Pseudo-label Enhanced TransUNet for Robust Landmark Localization in Intrapartum Ultrasound

Xuezhi Zhang[1,2,3,4], Xi Chen[1,2,3,4], Hao Yan[5], Lyuyang Tong[1,2,3,4*✉], and Bo Du[1,2,3,4*✉]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] National Engineering Research Center for Multimedia Software, Wuhan University
[3] Institute of Artificial Intelligence, School of Computer Science, Wuhan University
[4] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University
[5] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
[*] Corresponding authors: Lyuyangtong@whu.edu.cn, dubo@whu.edu.cn

Abstract. Accurate and reliable detection of anatomical landmarks in intrapartum ultrasound is a critical component of quantitative and objective assessment of fetal head descent, which plays an essential role in guiding clinical decision-making during labor. However, manual annotation of ultrasound images is time-consuming, requires expert knowledge, and suffers from inter-observer variability. Moreover, the scarcity of fully annotated datasets poses additional challenges for training high-performance deep learning models in this domain. To address these challenges, we propose a three-stage framework that effectively leverages both fully labeled and partially labeled data to improve landmark detection performance. In Stage 1, a TransUNet model is pre-trained on a large-scale video-derived segmentation dataset and iteratively fine-tuned on point-annotated images using an error-weighted loss strategy. Stage 2 incorporates high-confidence pseudo-labeled data generated by the refined model, with post-processing applied to ensure label quality. Stage 3 fuses predictions from three independently trained TransUNet models via averaging to enhance stability and robustness. Experimental results on the IUGC 2025 Landmark Detection Challenge test set demonstrate that our method achieves an Average Point Distance of 13.28 pixels and an AOP MAE of 3.87 degrees, demonstrating the effectiveness of semi-supervised learning and model ensembling for intrapartum ultrasound landmark detection.

Keywords: Intrapartum Ultrasound · Landmark Detection · Semi-supervised Learning · Pseudo-labeling

1 Introduction

Intrapartum ultrasound is an important imaging tool for real-time assessment of labor progression, offering more objective and reproducible information than

traditional clinical examinations. Key anatomical landmarks—such as the fetal head and pubic symphysis—are essential for deriving clinically relevant measurements like the angle of progression (AoP) and head–perineum distance, which guide decisions on labor management.

Manual annotation of these landmarks requires expert knowledge, is time-consuming, and suffers from inter- and intra-observer variability, limiting large-scale clinical adoption. Fully automated landmark localization has thus become a pressing research goal. Traditional image processing methods (e.g., active shape models, Hough transforms) struggle with the noisy, low-contrast, and variable appearance of intrapartum ultrasound.

Deep learning methods, particularly convolutional neural networks (CNNs) and Transformer-based models, have shown strong performance in medical image analysis. U-Net [16] is a widely adopted encoder–decoder architecture with skip connections that effectively fuses multi-scale features for precise pixel-level predictions. TransUNet [8] extends U-Net by incorporating Vision Transformers into the encoder, combining CNN-based local feature extraction with global context modeling—an advantage in ultrasound landmark detection, where relevant structures may be distant or partially occluded.

Despite these advances, several challenges remain in applying deep learning models to intrapartum ultrasound. Ultrasound images are inherently noisy and exhibit poor contrast, making it difficult to distinguish anatomical boundaries. Furthermore, the scarcity of large annotated datasets limits the effectiveness of supervised learning frameworks. Therefore, there is a growing demand for data-efficient and architecture-robust solutions that can achieve high localization accuracy while accommodating the unique characteristics of intrapartum ultrasound.

In this study, we propose a fully automated three-stage framework for fetal landmark localization and AoP estimation from intrapartum ultrasound images. Our method leverages a progressive pseudo-labeling strategy to exploit unlabeled data and improve robustness. Specifically, we first pretrain a TransUNet-based segmentation model using manually labeled video keyframes and then refine it with pseudo labels from point-supervised images (Stage 1). Next, we perform progressive pseudo labeling with confidence-based filtering to incrementally incorporate high-quality unlabeled samples into training (Stage 2). Finally, we apply a weighted ensemble strategy, where each model independently predicts the segmentation mask and three anatomical landmarks, and their outputs are averaged to compute the AoP (Stage 3). Our contributions are summarized as follows:

- We propose a fully automatic framework for fetal head progression assessment in intrapartum ultrasound, which jointly performs anatomical landmark localization and AoP estimation.
- We design a three-stage pseudo-labeling strategy to leverage both labeled and unlabeled data, enhancing the effectiveness of training.

