# Are you sure that medical image segmentation performance is good?

**Author name(s) withheld** EMAIL(S) WITHHELD
*Address withheld*

**Editors:** Under Review for MIDL 2023

## Abstract

Recently, deep learning-based image segmentation has recently developed highly effective and precise approaches, and the majority of them are based on Convolutional Neural Networks and Vision transformers for medical image segmentation, which provide detailed information. In other words, there is a difference between structure detection and segmentation. For instance, a crown-rump-length (CRL) measurement from a segmentation mask provides more information than CRL detection, which can be a significant indicator of possible fetal abnormalities. However, this framework is so sensitive and can be easily affected by the segmentation performance because any incorrect measurement from wrong segmentation results in a wrong medical report. The deep learning-based segmentation models can be sensitive because of the domain shift. Therefore, the main motivation of this study is to make sure that the segmentation performance of the model is good. I'll mention the final results!

**Keywords:** List of keywords, comma separated.

## 1. Introduction

Ultrasound which is one of the safest and noninvasive imaging modalities has long been acknowledged as a powerful and routine screening tool. In particular, due to real-time display, low cost, and its safety, fetal US imaging is widely employed as a first-line visual-based prenatal screening throughout the world to evaluate fetal growth. The researchers mentioned that an early pregnancy ultrasound at 12 to 14 weeks of pregnancy, called specifically a dating scan, is an effective way to screen and detect possible fetal anomaly (Drysdale et al., 2002). They also emphasized that five of the seven abnormalities were found before the 20-week anomaly scan in a pilot study (Drysdale et al., 2002). Basically, the fetal length can be measured during this scan using a crown-rump-length (CRL) view image, which is obtained from the top of the head to the bottom of the rump. This length can be used to calculate gestation age (GA), which helps to understand whether the baby is in the expected growth or is behind or ahead of the expected growth. Due to the fact that this calculation performed by a sonographer or directly calculated via segmentation masks are so sensitive, any incorrect measurement of CRL might cause inaccurate GA calculation, resulting in an incorrect growth report of the fetus. Briefly, precise CRL measurement has a key role in detecting any potential fetal anomalies. How are we sure that the fetal ultrasound image segmentation that we use for CRL measurement, is performed well?

Artificial intelligence (AI) is a new technological field that develops application systems to imitate human intelligence. The use of artificial intelligence in medical research has been a hot issue in modern science and technology for the last five years. Image classification,

© 2023 Anonymous.

segmentation, identification, detection of anomalies, and other elements of fetal ultrasound images require a wide variety of expertise, and AI may help in these areas. For instance, deep learning based-segmentation of anatomical structures such as the fetus (Yang et al., 2019), placenta (Yang et al., 2019; Zimmer et al., 2019; Looney et al., 2018; Schwartz et al., 2022), gestational sac (Yang et al., 2019), fetal CRL (Ryou et al., 2019; Cengiz and Yaqub, 2021), fetal head (Cengiz and Yaqub, 2021), body(Cengiz and Yaqub, 2021), palate (Cengiz and Yaqub, 2021), heart (Nurmaini et al., 2021), femur (Zhu et al., 2021), abdomen (Ravishankar et al., 2016). Most of them reported higher accuracy and precision, recall, F1, and dice scores after segmentation. However, even the greatest approach, though, can occasionally fail because of poor image quality which has noise, and artifacts or might exhibit unexpected behavior on a new and different dataset. For example, an AI-assisted fetal structure segmentation framework can provide higher accuracy in a specific dataset. However, the segmentation output generated from a different dataset acquired from a different machine might not be outstanding because of the domain shift. In this case, all fetal growth reports based on the poor segmentation masks might not show reality. Therefore, it is crucial to be able to recognize such failure segmentation masks on an individual basis in clinical settings, especially AI-assisted screening.

### 1.1. Contributions

We specifically focus on the automatic quality control of CRL segmentation in this study. The main contributions of this paper are as follows:

- We propose a method to understand that the fetal ultrasound image segmentation that we use for CRL measurement, is performed well.
- The proposed method also gives detailed insights into the class-based segmentation quality considering to dice score prediction, in addition to the whole segmentation quality performance.
- We demonstrate the effectiveness of the proposed method in a different dataset acquired from different ultrasound machines in different hospitals to show the power even if domain shift.
- We show how our method outperforms different methods based on the Siamese or Synergic methods. The code will be publicly available.

