END-TO-END FIRST TRIMESTER FETAL ULTRASOUND VIDEO AUTOMATED CRL AND NT SEGMENTATION

Robail Yasrab[1]   *Zeyu Fu*[1]   *Lior Drukker*[2,3]   *Lok Hin Lee*[1]   *He Zhao*[1]

Aris T. Papageorghiou[2]   *J. Alison Noble*[1]

[1]Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2]Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK
[3]Rabin Medical Center, Sackler Faculty of Medicine, Tel-Aviv University, Israel

ABSTRACT

This study presents a novel approach to automatic detection and segmentation of the Crown Rump Length (CRL) and Nuchal Translucency (NT), two essential measurements in the first trimester US scan. The proposed method automatically localises a standard plane within a video clip as defined by the UK Fetal Abnormality Screening Programme. A Nested Hourglass (NHG) based network performs semantic pixel-wise segmentation to extract NT and CRL structures. Our results show that the NHG network is faster ($19.52\% <$ GFlops than FCN32) and offers high pixel agreement (mean-IoU=80.74) with expert manual annotations.

Index Terms— First trimester, video segmentation, crown rump length (CRL), ultrasound, nuchal translucency (NT).

1. INTRODUCTION

Fetal ultrasound (US) is a non-invasive imaging method for assessing fetal growth and development. The prenatal first trimester US scan is carried out at 11^{+0} to 13^{+6} weeks gestation to evaluate fetal viability, pregnancy dating and assess the risk for chromosomal anomalies [1]. To accomplish these tasks, current clinical approaches rely on manual selection of the mid-sagittal plane with a measurement of the fetal Nuchal Translucency (NT) and Crown-Rump Length (CRL), which is subjective and requires extensive training and years of experience [2, 3].

Contribution. We present a two-stage deep learning architecture that automatically detects the mid-sagittal plane (MSP) and segments the key CRL and NT structures as shown in Fig. 1. As a pre-processing step, a real-time detection CNN predicts the class probabilities of key anatomical structures (nose, head, horizontal sagittal section, diencephalon and rump) to detect the best MSP view. Stage two is a novel nested encoder-decoder semantic segmentation architecture designed to segment the CRL and NT structures. The proposed design aims to ensure that the various levels of US image features extracted from the encoder are delivered to the decoder to discriminate more subtle anatomical structures at the cost of fewer trainable parameters (32.5% fewer than U-Net [4]). A class balancing based weighted loss function was employed to further improve the segmentation, which is reflected by an increase of 4.27% in the mean intersection-over-union (IoU) score.

Related Work. There are a limited number of studies on automated fetal biometry for the first-trimester US. Zhao et al. [5] presented a linear support vector machine (SVM)-based study to detect physical characteristics that ultimately help to detect Down Syndrome. Nirmala et al. [6] proposed NT detection method based on segmentation and edge detection. The authors utilized a shift procedure that clusters the features of pixels in an iterative way to get the segmentation mask of NT. More recently, Sobhaninia et al. [7] proposed a neural network-based multi-task fetal head circumference segmentation method for fetal biometry. However, none of the aforementioned methods considered the combined task of detecting a standard plane and segmentation of key anatomical structures from US video.

2. METHODS

2.1. Data Acquisition

The dataset of 250 full-length routine first-trimester free-hand fetal US scans containing midline sagittal view of fetus acquired under a large-scale study PULSE (Perception Ultrasound by Learning Sonographer Experience) at Fetal Medicine Unit, Oxford University Hospitals National Health Services (NHS) Foundation Trust. The scans were performed on a commercial Voluson E8 version BT18 *(General Electric Healthcare, Zipf, Austria)* US machine. The setup was equipped with customized video recording software through secondary video output of the ultrasound machine to record full-length video scans using screengrab [8]. The video data was saved by anonymizing the patient details. The full-length US scans were recorded with HD resolution

Table 1: Details of datasets and tasks used in this study.

