Unlabeled Data-Driven Fetal Landmark Detection in Intrapartum Ultrasound

Chen Ma$^{[0009-0001-3520-6258]}$, Yunshu Li$^{[0009-0002-4228-266X]}$, Bowen Guo$^{[0009-0007-7740-8425]}$, Jing Jiao$^{[0000-0003-3152-415X]}$, Yi Huang$^{[0000-0001-7834-8689]}$, Yuanyuan Wang, and Yi Guo*

Fudan University, Shanghai, China
guoyi@fudan.edu.cn

Abstract. The angle of progression (AoP) is a critical parameter for assessing fetal head descent during labor, requiring identification of three anatomical landmarks in intrapartum ultrasound images. Manual annotation is time-consuming and prone to inter- and intra-observer variability, while automated methods are hindered by limited labeled data and domain shifts across ultrasound devices. We propose an automated fetal biometry method for AoP calculation based on a modified TransUNet architecture with TinyViT backbone. The design integrates (i) MAE-assisted knowledge distillation from an Ultrasound Foundation Model (USFM) for robust representation learning, (ii) label perturbation to enhance robustness and cross-device generalization, and (iii) semi-supervised learning with pseudo-labeling to leverage unlabeled data. The network predicts heatmaps for landmark localization and calculates AoP from the detected coordinates. On the IUGC2025 Challenge test set, the proposed method achieved a mean radial error of 11.6749 pixels and a mean absolute AoP error of 3.8061 degrees.

Keywords: Intrapartum Ultrasound · Angle of Progression · Keypoint Detection · Pseudo Labels · Model Pretraining.

1 Introduction

Labor monitoring is essential for ensuring maternal and fetal safety during childbirth. The World Health Organization's Labour Care Guide (LCG) emphasizes the importance of systematic assessment of fetal head position and progression for timely clinical decision-making. Among the critical parameters, the angle of progression (AoP) measured from intrapartum ultrasound provides crucial insights into fetal descent and directly influences intervention decisions [6, 12]. The AoP is calculated by identifying three anatomical landmarks: two points along the pubic symphysis (PS1, PS2) and one point where a tangent from PS1 touches the fetal head (FH1).

* Corresponding author: Yi Guo

In recent years, deep learning has shown great promise for analyzing intrapartum ultrasound images, particularly for pubic symphysis and fetal head segmentation. Traditional approaches generally follow a two-stage pipeline: first segmenting the anatomical structures, then calculating the angle of progression. Lu et al. [14] proposed a multitask deep neural network combining image segmentation, endpoint detection, and angle calculation using a shared encoder with multiple decoders. Bai et al. [1] developed a dual-branch network for computing AoP from transperineal ultrasound images. Chen et al. [5] introduced direction-guided and multi-scale feature screening methods for improved segmentation-based AoP computation. Chen et al. [4] proposed a dual-path boundary-guided residual network (DBRN) with attention mechanisms. Furthermore, Ou et al. [15] developed RTSeg-Net, a lightweight model for real-time fetal head–pubic symphysis segmentation in clinical settings.

Despite recent progress, segmentation-first pipelines face notable limitations. First, the multi-stage process enables errors from the segmentation step to propagate to landmark localization and AoP calculation, ultimately degrading final accuracy. Second, segmentation requires dense pixel-level annotations, which are costly to obtain and prone to inter-observer variability. These drawbacks motivate the exploration of direct landmark detection, which can mitigate error accumulation, reduce annotation burden, and improve computational efficiency.

However, direct landmark detection in intrapartum ultrasound introduces its own challenges. The images are inherently noisy with low contrast [7], and substantial domain shifts across devices and manufacturers hinder model generalization [2, 18]. In addition, expert landmark annotation is time-consuming, and the scarcity of labeled datasets makes it difficult to train robust models, especially under cross-device variations.

To address these challenges, we propose a novel automated fetal biometry method for AoP calculation that integrates self-supervised pretraining, lightweight architecture design, and cross-device adaptation.

The main contributions of this paper are as follows:

- A MAE-assisted knowledge distillation framework using USFM as teacher to train a lightweight TinyViT backbone specifically adapted for intrapartum ultrasound;
- A modified TransUNet architecture combining ResNet-50 encoder, transformer bottleneck, and UNet decoder for precise heatmap-based landmark detection;
- Training strategies incorporating cross-device adaptation through label perturbation and semi-supervised learning via iterative pseudo-labeling to leverage abundant unlabeled data.