### 1.2. Related Work

In this section, we review prediction quality of any segmentation task and fetal ultrasound segmentation topics under two topics.

**Fetal Ultrasound Segmentation**

Segmenting or detecting objects in fetal ultrasound images has several advantages such as allowing for better visualisation of structures, measuring structure length, area or volume of clinically important objects, and assessing these structures against clinical guidelines. Segmentation and detection of fetal brain structures are the most widely researched topic in fetal ultrasound such as (Yang et al., 2019; Yaqub et al., 2017; Venturini et al., 2020; Namburete et al., 2018; Yang et al., 2020; Lin et al., 2019; Wyburd et al., 2020; Liu et al.,

2019). These articles have mainly focused on detecting and segmenting fetal brain structures such as the ventricles, cavum septum, cerebellum and the midline falx. Other fetal structures have been investigated in the literature such as the fetal heart (Philip et al., 2019), abdomen (Ravishankar et al., 2016), placenta (Schwartz et al., 2022), fetal head for head circumference (Sobhaninia et al., 2019), and fetal CRL (Ryou et al., 2019; Cengiz and Yaqub, 2021).

**Segmentation Quality**

It can be difficult to guarantee the accuracy of an automatically generated segmentation in an image analysis framework especially if it is real-time deployed or is domain shift. Robinson et al. conducted a study for real-time automated quality control for cardiovascular MR segmentations (Robinson et al., 2018). In 2022, the same group (Li et al., 2022) tried to understand that how the model has confident predictions on unlabelled test data based on the deployment's domain. I can add more if needed...

## 2. Materials and Methods

### 2.1. Dataset and Preprocessing

We conducted our experiments with two different fetal ultrasound datasets acquired from different machine brands at different hospitals. Both datasets include original 2D fetal ultrasound images and masks which were segmented into 4 different structures including the body, head, the gap between the chin and chest, and the fetal palate. The first dataset includes 696 images and masks. Firstly, 9 different segmentation masks were generated by using the ground truth segmentation mask. 5 of these masks represent the segmentation performed poorly, and 4 of them represent the segmentation performed well with very small changes applied to the original segmentation masks.

Figure 1 shows poor segmentation mask examples of the original image. The first row shows the original, over estimation done with over dilation, vertical mirror by flipping of randomly selected label, and the second one shows under estimation done with over erosion, wrong label tag of the selected label, randomly selected label deletion from left to right respectively.

As a result of synthetic data generation, 6970 images were used for training and validation. The second dataset acquired from a different brand ultrasound machine at a different hospital was used for testing. Unfortunately,the second dataset doesn't have manual segmentation masks. Therefore, the segmentation masks of the second dataset were generated by using Fetal-TransUnet [cite:cengiz] and an expert tagged visually the outputs that the segmentation of ultrasound has good or poor quality.

### 2.2. Implementation Details

The samples in the first dataset were divided into the training and validation sets with percentage of 90%, and 10&. We make sure that a set of generated masks and the original segmentation mask of a patient were used for only training or only validation. The second dataset was only used for testing because our aim was to see the performance of the model when we use a different dataset that has different data characteristics than training. Data augmentation with scaling and horizontal flipping was applied.

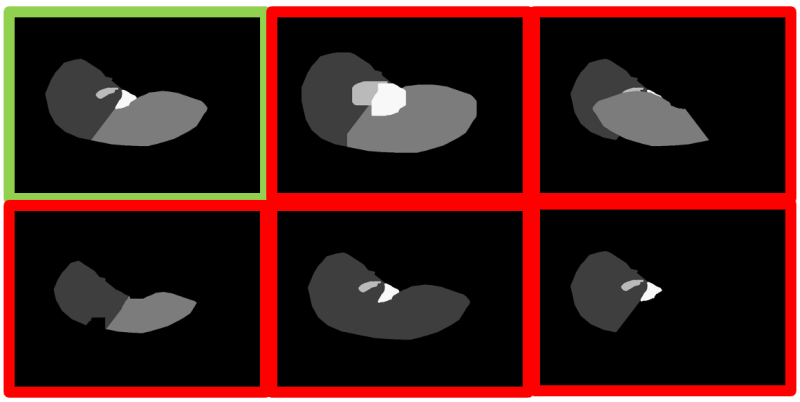

Figure 1: Poorly generated segmentation example

In this section, we introduce our proposed method that uses ultrasound images and generated segmentation masks as an input in the different R, G, and B channels. We loaded the pre-trained model, reset the final fully connected layer, and started training (Fig 2). Our proposed model was implemented in Pytorch and trained on a single NVIDIA RTX 1080. We have experimented with multiple hyper-parameters. We utilize the CrossEntropy Loss. Other hyper-parameters will be detailed in the acknowledgment.