– We integrate a simple yet robust geometry-based module for AoP measurement based on three key landmarks, ensuring interpretability and clinical relevance.

In this work, we aim to support the technical implementation of the WHO Labour Care Guide and promote safer, more standardized labor monitoring practices via automated ultrasound analysis.

2 Method

As illustrated in Figure 1, our approach consists of three main components: 1) pretraining and label refinement; 2) progressive pseudo labeling with confidence filtering; and 3) model ensembling and final AoP estimation.

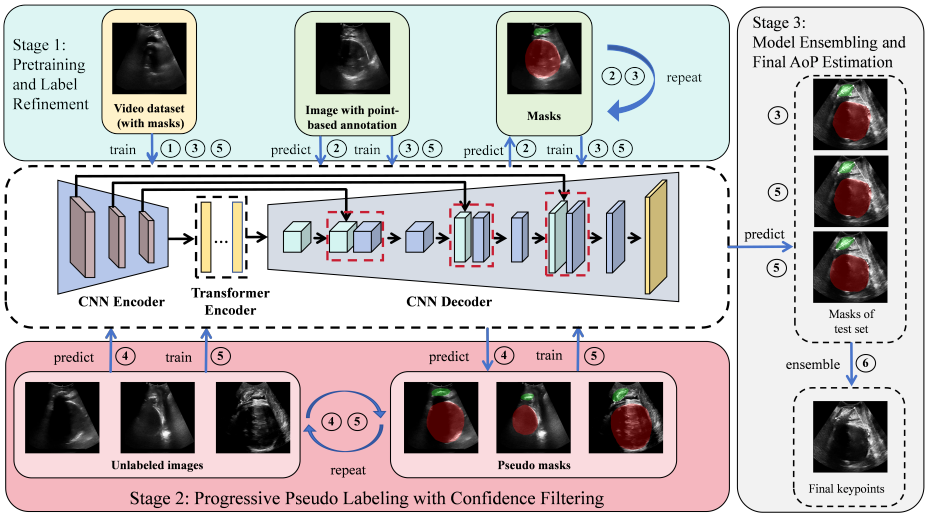

Fig. 1. Overview of the proposed framework.

2.1 Stage 1: Pretraining and Label Refinement

The first stage of our framework, illustrated in Figure 2, consists of a pretraining phase followed by iterative label refinement. We leverage the video dataset provided by the MICCAI IUGC 2024: Intrapartum Ultrasound Grand Challenge [2,12,13,4,10]. Specifically, we extract 2,562 key frames with corresponding segmentation masks from the videos as a pretraining dataset. A TransUNet model [8] is pre-trained on this dataset to learn robust anatomical representations. The pre-trained model is then used to generate masks for 300 images

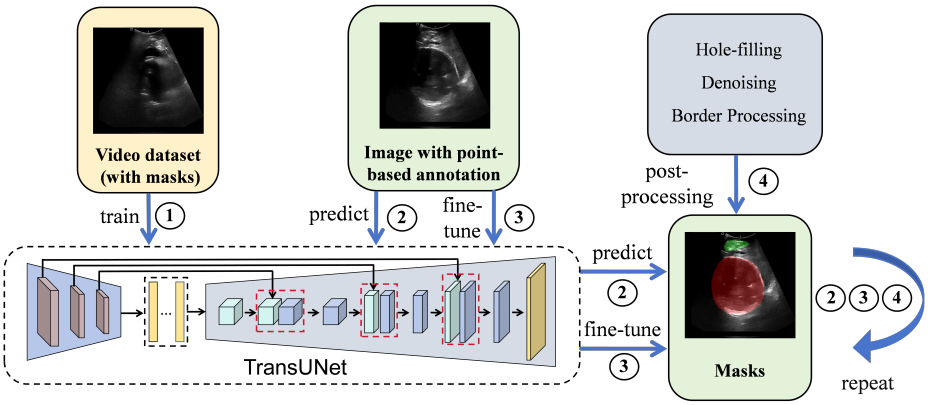

Fig. 2. Illustration of stage 1: pe-training and label refinement.

with point-based annotations (i.e., three annotated keypoints but no segmentation masks). To ensure the quality of these predictions, we apply a three-step post-processing pipeline:

- **Hole-filling:** Connected component analysis is used to fill holes within the segmented regions.
- **Denoising:** Only the largest connected component is retained, and small isolated regions are removed.
- **Border Processing:** Boundary-connected components are corrected to ensure the integrity of the segmentation mask.

