilities in network training, increases landmark localization accuracy, thus yields accurate and robust biometric measurements. To validate our method, we assembled a dataset of 3,398 ultrasound images from 1,829 subjects acquired in three clinical sites with seven different ultrasound devices. Comparison and cross-validation of three different biometric measurements on two independent datasets shows that BiometryNet is robust and yields accurate measurements whose errors are lower than the clinically permissible errors, outperforming other existing automated biometry estimation methods. Code is available at <https://github.com/netanellavisdris/fetalbiometry>.

Keywords: Fetal biometry estimation · Fetal ultrasound · Anatomical landmarks' localisation · Computer-assisted diagnosis.

1 Introduction

Ultrasound (US) based estimation of fetal biometry is widely used to monitor fetal growth assessed relative to the expected gestational age and to diagnose prenatal abnormalities. Reliable estimation of fetal biometry depends on the localization of fetal landmarks on standard planes (SPs) by the sonographer. Common SPs include the trans-ventricular plane to measure the fetal head circumference, the trans-abdominal plane to measure the fetal abdominal circumference, and femoral plane to measure the femur length. Obtaining manual measurements can be time-consuming and operator-dependent, especially for trainees [9], and results in high inter- and intra-operator variability [13]. In fetal biometry, inter-operator measurements variability range between $\pm 4.9\%$ and $\pm 11.1\%$ and an intra-operator variability between $\pm 3\%$ and $\pm 6.6\%$ [19]. This variability contributes to biometry uncertainty and hampers fetal growth evaluation in the clinic. Automating fetal biometry can reduce measurement variability and improve fetal monitoring and clinical decision making.

Numerous methods have been developed for automatic computation of different measurements in US-based fetal biometry [21], e.g., biparietal diameter [1,24], head circumference [12], and femur length [14]. Recently, methods for comprehensive biometry required for fetal weight estimation have been proposed [6,16]. Similar methods have been developed for other imaging modalities, e.g., fetal MRI [5] and 3D US [17]. These methods rely on obtaining segmentation masks to train fetal anatomy segmentation models, which is time-consuming, and on developing geometric methods for each measurement.

An alternative approach is the detection of landmarks directly on the image without relying on segmentation masks. The advantages of this approach are that it follows the clinical workflow – the sonographer locates on the two landmarks required for biometric measurement directly on the image, and that obtaining ground-truth data is fast and straightforward. Landmark-based approaches have been proposed for various medical imaging modalities [23], e.g. CT and MRI, to quantify lesions progression [20], for fetal MRI biometry estimation [4], and for fetal landmark detection in US using reinforcement learning agents [2]. However, this approach has not been used for automating fetal biometry in US SPs.

Automating fetal US biometry presents numerous challenges. First, imaging quality varies across different US devices, heterogeneous in overall appearance, particularly regarding gain and zoom. Second, some fetal anatomical structures may become difficult to image with increasing gestational age. Third, anatomical structure position, size, orientation, and appearance present significant variability (Fig. 1). Since the fetus can lie in any of a wide variety of positions, anatomical landmarks can be in orientations that are off the horizontal. Finally, biometric parameters have different geometric characteristics and performed on different SPs, thus each requires its own individual method.

We propose BiometryNet, an end-to-end framework that automates US fetal biometry for multiple fetal structures using direct landmark detection that overcomes these challenges. Unlike other methods [6,1,17,24,21], BiometryNet

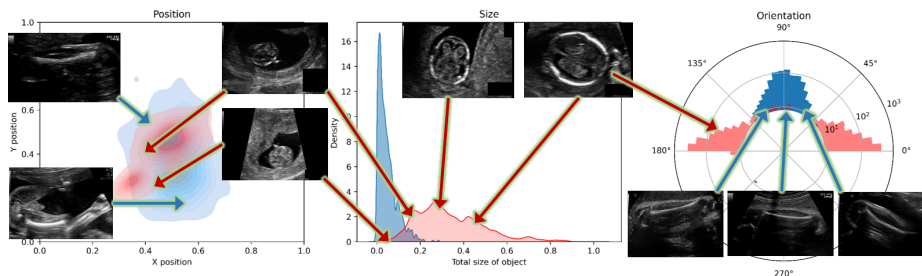


Fig. 1. Fetal biometry variability for the head (red) and femur (blue) SPs. The graphs show (left to right) the distributions of the position (structure center point), size (structure area with respect to image area) and orientation (structure angle with respect to horizontal plane). Sample US image inserts illustrate the distribution.