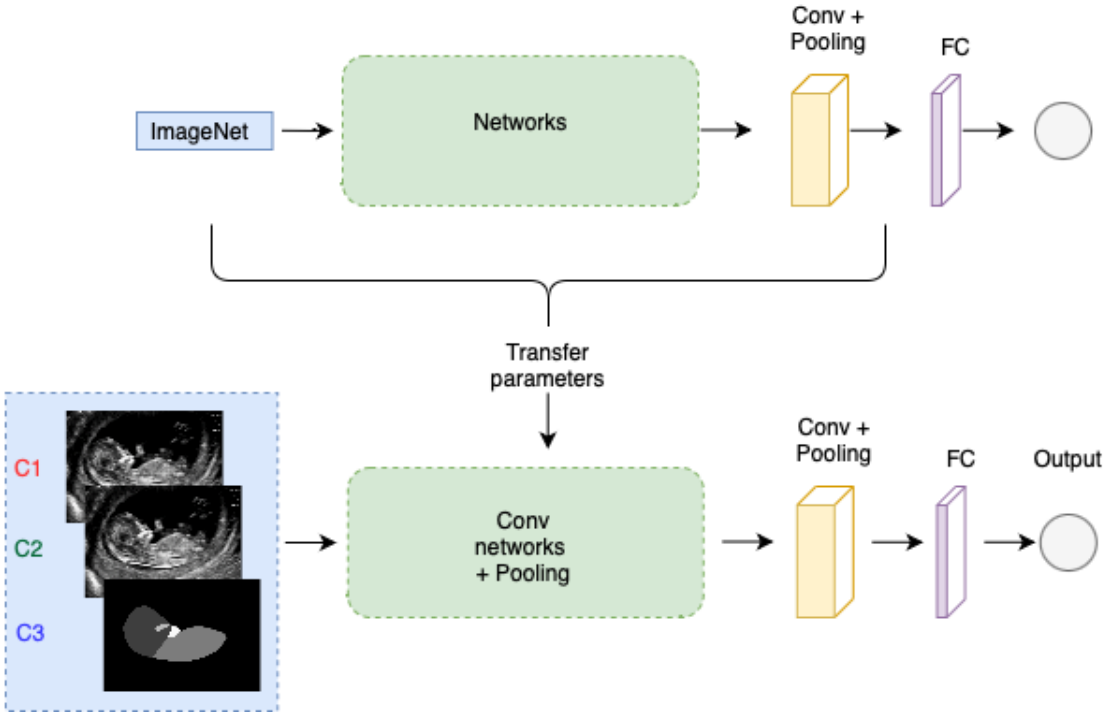


Figure 2: Proposed network

## 3. Experiments

**Experiment setup 1:** To compare our proposed solution with other CNN-based networks (Fig 2), Siamese network (Fig 3) and Synergic Deep Learning model (SDL) (Fig 3) were adopted to understand whether the segmentation mask has good or bad quality performance.

Siamese network was composed of two identical networks to learn features from the inputs, one of them for fetal ultrasound images and the other for segmentation masks. The output of these networks was concatenated and we passed it through a fully connected layer. The output from the layer was sent to binary cross-entropy loss.

For SDL, each pair of identical networks learned image representation from 2 image datasets and concatenated them to forward a fully connected layer to predict whether the segmentation of ultrasound images is good. In comparison to the Siamese network, 2 separate identical networks were used but they didn't share the weights.

**Experiment setup 2:** By using another dataset that is different than training and validation dataset, CRL measurement and GA calculated based on the segmentation mask if the segmentation quality is good, and then the measurements were compared to the real reported CRL and GA measurements.