Following post-processing, we compute the anatomical keypoints from each predicted mask (detailed in Stage 3), and measure the Euclidean distance between each predicted keypoint and its corresponding annotated location. These distances are then used to assign sample-specific loss weights during fine-tuning, where samples with larger prediction errors contribute less to the overall loss. Specifically, for the i-th sample, the loss weight w_i is defined as:

$$w_i = \exp\left(-\lambda \cdot \frac{1}{3} \sum_{j=1}^{3} \left| \mathbf{p}_{i,j} - \hat{\mathbf{p}}_{i,j} \right|_2 \right) \tag{1}$$

where $\mathbf{p}_{i,j}$ and $\hat{\mathbf{p}}_{i,j}$ denote the predicted and annotated coordinates of the j-th keypoint for the i-th sample, and λ is a scaling factor controlling the sensitivity to prediction error. This re-weighted fine-tuning process is iterated for three rounds, gradually refining the model using both the initial pseudo-labels and the spatial alignment between predicted and annotated landmarks. The final output of Stage 1 is a refined TransUNet model with improved segmentation quality tailored to point-supervised images.

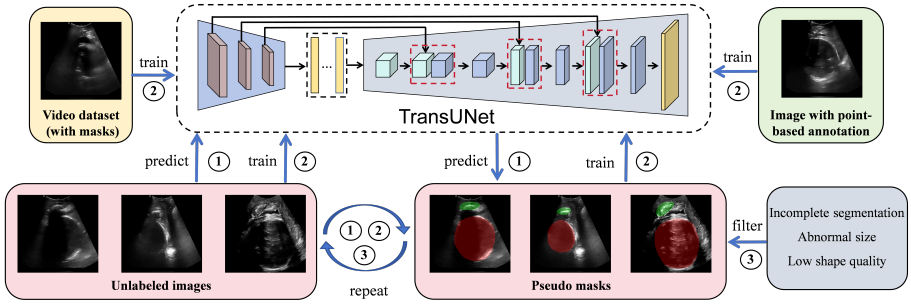

Fig. 3. Illustration of stage 2: progressive pseudo labeling with confidence filtering.

2.2 Stage 2: Progressive Pseudo Labeling with Confidence Filtering

In the second stage, we implement a progressive pseudo labeling strategy to exploit the unlabeled dataset more effectively. As illustrated in Figure 3, all unlabeled images are evenly split into three subsets. The fine-tuned model from Stage 1 is then used to generate segmentation masks for the first subset.

To ensure the quality of the pseudo labels, we apply a filtering process to remove unreliable masks based on the following criteria:

- **Incomplete segmentation:** Masks that contain only one or zero connected regions are discarded.
- **Abnormal size:** Masks whose area significantly deviates from the mean area of all labeled masks are discarded. Specifically, we discard masks whose area is either larger than 1.5× the mean or smaller than 0.5× the mean.
- **Low shape quality:** Masks with a fetal head mask shape factor less than 0.8 are removed. The shape factor is defined as:

$$\text{Shape Factor} = \frac{4\pi \times \text{Area}}{(\text{Perimeter})^2} \tag{2}$$

After filtering, the remaining high-quality pseudo-labeled samples from the first subset are combined with two fully labeled datasets: the manually labeled video keyframes and the corrected masks from Stage 1. To balance the contribution of different data sources, we apply different loss weights during training. Specifically, the total loss is computed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{stage1}} + 0.5 \times \mathcal{L}_{\text{pseudo}} \tag{3}$$

where $\mathcal{L}_{\text{video}}$ refers to the loss from the manually labeled video keyframes, $\mathcal{L}_{\text{stage1}}$ denotes the loss from the corrected pseudo labels generated in Stage 1, and $\mathcal{L}_{\text{pseudo}}$ represents the loss from the newly generated pseudo-labeled data in this stage.