Anatomy	Task	Datasets	Video Segments	Frames
CRL	SPD and SPS	Training	100	12534 (77.9%)
		Validation	18	2385 (14.8%)
		Test	10	1174 (7.2%)
NT	SPS	Training	110	10174 (79.3%)
		Validation	27	2083 (16.2%)
		Test	9	564 (4.4%)

(1920 × 1080 pixels) at 30 frames per second and lossless compression. The average duration of acquired first-trimester US scans is 13.73 ± 4.18 minutes (24720 ± 7534 frames). Figure 2 shows an illustrative example of how an US scan was partitioned into video clips by an expert.

2.2. The Proposed Architecture

Figure 1 presents an overview of the proposed architecture. US frames are input to a pre-processing CNN to detect the best MSP. Next, the selected keyframes are fed into the proposed NHG with a weighted loss function for the segmentation of CRL and NT. The final predictions are refined using a dense Conditional Random Field (dCRF) model.

2.2.1. Sagittal Plane Detection (SPD)

For the sagittal plane detection (SPD) task, Table 1 summarises the CRL dataset manually annotated by an engineering researcher and a clinical fellow for five anatomical structures; a) head [Hd], b) horizontal sagittal section of the fetus [HS], c) echogenic tip of the nose [EN], d) rump [Ru], and e) translucent diencephalon [TD]. We applied Yolo-v5 [9] for high-speed (more than 30 frames per second (fps)) US anatomical object detection posed as a regression and classification problem; it returns class label and associated probabilities. The best MSP is detected when all anatomical classes are detected with a probability higher than > 70%. This threshold was selected after several experiments to ensure the presence of all key anatomical classes must be present as suggested by NHS Fetal Anomaly Screening Programme (FASP) guidelines [1].

2.2.2. Sagittal Plane Segmentation (SPS)

For sagittal plane segmentation (SPS), we designed a NHG network architecture that sandwiches a single Hourglass (HG) [10] between residual blocks [11], as shown in Fig. 3. The proposed architecture arranges the residual, pooling, and HG blocks appropriately during the encoder stage, and likewise, during the decoding stage to produce various levels of feature maps in the same block. This leads to a final segmentation mask extracted with the help of encoder pooling in-

dices. During NHG network training, extreme foreground-background class imbalance, especially classes such as NT, was found to be problematic. To address this we introduced a weighted-loss (WL) function that assigns weights to each class inversely proportional to the median frequency in which that class appears throughout the entire training set [12]. This offers a more customized loss calculation strategy than the general focal-loss approach [13]. This simple heuristic loss calculation improves segmentation performance by optimizing the network convergence (by adding focus to foreground pixels) without additional trainable parameters.

The proposed weighted loss W_L is defined as:

$$W_L = \frac{\alpha_c}{N} \sum_{n=1}^{N} \sum_{x=1}^{W} \sum_{y=1}^{H} \left[g_{xy}^n \log\left(\hat{g}_{xy}^n\right) + \left(1 - g_{xy}^n\right) \log\left(1 - \hat{g}_{xy}^n\right) \right],$$

where, N is the number of feature maps, $\hat{g}_{xy}^n$ is the predicted class, and g_{xy}^n is the ground truth. The weight of each class α_c is scaled by its frequency relative to the median frequency of all classes, calculated as:

$$\alpha_c = \frac{median\_freq}{freq(c)},$$

where, $freq(c)$ is the frequency of class c pixels occurrences divided by the number of pixels in any image containing that class, and $median\_freq$ is the median of these frequencies over all classes [12]. A dCRF is used as a post-processing step for smoothing and maximizing agreement between similar neighbouring pixels of the predicted segmentation masks at the inference stage.

3. EXPERIMENTS

3.1. Settings and Metrics

The pre-processing CNN (Yolo-v5 [9]) and NHG architectures were trained for 200 epochs to detect the sagittal plane and segment the CRL and NT. Training was initiated with a 0.1 learning rate (lr) and decreased by a factor of $\times 0.1$ every 30 epochs. The data augmentation policy included rotation $[-30°, 30°]$ and horizontal flipping. For evaluation of the SPD model, Recall (R), Precision (P), F1-score (F1), and Top-1 accuracy (Top-1) metrics are reported. For evaluation of SPS, Global Average Accuracy (GAA), Mean Accuracy (MA), and Mean Intersection Over Union (mIoU) metrics are reported.