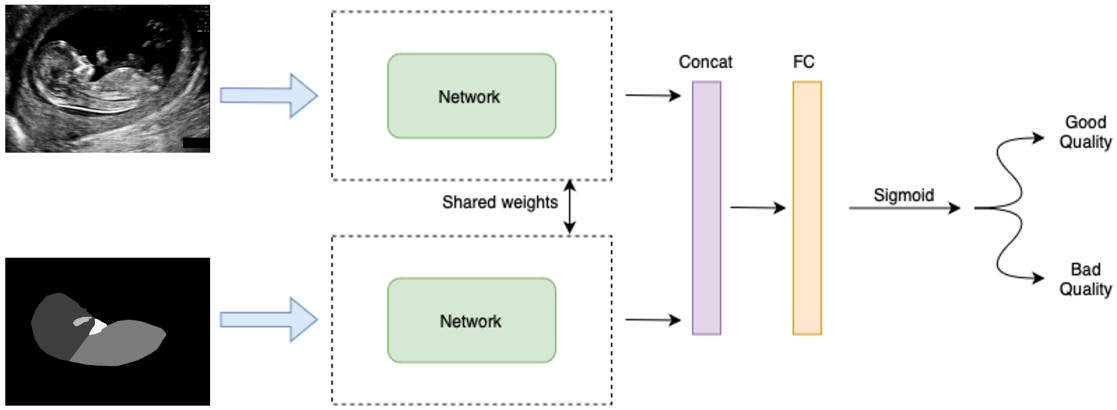


Figure 3: Siamese network

## 4. Results

The proposed model was trained using ResNet, DenseNet and AlexNet. Higher accuracy was observed with ResNet50 pretrained (Table 1). Therefore, the proposed model and other models including Siamese and Synergic were compared only with ResNet models.

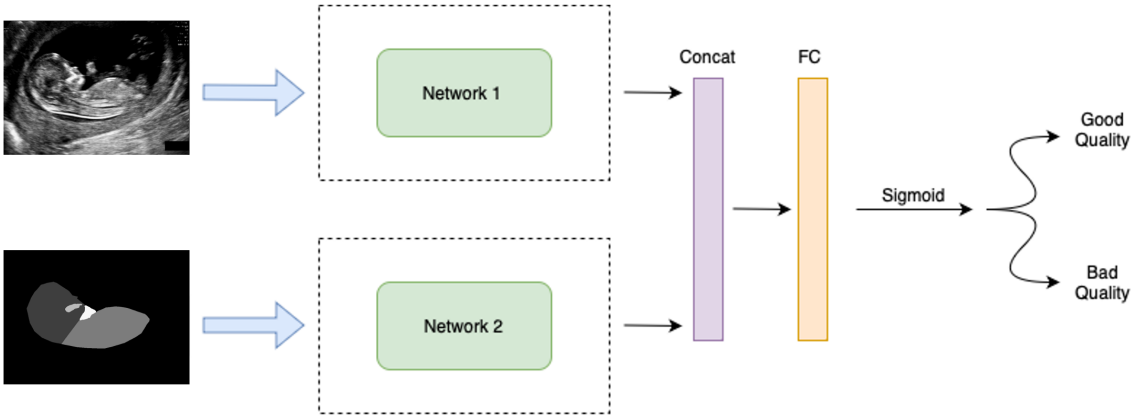


Figure 4: Synergic network

| Model | Pretrained | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| Siamese | ResNet18 False | 0.679 | 0.618 | 0.933 | 0.744 |
| | ResNet18 True | 0.75 | 0.673 | 0.97 | 0.795 |
| | ResNet34 False | 0.708 | 0.632 | 0.991 | 0.772 |
| | ResNet34 True | 0.6 | 0.555 | 0.995 | 0.713 |
| | ResNet50 False | 0.791 | 0.773 | 0.825 | 0.798 |
| | ResNet50 True | 0.5 | 0.5 | 1 | 0.66 |
| Synergic | ResNet18 False | 0.7104 | 0.6498 | 0.9125 | 0.759 |
| | ResNet18 True | 0.547 | 0.525 | 0.983 | 0.6851 |
| | ResNet34 False | 0.76 | 0.78 | 0.725 | 0.751 |
| | ResNet34 True | 0.7958 | 0.738 | 0.916 | 0.817 |
| | ResNet50 False | 0.756 | 0.7 | 0.895 | 0.786 |
| | ResNet50 True | 0.787 | 0.746 | 0.87 | 0.803 |
| Proposed | ResNet18 False | 0.7124 | 0.672 | 1 | 0.804 |
| | ResNet18 True | 0.871 | 0.795 | 1 | 0.886 |
| | ResNet34 False | 0.787 | 0.664 | 1 | 0.798 |
| | ResNet34 True | 0.871 | 0.768 | 1 | 0.869 |
| | ResNet50 False | 0.699 | 0.625 | 0.991 | 0.767 |
| | **ResNet50 True** | **0.902** | 0.804 | 0.982 | 0.870 |
| | Dense121 False | 0.8 | 0.672 | 1 | 0.804 |
| | Dense121 True | 0.898 | 0.773 | 1 | 0.872 |
| | AlexNet False | 0.752 | 0.662 | 0.938 | 0.776 |
| | AlexNet True | 0.862 | 0.733 | 1 | 0.846 |