only requires measurement landmark annotations for training. The main contributions of our work are:

- We show for the first time that US fetal biometry can be automated with high accuracy using only landmark annotations that can be acquired easily. We demonstrate our approach on a large fetal biometry dataset of 3,398 US images from 1,820 subjects acquired by various operators and US devices at multiple clinical centers.
- We propose a novel Dynamic Orientation Determination (DOD) method to determine measurement-wise orientation and provide consistent landmark class for various measurements. We show that DOD is robust for variabilities in the SPs orientations, leading to improved biometry estimates and generalization on unseen datasets.

2 Method

We propose BiometryNet, a framework for the estimation of fetal US biometry using a landmark regression convolution neural network (Fig. 2). BiometryNet locates two landmark points on an US image for three biometric parameters: biparietal diameter (BPD) and occipito-frontal diameter (OFD) in the trans-ventricular (head) plane, and femur length (FL) in the femur plane. It includes Dynamic Orientation Determination (DOD) used during training for consistent landmark class. At inference time, the trained BiometryNet predicts two biometry landmarks, followed by scale recovery in order to estimate image resolution needed to compute the actual biometric measure.

2.1 Landmark Regression Network

The landmark regression network is a modified HRNet [22], trained to predict a heat map for each landmark defined by a Gaussian function centered at the

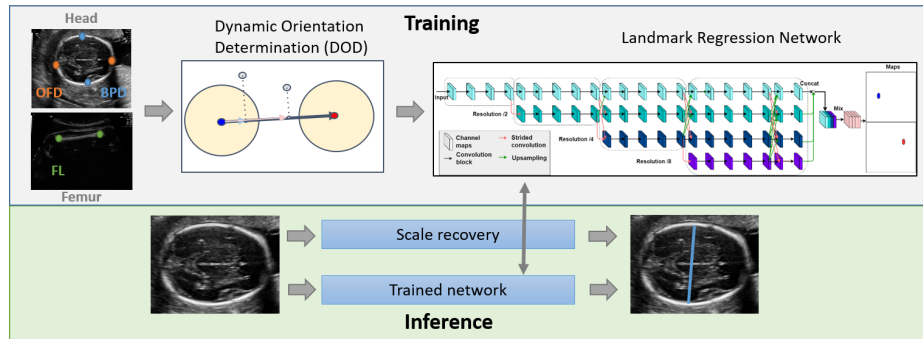


Fig. 2. BiometryNet framework. Top: during training, annotated head and femur planes are fed into the DOD module and to the landmark regression network to predict two landmarks per biometric measurement. Bottom: during inference, the trained model predicts the landmarks followed by scale recovery for biometric measurements estimation.

landmark coordinates whose co-variance describes the landmark location uncertainty. We used HRNet since it has achieved state-of-the-art performance for computer vision tasks, e.g. object detection, semantic segmentation, face landmark detection and human pose estimation. HRNet is a Convolutional Neural Network that combines the representations of multi-scale high-to-low resolution parallel streams into a single stream. The representations are then input to a two-layer convolution classifier. The first layer combines the feature maps of all four resolutions; the second layer computes a Gaussian heat map for each of the two landmarks. One network is trained for each biometric measurement (two landmarks points) with the Mean Squared Error loss between the Gaussian maps created from the ground-truth landmarks and the predicted heat maps. At inference time, the two landmarks' locations are defined by the coordinates of the pixel with the maximal value on each heat map.

To compensate for the high variability in acquisition orientation and object scales in fetal US images (Fig. 1), two training time augmentations are used: 1) rotations around the image center at randomly sampled angles in the range $[-180, 180]^\circ$ followed by cropping to preserve a fixed image size; 2) image scaling at randomly sampled scales in the range $[-5, +5]\%$.