Our method achieves superior performance compared to baseline approaches, with mean radial error of 11.6749 pixels and absolute parameter difference of 3.8061 degrees on the test set, demonstrating significant potential to streamline clinical workflows and improve diagnostic consistency in intrapartum care.

2 Methods

2.1 Overview

Our approach for automated fetal biometry in intrapartum ultrasound images consists of three main components: (1) a pretraining phase leveraging MAE-assisted knowledge distillation to learn domain-specific features from intrapartum ultrasound data, (2) a modified TransUNet architecture for keypoint detection, and (3) training strategies that incorporate cross-device adaptation and semi-supervised learning. This methodology is designed to address the core challenges of limited labeled data, cross-device generalization, and the spatial relationship modeling crucial for accurate anatomical landmark detection.

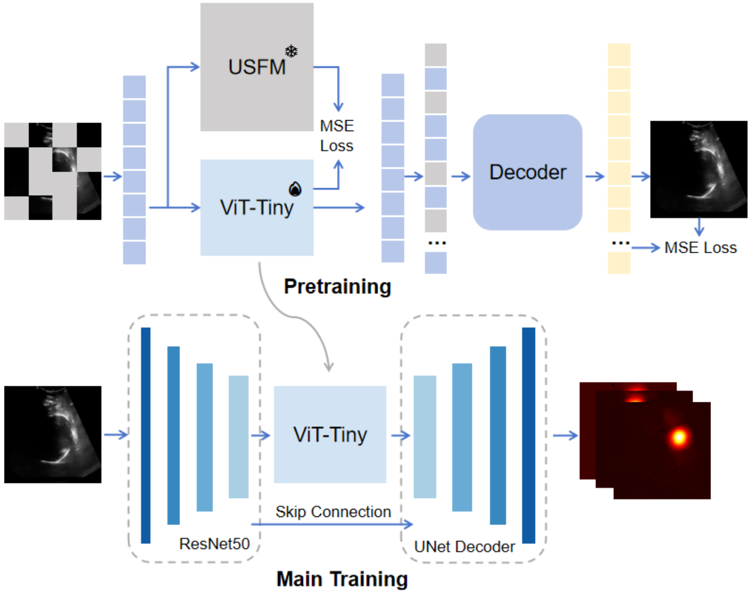

Fig. 1. Overview of the Pretraining Phase and Main Training Phase

2.2 Model Pretraining

To meet the demands of accurate and real-time keypoint detection in intrapartum ultrasound, we adopt a compact TinyViT backbone enhanced via MAE-assisted knowledge distillation, as shown in the upper part of Fig. 1. Specifically, the Ultrasound Foundation Model (USFM) [10] serves as the teacher network, transferring domain-specific anatomical representations to the lightweight student model.

The MAE pretraining method [8], which reconstructs masked image patches, is particularly suited for heatmap-based keypoint detection, as it captures spatial dependencies across image regions and facilitates precise landmark localization. This formulation also enables effective utilization of unlabeled intrapartum ultrasound images to enrich the learned representations.

The overall distillation loss is defined as:

$$\mathcal{L}_{distill} = MSE(f_{student}(x), f_{teacher}(x)) + \lambda \mathcal{L}_{MAE}(x_{masked}) \tag{1}$$

where $f_{student}$ and $f_{teacher}$ represent the feature extraction functions of the TinyViT and USFM models respectively, and L_{MAE} is the masked autoencoder reconstruction loss.

The resulting TinyViT model serves as the backbone for our keypoint detection network, providing a computationally efficient yet powerful feature extractor that has been specifically adapted to the characteristics of transperineal ultrasound images.

2.3 Network Architecture

Our keypoint detection model builds on a modified TransUNet [3] architecture that integrates a ResNet-50 [9] backbone encoder, a ViT-style transformer bottleneck using a distilled TinyViT [17], and a UNet-like decoder with skip connections, designed to predict heatmaps for anatomical landmark localization in intrapartum ultrasound images. As shown in the lower half of Fig. 1, the model structure is mainly divided into the following parts.