3.2. Evaluation of Sagittal Plane Detection

For the SPD task, we evaluated Yolo-v5 on the test set. The trained Yolo-v5 model statistics are P=0.88 ± 0.05, R=0.85 ± 0.03, F1=0.85 ± 0.10 and Top-1=0.87 ± 0.06. To further understand detection performance, we report the confusion matrix in Fig. 5. 'Hd' and 'HS' show little class confusion. 'EN', 'Ru' and 'TD' classes show some inter-class confusion.
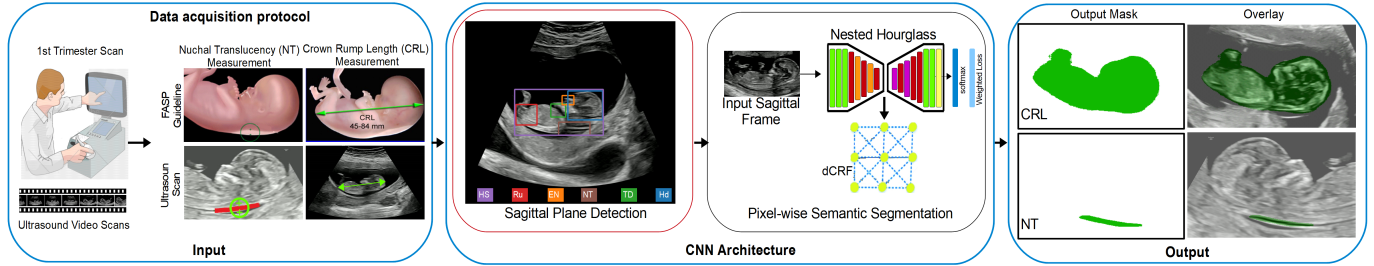
Fig. 1: An overview of the proposed architecture for automated fetal biometry in first-trimester US scans.

Table 2: Quantitative analysis of trained models on test dataset.

| Methods	Para.(M)	CRL			NT		
		GAA(%)	MA(%)	mIoU(%)	GAA(%)	MA(%)	mIoU(%)
FCN-16 [14]	134.27	79.05 ± 0.05	66.23 ± 0.10	54.48 ± 0.20	82.64 ± 0.21	55.11 ± 0.03	51.60 ± 0.15
FCN-32 [14]	144	81.68 ± 0.01	76.56 ± 0.02	63.87 ± 0.18	85.02 ± 0.14	56.97 ± 0.01	51.80 ± 0.11
U-Net [4]	30.72	83.64 ± 0.08	79.80 ± 0.07	67.41 ± 0.25	90.17 ± 0.02	60.41 ± 0.01	58.39 ± 0.01
SegNet [15]	15.27	85.08 ± 0.10	83.82 ± 0.10	70.05 ± 0.33	89.66 ± 0.31	56.61 ± 0.24	48.18 ± 0.24
HG (B=1, S=2) [10]	35.08	89.05 ± 0.09	82.70 ± 0.20	70.83 ± 0.05	94.22 ± 0.01	64.10 ± 0.01	63.10 ± 0.05
NHG (ours)	11.46	92.32 ± 0.03	85.01 ± 0.01	74.42 ± 0.04	92.49 ± 0.05	66.37 ± 0.11	67.37 ± 0.01
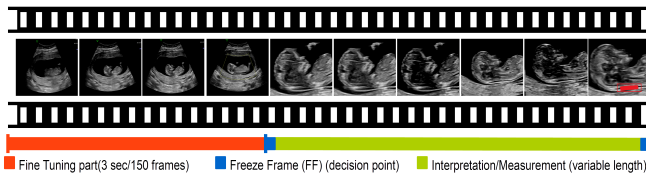
Fig. 2: Illustration of expert annotation process: video frames are annotated as Frozen (FF) video segments (blue), measurements (technical annotation) segment (green) and fine-tune segment (red). A three-second pre-frozen (fine-tune) state was added to incorporate a wide variety of anatomical views.
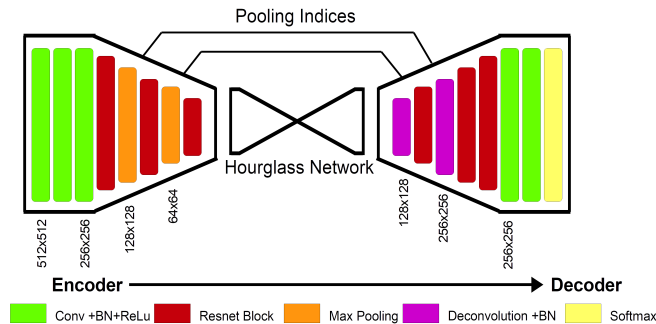

Fig. 3: Nested Hourglass (NHG) deep learning architectures.