2.2 Dynamic Orientation Determination (DOD)

Since fetal structure may appear in various orientations (Fig. 1) and CNN are not rotation-invariant, orientation variability must be handled properly. A common approach to handle such variability is augmentation. However, rotation augmentation may cause landmark class labeling (e.g left/right landmarks) to be inconsistent with image coordinates, i.e. the left and right points may be erroneously switched, which will hamper network training (Fig. 1 Supp.). This

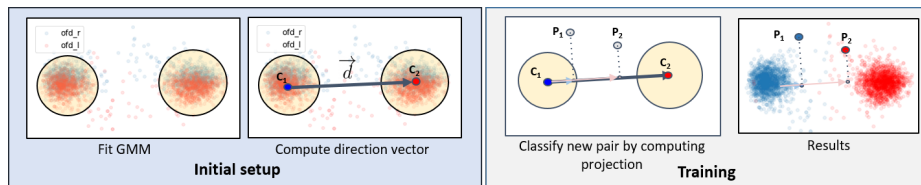


Fig. 3. Illustration of Dynamic Orientation Determination (DOD). Left: *Initial setup stage, measurement direction vector computation between GMM centroids fitted to normalized landmark points*; Right: *Training stage, re-assignment of each landmark pair class by projecting the landmarks on the direction vector and ordering them*.

inconsistency can be corrected by landmark class reassignment (LCR) [4] to preserve horizontal (left/right) landmark class consistency after augmentation. However, different biometric measurements may have different spatial orientation, e.g., OFD is mostly vertical and BPD is mostly horizontal.

To overcome this issue, we introduced Dynamic Orientation Determination (DOD), a method that determine measurement-wise orientations and perform class reassignment accordingly, instead of computing only the horizontal measurement landmark pairs as in [4]. DOD consists of two stages (Fig. 3). In *initial setup* stage, biometry orientation is determined. First, Gaussian Mixture Model of two Gaussians was fitted onto the ground-truth landmarks of training dataset of each biometric measurements with Expectation-Maximization algorithm [8]. Next, biometry orientation is computed as directional vector between the two Gaussian centroids, $\vec{d} = \vec{C}_1\vec{C}_2$. In the network *training* stage, the learned orientation is used to enforce consistency. After all augmentations are performed, each resulting biometric measurement landmark pair ($P = p_1, p_2$) is projected on the directional vector $r_i = (p_i \cdot \vec{d})/|\vec{d}|$ and then ordered according to its projection $sort(|r_i|)$ to obtain their reassigned class.

2.3 Scale Recovery

To obtain actual measurements that can be compared to those obtained in the clinic, scale conversion from pixel to millimeter units is required. While this information is usually available during examination or is embedded in the raw image data, some retrospectively collected images may lack it. Therefore, we perform scale recovery using the approach presented in [6]. Briefly, it recovers true scale by detecting ruler markers using template matching.

3 Experimental Setup

Datasets and Annotations: we use two public datasets, Fetal Planes (FP) [7] and HC18 [12]. FP [7] was originally designed for US SPs classification challenge. The US images were acquired on six US devices: three GE Voluson E6, one Voluson S8, one Voluson S10 and one Aloka at two clinical sites in Barcelona, Spain.

Since not all images in FP qualified as SPs for fetal biometry [18], we selected 1,638 (909 subjects) fetal head and 761 (630 subjects) fetal femur SPs. Not all subjects are included in both planes datasets, resulting in a total of 2,399 images from 1,014 unique subjects. An obstetrician then manually annotated the landmarks on each image with the VIA annotation tool [10]. On average, each plane annotation for landmarks took 20 seconds, which is far less than 70 seconds required for manual structure delineation. HC18 [12] was designed for fetal head circumference (HC) challenge. The US images were acquired with two US devices: GE Voulson E8 and 730, in one clinical site in the Netherlands. All HC18 training set, 999 (806 subjects) fetal head SPs, were annotated with a HC measurement. We computed the BPD and OFD biometric measurements from the major and minor axes of an ellipse by least-square fitting [11] onto the ground-truth mask.