Encoder The encoder employs a ResNet-50 pretrained backbone to extract hierarchical feature maps at multiple scales. This multi-scale representation provides rich local details from shallow layers and abstract semantic information from deeper layers, facilitating precise anatomical localization.

TinyViT Bottleneck The last encoder layer is projected to a lower-dimensional embedding and then fed into the pretrained TinyViT from Section 2.2 for further processing. The output tokens are reshaped back into spatial features for decoding.

Decoder The decoder reconstructs high-resolution feature maps through a series of transposed convolutional upsampling layers. At each upsampling stage, features are concatenated with the corresponding encoder features via skip connections to fuse local and global information effectively. Each concatenated feature map is refined by a lightweight double convolution block composed of two sequential convolution, batch normalization, and ReLU activation layers. The decoder progressively upsamples features until the spatial resolution matches the desired heatmap size.

Output Head and Coordinate Extraction A final 1×1 convolution layer projects the decoder's output to K heatmaps ($K = 3$ in our case), each representing the predicted spatial probability distribution of one anatomical landmark. The heatmaps have a fixed spatial resolution of 64×64. Landmark coordinates are extracted by locating the spatial maxima of each heatmap via an argmax operation, followed by normalization to relative coordinates.

Loss Funtion The model is trained using a single heatmap loss:

$$\mathcal{L}_{heatmap} = MSE(heatmap_{pred}, heatmap_{gt}) \tag{2}$$

where $heatmap_{pred}$, and $heatmap_{gt}$ denote the predicted and ground-truth heatmaps, respectively. The loss is computed as mean squared error (MSE) between them.

2.4 Training Strategies

Device-Domain Adaption The dataset presents a significant domain shift, as the training images are acquired from two ultrasound machines, whereas the test set consists of images from two different devices. To improve the model's robustness across domains, we apply label perturbation during training by injecting Gaussian noise into the ground truth landmark coordinates:

$$(\tilde{x}_i, \tilde{y}_i) = (x_i, y_i) + \mathcal{N}(0, \sigma^2 I) \tag{3}$$

where $\sigma = 2$ pixels. This perturbation acts as a regularizer, encouraging the network to learn features that are invariant to small spatial variations caused by differences in machine calibration and imaging protocols, thus enhancing generalization to unseen domains.

Pseudo Labeling Given the substantial imbalance between labeled and unlabeled images, we employ iterative pseudo-labeling to leverage the abundant unlabeled data. We generate pseudo-labels using the device-domain-adapted model and select high-quality samples based on multiple criteria: (1) prediction confidence measured by heatmap peak sharpness and (2) geometric plausibility enforcing anatomical constraints between landmarks.

3 Experiments

3.1 Datasets

The challenge organizers provide a training set comprising 300 labeled intrapartum ultrasound images, each annotated with three anatomical landmarks (PS1, PS2, FH1) and the angle of progression (AoP). The corresponding annotations are stored in a CSV file containing the image filenames, landmark coordinates (x, y), and AoP values. In addition, the dataset includes 31,421 unlabeled intrapartum ultrasound images. To facilitate standard-plane identification, 2,045 reference images depicting the standard acquisition view are also provided.

We further incorporate the FH-PS-AoP public dataset [11], containing 4,000 annotated intrapartum ultrasound images at a native resolution of 256 × 256, for model pretraining. The detailed data statistics are shown in Table 1.

For each main training phase, the available labeled data are randomly split into training and test subsets with a ratio of 4:1.

Table 1. Summary of datasets used in this study.

Dataset	**# Images**	**Annotations**	**Resolution**
Labeled cases	300	PS1, PS2, FH1, AoP	512 × 512
Unlabeled cases	31,421	None	512 × 512
Standard-plane examples	2,045	None	512 × 512
FH-PS-AoP	4,000	None	256 × 256

Preprocessing All images are resized to 224×224 during pretraining. For main training and evaluation phases, images are resized to 512 × 512, and landmark coordinates are scaled accordingly.

For each annotated landmark, we generate a Gaussian heatmap representation to serve as the regression target. Given the landmark location (x_0, y_0), the heatmap H at pixel location (x, y) is defined as:

$$H(x,y) = exp(-\frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma^2}) \tag{4}$$

where $\sigma = 4$ pixels controls the spatial spread. This continuous spatial encoding provides localized supervision, allowing the network to learn more precise keypoint localization compared to direct coordinate regression.