3.3. Evaluation of Sagittal Plane Segmentation

For the SPS task, we trained and tested benchmark CNNs (FCN [14], UNet [4], SegNet [15] and Hourglass [10]) which were selected due to their high benchmark segmentation performance on the public computer vision datasets. Experimental results are reported in Table 2. The results showed that the proposed low compute design of the NHG network outperforms other benchmark CNN architectures. NHG offered 3.07% higher GAA scores than the standard HG (block=1,

Table 3: Quantitative results of NHG for NT and CRL segmentation on test dataset.

Architecture	CRL-Mean IoU	NT-Mean IoU	mean
NHG-Focal-Loss	76.14 ± 0.01	69.51 ± 0.05	72.82 ± 0.05
NHG-Focal-Loss-dCRF	76.92 ± 0.21	71.01 ± 0.22	73.96 ± 0.13
NHG-Weighted-Loss	78.69 ± 0.27	72.89 ± 0.20	75.79 ± 0.01
NHG-Weighted-Loss+dCRF	80.02 ± 0.19	73.71 ± 0.02	76.86 ± 0.02

stack=2). The effectiveness of NHG-based segmentation can be attributed to its layers arrangement, which offers repeated bottom-up, top-down processing with intermediate supervision.

Addition of weighted loss reflects 0.87% increase in the mIoU score, as shown in Table 3. These empirical results showed that the proposed NHG network with weighted loss performs consistently better than a class balancing ('focal loss') strategy based on standard cross-entropy. The quantitative metrics indicate that the majority of pixels have been classified correctly, depicted in Fig. 4. Figure 4-f shows that class balancing and dCRF yield considerable improvements by maximising agreement and smoothing between similar neighbouring pixels. These methods helped improve the segmentation of conflicting regions of pixels where the image was cluttered. However, in segmentation classes such as 'NT', the GAA score is higher, whereas the mIoU score is lower in comparison to the 'CRL' class, which certainly happens due to an imbalance between foreground and background classes. The dCRF also offers a well-defined separation between foreground and background pixels, specifically in the NT class, which is reflected in increased mIoU$= 1.07\%$ scores for each class. The test set automated semantic pixel-wise segmentation showed a high correlation Pearson Correlation Coefficient (PCC) value ($\rho = 0.93, p = 0.0003$) with
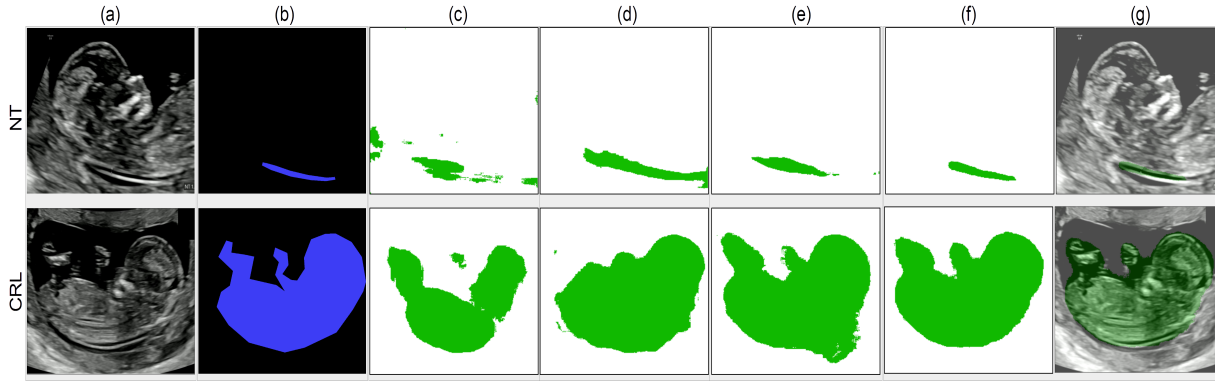
Fig. 4: Example results: a) input video frame, b) ground truth mask, c) NHG output (no weighted loss), d) NHG output with additional weighted loss (WL) function, e) NHG with additional post-processing with dCRF, f) NHG model with WL and dCRF, g) (f) overlaid on the input image.
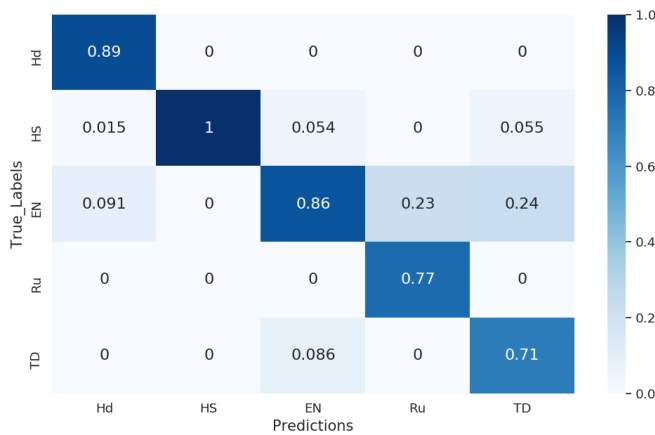