Evaluation Metrics: we use the mean and median of L_1 difference, bias and agreement. For two sets of n biometric measurements, $M_1 = \{m_i^1\}$, $M_2 = \{m_i^2\}$, let m_i^1 and m_i^2 ($1 \leq i \leq n$) be two measurement values, the ground-truth and the computed one, respectively. The difference between each pair of measurements is defined as $d_i = m_i^1 - m_i^2$. The mean and median differences for M_1, M_2 are defined as $\overline{L}_1(M_1, M_2) = 1/n \sum_{i=1}^n |d_i|$ and $\widetilde{L}_1(M_1, M_2) = \text{median}_i(|d_i|)$, respectively. We use the Bland-Altman method [3] to estimate the bias and agreement between two biometry measurement sets. Agreement is defined by the 95% confidence interval $CI_{95}(M_1, M_2) = 1.96 \times \sqrt{1/n \sum_{i=1}^n (\overline{L}_1(M_1, M_2) - d_i)^2}$. The measurements bias is defined as $Bias(M_1, M_2) = 1/n \sum_{i=1}^n d_i$.

Study Setup: the two datasets FP and HC18 were split into training and test sets. For the FP head and femur SPs, we selected 757 (449 subjects) and 437 (368 subjects) images for training and 881 (460 subjects) and 324 (262 subjects) images for testing, respectively. For a fair comparison of head planes in FP and HC18, we selected from HC18, similar number of 737 images (600 subjects) for training. The remaining 262 images (206 subjects) were used for the test set. BiometryNet was implemented in PyTorch and trained for 200 epochs in about 2 hours on a single NVIDIA 1080Ti GPU with a batch size of 16 using the ADAM optimizer [15], an initial learning rate of 10^{-4} and a drop factor of 0.2 in epochs 10, 40, 90, and 150. For validation, a comparison to four networks was performed: (a) Vanilla HRNet [22]; (b) HRNet with fixed horizontal orientation determination; (c) HRNet with fixed vertical orientation determination for biometric measurements with horizontal orientation (OFD, FL) and vertical orientation (BPD); and, (d) FMLNet [4], which is an HRNet with fixed horizontal orientation determination, test-time augmentation and a method for prediction reliability estimation.

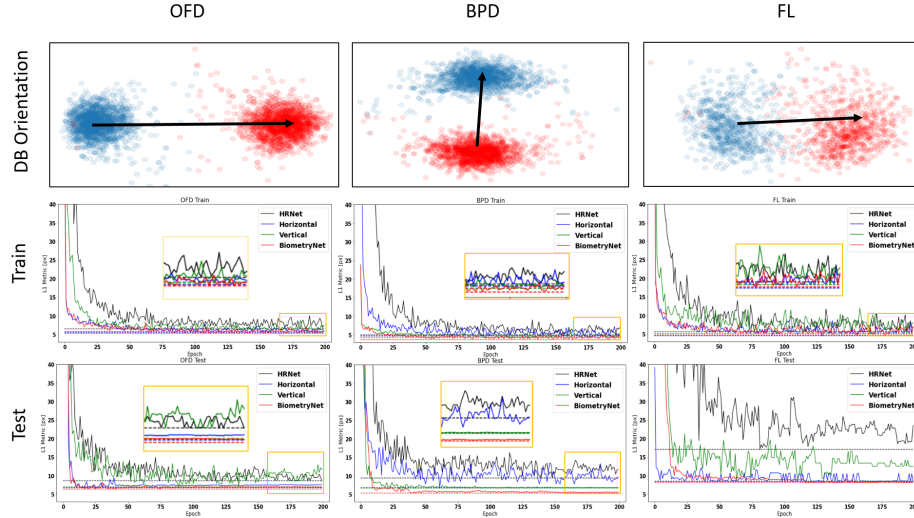


Fig. 4. Effect of learned dynamic orientation (first row) on training (2nd row) and inference (3rd row) for three measurements (columns) on convergence (yellow boxes): OFD, BPD and FL, in the FP dataset. Four models were trained: *HRNet* (black), with *horizontal* (blue) fixed orientation, *vertical* (green) fixed orientation and *BiometryNet* with DOD (red). Dotted lines denote the best performing epoch metric for each model.

4 Results

We conducted three experimental studies on the FP and HC18 datasets to quantify the performance of BiometryNet and compare it to other existing methods.

Study 1: Effect of DOD on training: we evaluated the effect of DOD on training and inference dynamics on three biometric measurements: OFD, BPD, and FL in the FP dataset. OFD and FL have mostly horizontal orientations and BPD has mostly vertical orientation. Fig. 4 shows the results. Note that for all biometric measurements, the learned orientation is similar to the preferred orientation. BiometryNet converged faster and performed better than networks trained on fixed orientations of both the training and test sets, similar to those with the preferred fixed orientation, and better than the Vanilla HRNet, which performed poorly. This advantage becomes more evident when the landmarks' locations are more disperse, e.g. for FL. This shows that DOD provides robustness with respect to orientation variability and yields an improved performance.