Data Augmentation To improve robustness to acquisition variability and mitigate overfitting, we employ both geometric and photometric augmentations.

Geometric: In-plane rotations, and random scaling with cropping are applied jointly to images and landmark coordinates to preserve spatial alignment, with AoP values recalculated from the transformed points.

Photometric: Gamma correction and contrast adjustment are applied to simulate device- and operator-induced appearance variations, without altering landmark positions.

This augmentation strategy strengthens the model's resilience to spatial perturbations and intensity variations, facilitating cross-device generalization in intrapartum ultrasound analysis.

3.2 Experimental Settings

During the pretraining phase, we set the probability of spatial domain masking at 0.75 and employed the AdamW optimizer with an initial learning rate of 1e-4, complemented by a cosine learning rate decay strategy. The experiments were trained for 400 epochs with a batch size of 1024.

In the main training, we employed the Adam optimizer with an initial learning rate of 3e-5 and adopt a StepLR learning rate decay strategy, training for 100 epochs with a batch size of 4.

All experiments were conducted on an AMD EPYC 7763 CPU and NVIDIA A100 GPU. All models were developed using PyTorch.

4 Results and Discussion

4.1 Quantitative Results on Validation Phase

We evaluate all methods on two sets of metrics: one computed on our own test split, and another obtained from the challenge platform's validation phase. Both sets report **Mean Radial Error (MRE)**—the average Euclidean distance between predicted and ground-truth landmarks—and **Absolute Parameter Difference (APD)**—the absolute difference in the predicted AoP angle. Lower values indicate better performance.

Table 2 summarizes the results. Our proposed method consistently outperforms the baseline(a UNet [16] based model) across both test and validation sets. Pretraining on external datasets and leveraging pseudo-labels further improve accuracy, with label perturbation providing additional robustness and the best overall results. As illustrated in Fig. 2, the heatmaps generated by our method show significantly improved localization capability and reduced uncertainty compared to other models.

Table 2. Comparison of different methods on test split and challenge platform validation phase. Metrics reported are Mean Radial Error (MRE) in pixels and Absolute Parameter Difference (APD) in degrees; lower is better. (PL: pseudo-labeling; DDA: device-domain adaptation)

Method	**Test Split**		**Validation Phase**	
	MRE ↓	APD ↓	MRE ↓	APD ↓
Baseline (Heatmap U-Net)	14.3266	6.2033	22.8510	8.8178
TransUNet	14.1093	5.8906	20.3387	7.4982
TransUNet + Pretrain	13.5632	5.5691	17.5204	6.6031
TransUNet + Pretrain + PL	12.8841	5.2377	15.6327	5.9873
TransUNet + Pretrain + PL + DDA	**12.3115**	**4.9815**	**14.3584**	**5.1569**

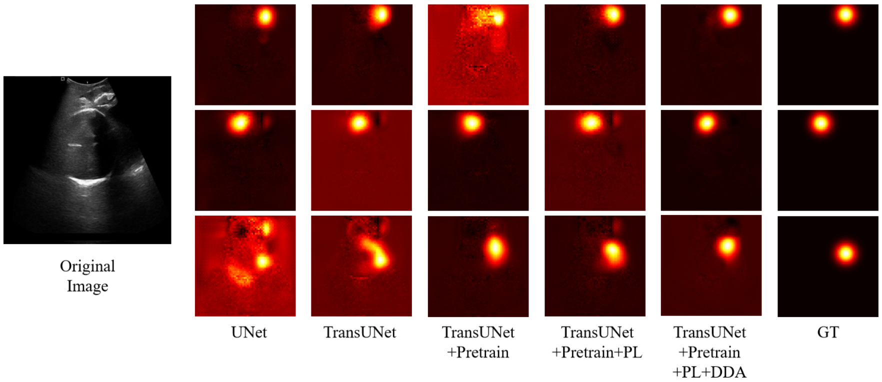

Fig. 2. Comparison of heatmaps generated by different methods.(PL: pseudo-labeling; DDA: device-domain adaptation)

4.2 Quantitative Results on Test Phase

We evaluated our final model, which incorporates pretraining, pseudo-labeling, and device-domain adaptation, on the final test phase. The performance metrics include the Mean Radial Error (MRE) between predicted and ground truth landmark points, and the Absolute Parameter Difference (APD) for the angle of progression (AoP). Results, as shown in Table 3, demonstrate that the model achieves robust and accurate landmark localization and AoP estimation on unseen test data.