The model trained on this combined dataset is then used to generate masks for the second subset of unlabeled data, followed by the same filtering and retraining process. This process is repeated once more to handle the third subset,

forming a three-step progressive refinement framework that incrementally improves pseudo-label quality and model performance.

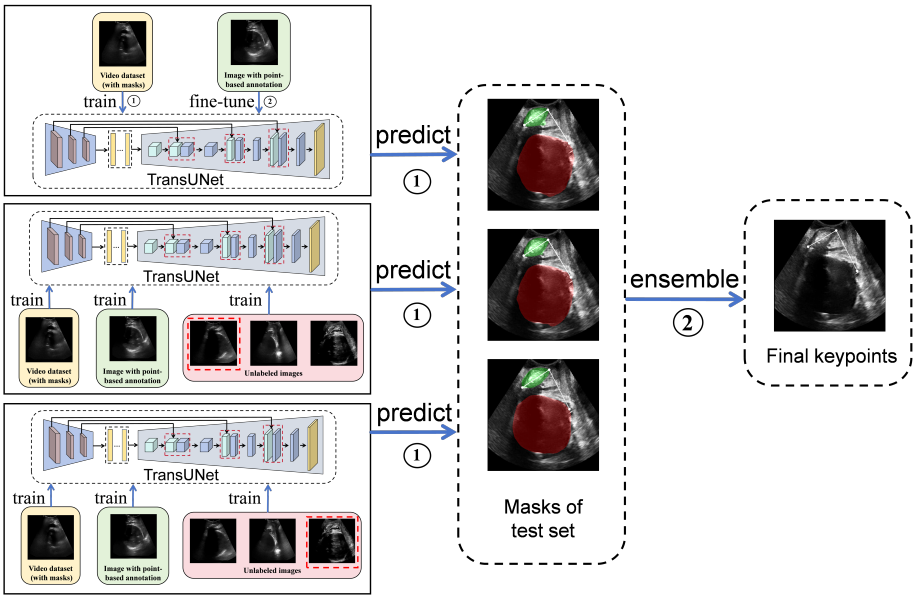

Fig. 4. Illustration of stage 3: model ensembling and final AoP estimation.

2.3 Stage 3: Model Ensembling and Final AoP Estimation

In the final stage, we adopt a model ensembling strategy to enhance segmentation robustness and measurement consistency. As shown in Figure 4, we employ an ensemble of three models: (1) the fine-tuned model from Stage 1, and (2)(3) two models trained respectively on the first and third subsets in Stage 2. Each model independently generates segmentation masks for the test set.

The predicted masks are first refined using the same post-processing pipeline introduced in Stage 1, which includes hole filling, denoising, and boundary correction. Based on the refined binary masks, we extract three anatomical keypoints required for computing the Angle of Progression (AoP): two points on the pubic symphysis contour (PS1 and PS2), and one tangential point on the fetal head contour (FH1). The extraction procedures are described below as pseudocode (Algorithm 1).

Each model predicts a set of coordinates for the three keypoints: PS1, PS2, and FH1. To reduce prediction variance, we adopt a coordinate-wise averaging

strategy across the three models:

$$\text{Keypoint}_{final} = \frac{1}{3}\sum_{m=1}^{3}\text{Keypoint}_m \tag{4}$$

where Keypoint_m represents the coordinates predicted by the m-th model. This straightforward fusion improves robustness and ensures stable keypoint localization results.