Fig. 5: Confusion Matrix. Automatic v/s Manual Labeling.

manually segmented video.

4. CONCLUSION

We have presented a deep-learning based architecture that takes an ultrasound video as an input and outputs key structure segmentations that are used for fetal biometry in first-trimester US in one step. At the segmentation stage, our NHS based network outperformed all benchmark architectures in terms of accuracy, speed, and parameter efficiency. A good correlation was found between manually labelled and automatically segmented anatomical structures. The future work will examine downstream automated biometry and translational issues in terms of algorithm evaluation and its application in the clinical setting.

5. COMPLIANCE WITH ETHICAL STANDARDS

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

7. REFERENCES

[1] D Kirwan, "NHS Fetal Anomaly Screening Programme," *National Standards and Guidance for England*, vol. 18, no. 0, 2010.

[2] P Taipale et al., "Learning curve in ultrasonographic screening for selected fetal structural anomalies in early pregnancy," *Obstetrics & Gynecology*, vol. 101, no. 2, pp. 273–278, 2003.

[3] L Drukker et al., "Vp18. 07: First trimester scans: how much time does it take to acquire the crl and nt?," *Ultrasound in Obstetrics & Gynecology*, vol. 58, pp. 174–174, 2021.

[4] O Ronneberger et al., "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Springer, 2015, pp. 234–241.

[5] Q Zhao et al., "Automated down syndrome detection using facial photographs," in *Proc. IEEE EMBC*. IEEE, 2013, pp. 3670–3673.

[6] S Nirmala et al., "Measurement of nuchal translucency thickness in first trimester ultrasound fetal images for detection of chromosomal abnormalities," in *Proc. IN-CACEC*. IEEE, 2009, pp. 1–5.

[7] Z Sobhaninia et al., "Fetal ultrasound image segmentation for measuring biometric parameters using multitask deep learning," in *Proc. IEEE EMBC*. IEEE, 2019, pp. 6545–6548.

[8] L Drukker et al., "Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[9] G Jocher et al., "Yolov5," *Code repository https://github. com/ultralytics/yolov5*, 2020.

[10] A Newell et al., "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*. Springer, 2016, pp. 483–499.

[11] K He et al., "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[12] R Yasrab et al., "Rootnav 2.0: Deep learning for automatic navigation of complex plant root architectures," *GigaScience*, vol. 8, no. 11, pp. giz123, 2019.

[13] T Lin et al., "Focal loss for dense object detection," in *Proc. IEEE ICCV*, 2017, pp. 2980–2988.

[14] J Long et al., "Fully convolutional networks for semantic segmentation," in *Proc. IEEE CVPR*, 2015, pp. 3431–3440.

[15] V Badrinarayanan et al., "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.