Table 3. Test phase performance of the final model. Metrics reported are Mean Radial Error (MRE) in pixels and Absolute Parameter Difference (APD) in degrees; lower is better.

Model	**MRE ↓**	**APD ↓**
TransUNet + Pretrain + PL + DDA	**11.6749**	**3.8061**

4.3 Limitation and Future Work

Despite the promising results, our device-domain adaptation (DDA) method exhibits some instability. This is primarily due to the random noise introduced in the label perturbation process, which can cause fluctuations in training performance and occasionally result in poor outcomes on certain samples. For example, a few images in the validation phase showed significantly degraded performance.

In future work, we plan to explore more stable and robust unsupervised domain adaptation techniques to mitigate this issue.

Additionally, constrained by the inference time limitations imposed by the challenge and the need for rapid model iteration, we selected a lightweight TinyViT model as our backbone. We also experimented with larger Vision Transformer variants such as USFM and SAM [13], integrated as backbones within the TransUNet architecture. These larger models achieved performance comparable to the TinyViT model, which benefited from MAE-assisted knowledge distillation specialized for intrapartum ultrasound images, despite not undergoing the distillation process themselves. This not only validates the effectiveness of our MAE-assisted distillation strategy but also suggests that further optimizing these larger models specifically for intrapartum data may yield even better results. Future research will focus on targeted enhancements of USFM and SAM models on intrapartum ultrasound images to further boost performance.

5 Conclusion

In this paper, we proposed a unified framework for precise anatomical landmark localization and angle of progression estimation in intrapartum ultrasound images. Our method incorporates a lightweight TinyViT backbone—pretrained with knowledge distillation on USFM using related ultrasound data—integrated within a heatmap TransUNet architecture. Together with pseudo-labeling and device-domain adaptation strategies, this approach significantly enhances accuracy and generalization across varied datasets. Experimental results demonstrate strong performance and efficiency, highlighting the potential of our method for practical intrapartum ultrasound analysis.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (82227803, 62371139), and the Foundation of the Shanghai Municipal Education Commission (24KXZNA09).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 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)
2. Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A.: Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE journal of biomedical and health informatics **19**(5), 1627–1636 (2015)
3. 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)

4. 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)
5. 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)
6. Dall'Asta, A., Angeli, L., Masturzo, B., Volpe, N., Schera, G.B.L., Di Pasquo, E., Girlando, F., Attini, R., Menato, G., Frusca, T., et al.: Prediction of spontaneous vaginal delivery in nulliparous women with a prolonged second stage of labor: the value of intrapartum ultrasound. American journal of obstetrics and gynecology **221**(6), 642–e1 (2019)
7. Duarte-Salazar, C.A., Castro-Ospina, A.E., Becerra, M.A., Delgado-Trejos, E.: Speckle noise reduction in ultrasound images for improving the metrological evaluation of biomedical applications: an overview. IEEE Access **8**, 15983–15999 (2020)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical image analysis **96**, 103202 (2024)
11. Jieyun, B., ZhanHong, O.: Pubic symphysis-fetal head segmentation and angle of progression (Apr 2023), https://doi.org/10.5281/zenodo.7851339
12. Kalache, K.D., Dückelmann, A., Michaelis, S., Lange, J., Cichon, G., Dudenhausen, J.: Transperineal ultrasound imaging in prolonged second stage of labor with occipitoanterior presenting fetuses: how well does the 'angle of progression'predict the mode of delivery? Ultrasound in Obstetrics and Gynecology **33**(3), 326–330 (2009)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. 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)
15. 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)
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. Springer (2015)
17. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)

18. Zhou, M., Wang, C., Lu, Y., Qiu, R., Zeng, R., Zhi, D., Jiang, X., Ou, Z., Wang, H., Chen, G., et al.: The segmentation effect of style transfer on fetal head ultrasound image: a study of multi-source data. Medical & biological engineering & computing **61**(5), 1017–1031 (2023)