Algorithm 1 Extraction of PS1, PS2, and FH1

Require: Binary mask M_{ps} of pubic symphysis (label 1), binary mask M_{fh} of fetal head (label 2)
Ensure: Coordinates $(PS1_x, PS1_y)$, $(PS2_x, PS2_y)$, $(FH1_x, FH1_y)$

$C_{ps} \leftarrow \text{FindLargestContour}(M_{ps})$
$P_{ps} \leftarrow \text{ExtractContourPoints}(C_{ps})$
$(p_a, p_b) \leftarrow \text{FindFurthestPointPair}(P_{ps})$
if $p_a.x > p_b.x$ **then**
    $PS1 \leftarrow p_a$, $PS2 \leftarrow p_b$
else
    $PS1 \leftarrow p_b$, $PS2 \leftarrow p_a$
end if
$C_{fh} \leftarrow \text{ExtractContourPoints}(M_{fh})$
$FH1 \leftarrow \arg\min_{p \in C_{fh}} \text{Angle}(\overrightarrow{\text{PS1}p}, \text{NormalVector}(p))$
if $FH1$ is not valid **then**
    $FH1 \leftarrow \text{FindRightmostPoint}(C_{fh})$
end if
return $PS1, PS2, FH1$

3 Experiments

3.1 Dataset Description

Our experiments are conducted on the benchmark dataset provided by the **Landmark Detection Challenge for Intrapartum Ultrasound Measurement (IUGC 2025)** [6,10,12,13,4,17,11,14,15,7,9,2,3,1,5] , which focuses on automatic landmark detection in fetal ultrasound images to assist clinical assessment of labor progression. The dataset is divided into the following subsets:

- **Training Set:** 31,421 ultrasound images in total, among which 300 images are manually annotated with three anatomical landmarks for supervised learning. The remaining unlabeled images are used for semi-supervised learning.
- **Validation Set:** 100 annotated images used to validate model performance during training.
- **Test Set:** 501 hidden images used for final evaluation by the challenge organizers.

In addition, we utilize an external dataset from the **MICCAI IUGC 2024: Intrapartum Ultrasound Grand Challenge** to improve the robustness of our segmentation model through pretraining. This dataset consists of:

- **Standard Plane Videos:** 288 videos composed entirely of standard planes, from which 24,434 frames are extracted, including 2,906 frames with segmentation masks.
- **Non-standard Plane Videos:** 168 videos consisting of non-standard planes, contributing 31,450 additional frames without segmentation labels.

This external video dataset serves as the foundation for initial model pretraining and label refinement in Stage 1 of our framework.

3.2 Experimental Setup

All experiments are conducted on two NVIDIA GeForce RTX 4090 GPUs. The network is trained in three stages: (1) 200 epochs of pretraining, (2) 100 epochs of fine-tuning on labeled data, and (3) 300 epochs of pseudo-label-based training.

We adopt stochastic gradient descent (SGD) with an initial learning rate of 0.07, which is decayed during training. The batch size is set to 16 throughout all stages. The input images are resized to 512×512, and the following data augmentation strategies are employed:

- **Random Horizontal Flip:** Applied with 50% probability.
- **Random Rotation:** Random rotation within $\pm 10^\circ$ using bilinear interpolation.
- **Color Jittering:** Brightness and contrast adjusted within a variation range of 0.1.

The base segmentation loss is a weighted combination of cross-entropy loss and Dice loss, defined as:

$$\mathcal{L}_{seg} = 0.5 \cdot \mathcal{L}_{CE} + 0.5 \cdot \mathcal{L}_{Dice} \tag{5}$$

To account for different data sources, we apply stage-specific loss scaling factors λ to balance contributions from labeled and pseudo-labeled samples. The detailed formulation can be found in the Method section.

3.3 Experimental Results

Table 1 presents the quantitative evaluation results for all models across different evaluation metrics. We report the overall Mean Squared Error (MSE), Mean Absolute Error (MAE), average distance error of three keypoints (PS1, PS2, and Tangency), as well as the MSE and MAE of the AoP angle.

As shown in Table 1, the fine-tuned model significantly outperforms the pretrained model, with the average keypoint distance reduced from 17.37 px to

Table 1. Quantitative evaluation results on the validation set.

Models	MSE	MAE	Average Point Distance	PS1 Distance	PS2 Distance	Tangency Distance	AOP MSE	AOP MAE
Pre-trained Model	417.60	10.96	17.37	10.89	14.60	26.60	213.26	8.16
Fine-tuned Model	260.55	8.68	13.80	9.10	8.99	23.32	152.36	5.98
Model with Pseudo-labels	247.50	8.23	13.04	8.04	8.63	22.45	124.75	5.31
Ensemble Model	225.51	7.99	12.67	7.77	8.56	21.67	128.00	5.30

13.80 px. Incorporating pseudo-labeled data further boosts performance, especially in the AoP angle estimation, where the MSE decreases from 152.36 to 124.75. The ensemble model achieves the best overall performance, reducing the average keypoint error to 12.67 px and the AoP MAE to 5.30, demonstrating that model fusion effectively mitigates prediction variance and enhances robustness.