| | CRL diff. in mm | GA diff. days |
|---|---|---|
| Segmentation based | 2.64 ± 3.20 | 1.45 ± 1.74 |

| | | | | Accuracy |
|---|---|---|---|---|
| **Segmentation Quality** | **Pixel-Gap-Tag** | **Fetalpalateexist** | **Magnification** | **FetalFacedirection** |
| Good | 0.203703704 | 0.064814815 | 0.009259259 | 0.009259259 |
| Poor | 0.324074074 | 0.351851852 | 0.25 | 0.240740741 |

## 5. Discussion

## 6. Conclusion

## Acknowledgments

Acknowledgments withheld.

## References

Sevim Cengiz and Mohammad Yaqub. Automatic fetal gestational age estimation from first trimester scans. volume 12967 LNCS, 2021. doi: 10.1007/978-3-030-87583-1_22.

Karen Drysdale, D Ridley, Karry Walker, B Higgins, and Taraneh Dean. First-trimester pregnancy scanning as a screening tool for high-risk and abnormal pregnancies in a district general hospital setting. *Journal of Obstetrics and Gynaecology*, 22:159–165, 2002. doi: 10.1080/01443610120113300. URL https://doi.org/10.1080/01443610120113300.

Zeju Li, Konstantinos Kamnitsas, Mobarakol Islam, Chen Chen, and Ben Glocker. Estimating model performance under domain shifts with class-specific confidence scores. 7 2022. URL http://arxiv.org/abs/2207.09957.

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, 2019. ISSN 13618423. doi: 10.1016/j.media.2019.101548.

Chaoyue Liu, Guohao Dong, Muqing Lin, Yaoxian Zou, Tianzhu Liang, Xujin He, Zhijie Chen, Dong Ni, Yi Xiong, and Lei Zhu. Ars-net: Adaptively rectified supervision network for automated 3d ultrasound image segmentation. volume 11766 LNCS, 2019. doi: 10.1007/978-3-030-32248-9_42.

Pádraig Looney, Gordon N. Stevenson, Kypros H. Nicolaides, Walter Plasencia, Malid Molloholli, Stavros Natsis, and Sally L. Collins. Fully automated, real-time 3d ultrasound segmentation to estimate first trimester placental volume using deep learning. *JCI insight*, 3, 2018. ISSN 23793708. doi: 10.1172/jci.insight.120178.

Ana I.L. Namburete, Weidi Xie, Mohammad Yaqub, Andrew Zisserman, and J. Alison Noble. Fully-automated alignment of 3d fetal brain ultrasound to a canonical reference space using multi-task learning. *Medical Image Analysis*, 46, 2018. ISSN 13618423. doi: 10.1016/j.media.2018.02.006.

Siti Nurmaini, Muhammad Naufal Rachmatullah, Ade Iriani Sapitri, Annisa Darmawahyuni, Bambang Tutuko, Firdaus Firdaus, Radiyati Umi Partan, and Nuswil Bernolian. Deep learning-based computer-aided fetal echocardiography: Application to heart standard view segmentation for congenital heart defects detection. *Sensors*, 21, 2021. ISSN 14248220. doi: 10.3390/s21238007.

Manna E. Philip, Arcot Sowmya, Hagai Avnet, Ana Ferreira, Gordon Stevenson, and Alec Welsh. Convolutional neural networks for automated fetal cardiac assessment using 4d b-mode ultrasound. volume 2019-April, 2019. doi: 10.1109/ISBI.2019.8759377.