Study 2: Performance comparison: we evaluate the performance of BiometryNet on the FP and HC18 datasets. Table 1 lists the results. HRNet performed poorly on all biometric measurements, with a high CI_{95} and error. Fixed orientation determination yielded better results in preferred orientation of measurement, i.e. horizontal in OFD and FL, and vertical in BPD. FMLNet [4]

Table 1. Study 2 and 3 results for head (OFD, BPD) and femur (FL) biometric measurements for the FP and HC18 datasets. For each biometric measurement, bias, agreement CI_{95} , and mean \bar{L}_1 and median L_1 differences with respect to expert annotations are listed. The best performance for each metric and dataset is indicated in Bold. (*) indicates only images with reliable predictions were included in FMLNet evaluation.

Train DB	Test DB	Method	Head - OFD				Head - BPD				Femur - FL			
			Bias [mm]	CI_{95} [mm]	\bar{L}_1 [mm]	L_1 [mm]	Bias [mm]	CI_{95} [mm]	\bar{L}_1 [mm]	L_1 [mm]	Bias [mm]	CI_{95} [mm]	\bar{L}_1 [mm]	L_1 [mm]
FP	FP	HRNet [22]	6.23	26.40	6.30	3.30	2.84	22.57	3.20	0.80	1.80	18.40	2.70	0.62
		Horizontal	2.65	10.23	2.87	1.90	2.36	21.60	2.78	0.76	0.17	3.27	0.99	0.59
		Vertical	4.73	23.51	4.86	2.46	0.77	8.28	1.28	0.65	0.33	10.27	1.47	0.65
		FMLNet* [4]	1.96	7.80	2.16	1.43	1.30	14.56	1.71	0.65	0.14	3.00	1.02	0.68
		BiometryNet	0.21	2.75	1.01	0.71	0.04	2.50	0.77	0.58	0.18	3.03	0.97	0.62
	HC18	BiometryNet	2.31	5.21	2.46	1.85	0.84	2.70	1.06	0.91				
HC18	HC18	HRNet [22]	0.64	6.01	1.51	0.92	2.64	21.48	3.10	0.71				
		Horizontal	2.82	23.9	3.69	0.93	1.35	12.86	1.75	0.59				
		Vertical	4.02	29.15	4.92	0.97	0.50	5.13	0.98	0.65				
		FMLNet* [4]	2.23	17.48	2.61	1.02	0.73	4.00	0.93	0.64				
		BiometryNet	0.56	4.43	1.39	0.84	0.16	3.54	0.88	0.61				
	FP	BiometryNet	-3.24	6.01	3.54	2.72	-1.11	3.35	1.40	1.08				

performed better than fixed horizontal in all cases, but at the expense of discarding 8% of test inputs in average. BiometryNet outperformed other methods in OFD and BPD biometric measurements in all metrics. While FL has a slightly better bias and CI_{95} in FMLNet, considering only included cases by all methods, the paired t-test ($p = 0.02$) shows that BiometryNet outperforms all other methods, including FMLNet.

BiometryNet yields a median error ($< 0.84\text{mm}$ (OFD), $< 0.61\text{mm}$ (BPD) and $< 0.62\text{mm}$ (FL)) that is better than the error reported in [6] (1.30mm (OFD), 0.80mm (BPD) and 2.1mm (FL)). In addition, the results are better than [1], which was the best BPD performer in a 2022 review on fetal US biometry [21], that achieved a mean error of 2.33mm, bias of 1.49mm and CI_{95} of 5.55mm. Furthermore, the variance of all methods except BiometryNet is higher than the reported inter-observer variability [24] of 5.0mm (OFD), 3.0mm (BPD) and 4.3mm (FL) [19]. These results suggest that overall BiometryNet performs better and is more stable (lower variance) than the existing methods.