Figure 5 visualizes representative segmentation results and predicted keypoints for different models. As shown, the pseudo-label model produces more accurate and stable keypoint locations, closely aligning with the ground truth and yielding smoother AoP angle estimation.

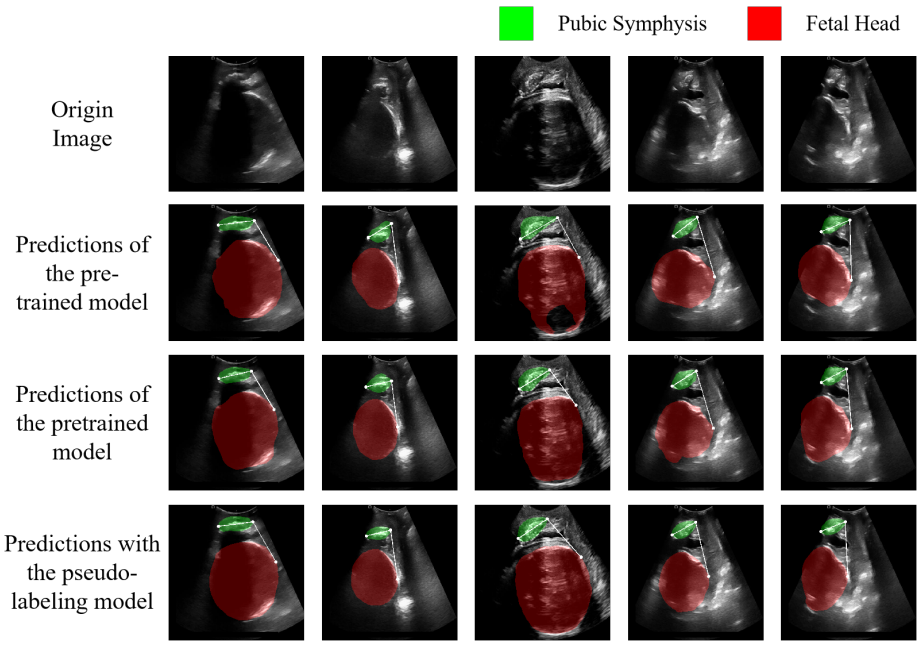

Fig. 5. Visualization results of segmentation and keypoints.

4 Conclusion

In this work, we proposed a three-stage framework for accurate landmark detection in intrapartum ultrasound images. Starting from a pre-trained TransUNet model, we refined labels through an iterative error-weighted fine-tuning strategy and further leveraged pseudo-labeled data to enhance generalization. Experimental results on the IUGC 2025 dataset demonstrate consistent performance gains at each stage. The proposed approach effectively bridges the gap between limited high-quality annotations and abundant unlabeled data, offering a practical and scalable solution for clinical labor progress assessment.

5 Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grants 2023YFC2705700, the National Natural Science Foundation of China under Grants 62306217 and 62225113, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20231987, the China Postdoctoral Science Foundation under Grant Number 2024T170686 and 2024M752471, the Major Program (JD) of Hubei Province (2023BAA017), the Innovative Research Group Project of Hubei Province under Grants 2024AFA017. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