Hariharan Ravishankar, Sahana M. Prabhu, Vivek Vaidya, and Nitin Singhal. Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning. volume 2016-June, 2016. doi: 10.1109/ISBI.2016.7493382.

Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V. Valindria, Mihir M. Sanghvi, Nay Aung, José M. Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Bernhard Kainz, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Daniel Rueckert, and Ben Glocker. Real-time prediction of segmentation quality. volume 11073 LNCS, 2018. doi: 10.1007/978-3-030-00937-3_66.

Hosuk Ryou, Mohammad Yaqub, Angelo Cavallaro, Aris T. Papageorghiou, and J. Alison Noble. Automated 3d ultrasound image analysis for first trimester assessment of fetal health. *Physics in Medicine and Biology*, 64, 2019. ISSN 13616560. doi: 10.1088/1361-6560/ab3ad1.

Nadav Schwartz, Ipek Oguz, Jiancong Wang, Alison Pouch, Natalie Yushkevich, Shobhana Parameshwaran, James Gee, Paul Yushkevich, and Baris Oguz. Fully automated placental volume quantification from 3d ultrasound for prediction of small-for-gestational-age infants. *Journal of Ultrasound in Medicine*, 41, 2022. ISSN 15509613. doi: 10.1002/jum.15835.

Zahra Sobhaninia, Shima Rafiei, Ali Emami, Nader Karimi, Kayvan Najarian, Shadrokh Samavi, and S. M. Reza Soroushmehr. Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning. 2019. doi: 10.1109/EMBC.2019.8856981.

Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I.L. Namburete. Multi-task cnn for structural semantic segmentation in 3d fetal brain ultrasound. volume 1065 CCIS, 2020. doi: 10.1007/978-3-030-39343-4_14.

Madeleine K. Wyburd, Mark Jenkinson, and Ana I.L. Namburete. Cortical plate segmentation using cnns in 3d fetal ultrasound. volume 1248 CCIS, 2020. doi: 10.1007/978-3-030-52791-4_5.

Xin Yang, Lequan Yu, Shengli Li, Huaxuan Wen, Dandan Luo, Cheng Bian, Jing Qin, Dong Ni, and Pheng Ann Heng. Towards automated semantic segmentation in prenatal volumetric ultrasound. *IEEE Transactions on Medical Imaging*, 38, 2019. ISSN 1558254X. doi: 10.1109/TMI.2018.2858779.

Xin Yang, Xu Wang, Yi Wang, Haoran Dou, Shengli Li, Huaxuan Wen, Yi Lin, Pheng Ann Heng, and Dong Ni. Hybrid attention for automatic segmentation of whole fetal head in prenatal ultrasound volumes. *Computer Methods and Programs in Biomedicine*, 194, 2020. ISSN 18727565. doi: 10.1016/j.cmpb.2020.105519.

Mohammad Yaqub, Brenda Kelly, Aris T. Papageorghiou, and J. Alison Noble. A deep learning solution for automatic fetal neurosonographic diagnostic plane verification using clinical standard constraints. *Ultrasound in Medicine and Biology*, 43, 2017. ISSN 1879291X. doi: 10.1016/j.ultrasmedbio.2017.07.013.

Fengcheng Zhu, Mengyuan Liu, Feifei Wang, Di Qiu, Ruiman Li, and Chenyang Dai. Automatic measurement of fetal femur length in ultrasound images: A comparison of random forest regression model and segnet. *Mathematical Biosciences and Engineering*, 18, 2021. ISSN 15510018. doi: 10.3934/mbe.2021387.

Veronika A. Zimmer, Alberto Gomez, Emily Skelton, Nicolas Toussaint, Tong Zhang, Bishesh Khanal, Robert Wright, Yohan Noh, Alison Ho, Jacqueline Matthew, Joseph V. Hajnal, and Julia A. Schnabel. Towards whole placenta segmentation at late gestation using multi-view ultrasound images. volume 11768 LNCS, 2019. doi: 10.1007/978-3-030-32254-0_70.

## Appendix A. Proof of Theorem 1

This is a boring technical proof of

## Appendix B. Proof of Theorem 2

This is a complete version of a proof sketched in the main text.