Study 3: Impact of training dataset: to demonstrate the generalization capabilities of BiometryNet, we analyzed its performance by cross-validation on unseen datasets. For this purpose, we train the OFD and BPD models on FP dataset and test each on HC18 dataset, and vice versa. Table 1 lists the results. Note that training on one dataset and testing on the other (rows 6 and 12) results in high mean and median errors, with significant and complementary (2.31mm vs -3.24mm in OFD, 0.81mm vs -1.11mm in BPD) bias and acceptable variance. In contrast, testing on the same dataset (rows 5 and 11) results in low bias, but similar variance (CI_{95}) as before. This bias can be explained by the differences in annotation protocol between the two datasets. In HC18, the landmark annotations

lie between the outer contours of the skull, while in FP, they are marked in the middle of the fetal skull contour, thus resulting in a consistent bias (Fig. 2 supp.). Clinically, both protocols are acceptable, and their selection depends on the specific clinical site. We also observe that using FP annotations for network training produced better and more consistent results than using those of HC18. This may occur because HC18 annotations are extracted from HC ellipse rather than annotated directly at the BPD/OFD landmarks. We conclude from these results that BiometryNet is capable of generalization, as it can learn to annotate like the annotation protocol, and can be tuned for use across many sites and protocols.

5 Conclusions

We proposed BiometryNet, an end-to-end network for automatic fetal biometry estimation from standard US planes. BiometryNet only required anatomical landmarks annotations for biometric measurements prediction without the need of any geometric methods as post-processing. To overcome the variability that is inherently present in the acquired US planes, we introduced a novel dynamic orientation determination mechanism which enforced measurement-wise orientation consistency in network training. This resulted in reduced variability and improved landmarks' localization, thus leading to more accurate biometric measurements compared to the state-of-the-art methods. Through the analysis of two large and independent datasets, we demonstrated the generalization ability of our proposed method. Moreover, BiometryNet resulted in minimal error which is lower than the clinically permissible error [19], thus showing the potential for clinical translation to improve fetal growth assessment.

In future work, we plan to analyze the abdominal plane and its measurement using BiometryNet. Furthermore, BiometryNet assumes SP is already available for the measurement, which is not always the case in clinical settings. Therefore, we will design a holistic network to jointly learn SP selection and fetal biometry.

Acknowledgements This research was partly supported by the Wellcome/EP-SRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; the Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; the Royal Academy of Engineering Chair in Emerging Technologies Scheme, and Horizon 2020 FET Open [863146]; Kamin Grants [63418, 72126] from the Israel Innovation Authority.


References

1. Al-Bander, B., Alzahrani, T., Alzahrani, S., Williams, B.M., Zheng, Y.: Improving fetal head contour detection by object localisation with deep learning. In: Annual Conference on Medical Image Understanding and Analysis. pp. 142–150. Springer (2019)
2. Alansary, A., Oktay, O., Li, Y., Le Folgoc, L., Hou, B., Vaillant, G., Kamnitsas, K., Vlontzos, A., Glocker, B., Kainz, B., et al.: Evaluating reinforcement learning