1. Bai, J., Khobo, I., Slimani, S., Lu, Y., Ni, D., Yaqub, M., Lekadir, K., Ma, J., Li, S.: Landmark detection challenge for intrapartum ultrasound measurement meeting the actual clinical assessment of labor progress. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI 2025). Springer (2025). https://doi.org/10.5281/zenodo.15172238 7
2. Bai, J., Lekadir, K., Ni, D., Slimani, S., Campello, V.M., Ohene-Botwe, B., Lu, Y., Chen, G., Hou, H., Qiu, D., Zhou, Z.: Intrapartum ultrasound grand challenge 2024. In: Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2024). Springer (2024). https://doi.org/10.5281/zenodo.10979813 3, 7
3. Bai, J., Ou, Z., Lu, Y., Ni, D., Chen, G.: Pubic symphysis-fetal head segmentation from transperineal ultrasound images. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2023). Springer (2023). https://doi.org/10.5281/zenodo.7861699 7
4. Bai, J., Sun, Z., Yu, S., Lu, Y., Long, S., Wang, H., Qiu, R., Ou, Z., Zhou, M., Zhi, D., et al.: A framework for computing angle of progression from transperineal ultrasound images for evaluating fetal head descent using a novel double branch network. Frontiers in physiology **13**, 940150 (2022) 3, 7
5. Bai, J., Yang, Z., Hasan, K., Gan, J., Liang, Z., Cai, W., Tan, T., Ye, J., Yaqub, M., Ni, D., Slimani, S., Ohene-Botwe, B., Roman Victor Manuel, C., Lekadir, K.: Fetal ultrasound grand challenge: Semi-supervised cervical segmentation (fugc25). In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI 2025). IEEE (2024). https://doi.org/10.5281/zenodo.14328192 7

6. Bai, J., Zhou, Z., Ou, Z., Koehler, G., Stock, R., Maier-Hein, K., Elbatel, M., Martí, R., Li, X., Qiu, Y., et al.: Psfhs challenge report: pubic symphysis and fetal head segmentation from intrapartum ultrasound images. Medical Image Analysis **99**, 103353 (2025) 7
7. Chen, G., Bai, J., Ou, Z., Lu, Y., Wang, H.: Psfhs: intrapartum ultrasound image dataset for ai-based segmentation of pubic symphysis and fetal head. Scientific Data **11**(1), 436 (2024) 7
8. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) 2, 3
9. Chen, Z., Lu, Y., Long, S., Campello, V.M., Bai, J., Lekadir, K.: Fetal head and pubic symphysis segmentation in intrapartum ultrasound image using a dual-path boundary-guided residual network. IEEE journal of biomedical and health informatics **28**(8), 4648–4659 (2024) 7
10. Chen, Z., Ou, Z., Lu, Y., Bai, J.: Direction-guided and multi-scale feature screening for fetal head–pubic symphysis segmentation and angle of progression calculation. Expert Systems with Applications **245**, 123096 (2024) 3, 7
11. Jiang, J., Wang, H., Bai, J., Long, S., Chen, S., Campello, V.M., Lekadir, K.: Intrapartum ultrasound image segmentation of pubic symphysis and fetal head using dual student-teacher framework with cnn-vit collaborative learning. In: International conference on medical image computing and computer-assisted intervention. pp. 448–458. Springer (2024) 7
12. Lu, Y., Zhi, D., Zhou, M., Lai, F., Chen, G., Ou, Z., Zeng, R., Long, S., Qiu, R., Zhou, M., et al.: Multitask deep neural network for the fully automatic measurement of the angle of progression. Computational and mathematical methods in medicine **2022**(1), 5192338 (2022) 3, 7
13. Lu, Y., Zhou, M., Zhi, D., Zhou, M., Jiang, X., Qiu, R., Ou, Z., Wang, H., Qiu, D., Zhong, M., et al.: The jnu-ifm dataset for segmenting pubic symphysis-fetal head. Data in brief **41**, 107904 (2022) 3, 7
14. Ou, Z., Bai, J., Chen, Z., Lu, Y., Wang, H., Long, S., Chen, G.: Rtseg-net: a lightweight network for real-time segmentation of fetal head and pubic symphysis from intrapartum ultrasound images. Computers in biology and medicine **175**, 108501 (2024) 7
15. Qiu, R., Zhou, M., Bai, J., Lu, Y., Wang, H.: Psfhsp-net: an efficient lightweight network for identifying pubic symphysis-fetal head standard plane from intrapartum ultrasound images. Medical & Biological Engineering & Computing **62**(10), 2975–2986 (2024) 7
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015) 2
17. Zhou, Z., Lu, Y., Bai, J., Campello, V.M., Feng, F., Lekadir, K.: Segment anything model for fetal head-pubic symphysis segmentation in intrapartum ultrasound image analysis. Expert Systems with Applications **263**, 125699 (2025) 7