- agents for anatomical landmark detection. *Medical image analysis* **53**, 156–164 (2019)
3. Altman, D.G., Bland, J.M.: Measurement in medicine: the analysis of method comparison studies. *J of the Royal Statistical Society: Series D (The Statistician)* **32**(3), 307–317 (1983)
 4. Avisdris, N., Ben Bashat, D., Ben-Sira, L., Joskowicz, L.: Fetal brain mri measurements using a deep learning landmark network with reliability estimation. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. pp. 210–220. Springer (2021)
 5. Avisdris, N., Yehuda, B., Ben-Zvi, O., Link-Sourani, D., Ben-Sira, L., Miller, E., Zharkov, E., Bashat, D.B., Joskowicz, L.: Automatic linear measurements of the fetal brain with deep neural networks. *Int J Comput Assist Radiol Surg* (2021)
 6. Bano, S., Dromey, B., Vasconcelos, F., Napolitano, R., David, A.L., Peebles, D.M., Stoyanov, D.: Autofb: Automating fetal biometry estimation from standard ultrasound planes. In: *Proc Int Conf on Medical Image Computing and Computer Assisted Intervention*. pp. 228–238. Springer (2021)
 7. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Sci. Rep.* **10**(1), 1–12 (2020)
 8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *J Royal Stat Society: Series B* **39**(1), 1–22 (1977)
 9. Dromey, B.P., Ahmed, S., Vasconcelos, F., Mazomenos, E., Kumpulainen, Y., Ourselin, S., Deprest, J., David, A.L., Stoyanov, D., Peebles, D.M.: Dimensionless squared jerk: An objective differential to assess experienced and novice probe movement in obstetric ultrasound. *Prenatal Diagnosis* (2020)
 10. Dutta, A., Zisserman, A.: The via annotation software for images, audio and video. In: *Proc. of the 27th ACM international conference on multimedia*. pp. 2276–2279 (2019)
 11. Fitzgibbon, A.W., Pilu, M., Fisher, R.B.: Direct least squares fitting of ellipses. In: *Proc Int Conf on Pattern Recognition*. vol. 1, pp. 253–257. IEEE (1996)
 12. van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., Ginneken, B.v.: Automated measurement of fetal head circumference using 2d ultrasound images. *PloS One* **13**(8), e0200412 (2018)
 13. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in ct. *European radiology* **29**(3), 1391–1399 (2019)
 14. Khan, N.H., Tegnander, E., Dreier, J.M., Eik-Nes, S., Torp, H., Kiss, G.: Automatic detection and measurement of fetal biparietal diameter and femur length—feasibility on a portable ultrasound device. *Open J Obstetr Gynecol* **7**(3), 334–350 (2017)
 15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: *Proc Int Conf Learning Representations* (2015)
 16. Prieto, J.C., Shah, H., Rosenbaum, A.J., Jiang, X., Musonda, P., Price, J.T., Stringer, E.M., Vwalika, B., Stamilio, D.M., Stringer, J.S.: An automated framework for image classification and segmentation of fetal ultrasound images for gestational age estimation. In: *Medical Imaging 2021: Image Processing*. vol. 11596, p. 115961N. International Society for Optics and Photonics (2021)
 17. Ryou, H., Yaqub, M., Cavallaro, A., Papageorghiou, A.T., Noble, J.A.: Automated 3d ultrasound image analysis for first trimester assessment of fetal health. *Phys. Med. Biol.* **64**(18), 185010 (2019)

18. Salomon, L., Winer, N., Bernard, J., Ville, Y.: A score-based method for quality control of fetal images at routine second-trimester ultrasound examination. *Prenatal Diagnosis* **28**(9), 822–827 (2008)
19. Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D., Papageorgiou, A., Fetal, I., for the 21st Century (INTERGROWTH-21st), N.G.C.: Intra-and interobserver variability in fetal ultrasound measurements. *Ultrasound Obstetr Gynecol* **39**(3), 266–273 (2012)
20. Tang, Y., Yan, K., Cai, J., Huang, L., Xie, G., Xiao, J., Lu, J., Lin, G., Lu, L.: Lesion segmentation and recist diameter prediction via click-driven attention and dual-path connection. In: *Proc Int Conf on Medical Image Computing and Computer Assisted Intervention*. pp. 341–351. Springer (2021)
21. Torres, H.R., Morais, P., Oliveira, B., Birdir, C., Rüdiger, M., Fonseca, J.C., Vilaça, J.L.: A review of image processing methods for fetal head and brain analysis in ultrasound images. *Comput. Methods Programs Biomed.* p. 106629 (2022)
22. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell* (2019)
23. Zhang, J., Liu, M., Shen, D.: Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Trans Pattern Anal Mach Intell* **26**(10), 4753–4764 (2017)
24. Zhang, L., Dudley, N.J., Lambrou, T., Allinson, N., Ye, X.: Automatic image quality assessment and measurement of fetal head in two-dimensional ultrasound image. *J Med Imaging (Bellingham)* **4**(2), 024001 (2017)

BiometryNet: Landmark-based Fetal Biometry Estimation from Standard Ultrasound Planes (Supplementary Material)

Netanell Avisdris ^{1,2}, Leo Joskowicz¹, Brian Dromey^{4,6}, Anna L. David⁶, Donald M. Peebles⁶, Danail Stoyanov^{4,5}, Dafna Ben Bashat^{2,3}, and Sophia Bano^{4,5}

¹ School of Computer Science and Engineering, The Hebrew U. of Jerusalem, Israel
{netana03,josko}@cs.huji.ac.il

² Sagol Brain Institute, Tel Aviv Sourasky Medical Center, Israel

³ Sagol School of Neuroscience and Sackler Faculty of Medicine, Tel Aviv U., Israel

⁴ Wellcome/EPSCRC Centre for Interventional and Surgical Sciences(WEISS),
University College London, London, UK

⁵ Department of Computer Science, University College London, London, UK

⁶ Elizabeth Garrett Anderson Institute for Women's Health, University College
London, London, UK

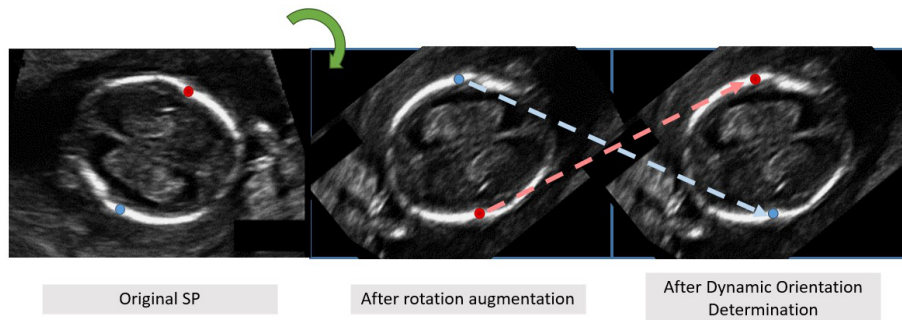


Fig. 1. Dynamic Orientation Determination (DOD): illustrative example. (a) original standard plane (SP) ultrasound (US) image of fetal trans-ventricular plane with the bi-parietal diameter (BPD) biometric measurement up (red) and down (green) ground truth landmarks; (b) SP after rotation showing an inconsistent landmark labeling in which the up and down landmarks are switched; (c) reassignment of the landmark up and down labels by DOD (blue and red intermittent lines).

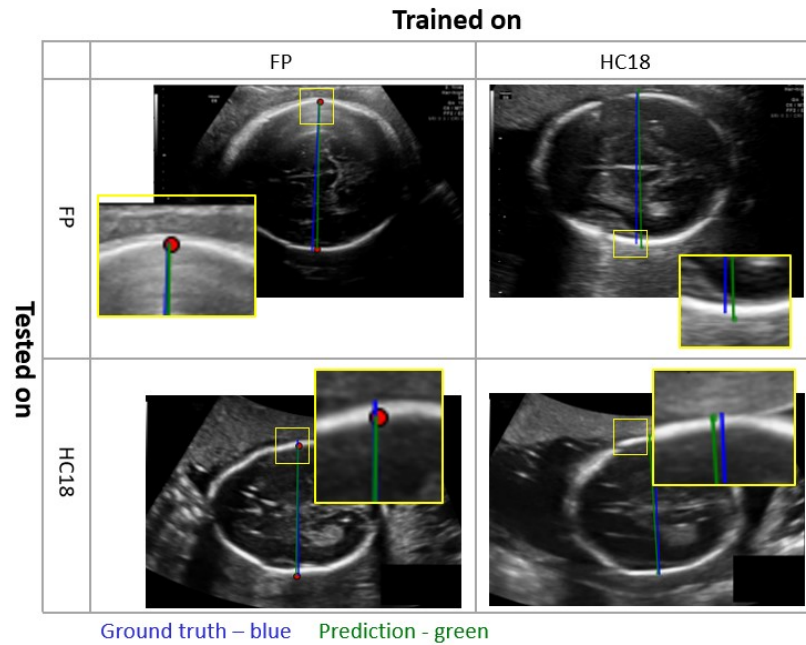


Fig. 2. Example showing different results of two annotation methods and the resulting BiometryNet predictions. Two models were trained, one on the FP training dataset (first column) and the second on the HC18 training dataset (second column). Both were run on the FP testing dataset (first row) and the HC18 testing dataset (second row). Landmarks are marked in the middle of the fetal skull in FP dataset and on the outer extrema of the skull in HC18 dataset. This introduces a bias that results in a relatively larger error, although still within the clinically acceptable limit when generalizing from FP→HC18 or HC18→FP. The quantitative results are presented in Table 1 and discussed in Study 2 and 3. The images show ground truth annotations (blue) and computed predictions (green) on representative examples of BPD biometry. The relevant areas are enlarged in the yellow box.