Heatmap Regression for Automated Angle of Progression Measurement: The Baseline Method for the IUGC2025

Yitong Tang[1], Zihao Zhou[1], Yaosheng Lu[1], Jieyun Bai[1,⋆], Shun Long[1], Yuxin Huang[2], Isaac Khobo[3], Shun Zhang[2], Zimo Zhou[2], and Lei Guo[2]

[1] College of Information Science and Technology, Jinan University, Guangzhou 510632, China
[2] Department of Obstetrics and Gynecology, Zhujiang Hospital, Southern Medical University, Guangzhou 510260, China
[3] University of Cape Town, Rondebosch, Cape Town, 7701, South Africa

Abstract. Angle of Progression (AoP) is a critical parameter for clinical assessment of fetal head descent and prediction of delivery mode, traditionally measured manually by experienced clinicians, which leads to efficiency and consistency issues. In this paper, we present a heatmap regression-based keypoint detection method as a baseline approach for the Intrapartum Ultrasound Grand Challenge (IUGC) 2025, designed to automatically measure the AoP in intrapartum ultrasound images. We employ a U-Net architecture for heatmap prediction to directly identify the three key points required for AoP measurement, followed by post-processing to extract precise coordinates. The method was evaluated on the IUGC 2025 dataset, trained with 300 annotated samples and tested on 501 samples, achieving an average AoP error of 8.37° and a MRE of 21.83 pixels. As a baseline method, we discuss current limitations and propose improvement directions, including semi-supervised learning to leverage unlabeled data, adoption of more advanced network architectures, and optimization of post-processing techniques. This study demonstrates the feasibility of automated AoP measurement in obstetric ultrasound imaging, potentially improving decision support tools in obstetric clinical practice.

Keywords: Intrapartum Ultrasound · Angle of Progression · Keypoint Detection · Heatmap Regression · U-Net

1 Introduction

The delivery modalities are primarily bifurcated into vaginal delivery and cesarean section[1]. The former respects maternal physiological mechanisms and demonstrates lower morbidity and mortality indices for the maternal-fetal dyad,

⋆ Corresponding Author: Jieyun Bai (jbai996@aucklanduni.ac.nz)

whereas the latter represents an alternative intervention when maternal or fetal pathophysiology precludes vaginal parturition[2]. Optimizing maternal and neonatal outcomes through reduction of unnecessary operative interventions while ensuring timely cesarean deliveries necessitates precise assessment of labor progression in contemporary obstetric practice.

Traditional labor monitoring methodologies predominantly utilize digital vaginal examinations, which the World Health Organization advocates performing at 4-hour intervals during the first stage of labor[3]. However, substantial evidence indicates that vaginal assessment of fetal head station and position demonstrates limited accuracy and significant subjectivity, particularly when cephalohematoma impedes palpation of cranial sutures and fontanels[4,5]. Moreover, repeated examinations potentially facilitate ascending microbial migration from the vagina to the cervix and uterus, presenting potential neonatal infectious risks[5].

Intrapartum ultrasonography has emerged as a superior methodological alternative for labor progression evaluation. Multiple investigations have demonstrated that sonographic measurements exhibit enhanced accuracy, objectivity, and reproducibility compared with digital examination[6,7]. Furthermore, ultrasonographic assessment neither elicits patient discomfort nor requires substantial additional clinical time[8]. Among various sonographic parameters,AoP has been identified as the most reproducible parameter for evaluating fetal head descent[9].

The AOP is defined as the angle formed by the two farthest points (PS1 and PS2) along the pubic symphysis contour and the point of tangency (FH1) where a tangent line drawn from the rightmost point (PS1) touches the fetal head .This measurement provides critical information regarding both the current position of the fetal head relative to the ischial spines and the trajectory of labor progression[10]. Research demonstrates that an AoP exceeding 120 degrees correlates significantly with successful spontaneous vaginal delivery probability, establishing it as a valuable predictive indicator of delivery modality[11].

The contemporary AoP measurement paradigm predominantly relies on manual assessments performed by experienced clinicians—a methodology characterized by temporal inefficiency and potential measurement inconsistencies[12]. The development of automated algorithms for AoP quantification therefore presents a significant opportunity to enhance both efficiency and precision in clinical labor assessment protocols. Nevertheless, significant challenges persist in the accurate segmentation of relevant anatomical structures due to inherent ultrasonographic limitations, including speckle noise, attenuation, artifacts, and suboptimal signal-to-noise ratios[13]. Additionally, transperineal ultrasound images frequently exhibit blurred anatomical targets, indistinct contours, and interference from adjacent tissues[14], with fetal head boundaries particularly susceptible to delineation difficulties due to normal sutures, sonographic artifacts, or interference from the similarly echogenic uterine wall[15].

Previous research on automated AoP measurement has primarily adopted segmentation-based methods, which first perform complete segmentation of the

pubic symphysis and fetal head contours in ultrasound images, and then calculate AoP based on the segmentation results.

Within these segmentation-based frameworks, the predominant strategy for extracting the keypoints (PS1, PS2, and FH1) involves elliptical modeling of the anatomical structures. Specifically, the segmented contours of the pubic symphysis and fetal head are typically subjected to ellipse fitting algorithms [16,17,18,19]. While some implementations perform dual ellipse fitting for both structures to derive the pubic symphysis endpoints (PS1 and PS2) and the fetal head tangent point (FH1) [16,19], others employ a hybrid approach by directly regressing the pubic symphysis endpoints (PS1 and PS2) and applying ellipse fitting only to the fetal head to determine FH1 [17,18]. Despite its widespread adoption, the ellipse fitting process constitutes an additional source of measurement error. This error arises because the anatomical structures may not conform perfectly to elliptical shapes, and the fitting accuracy is highly contingent upon segmentation quality. Moreover, the least-squares fitting algorithms themselves introduce numerical approximations. Consequently, these errors propagate to the subsequent AoP computation, potentially compromising measurement precision. Notably, all existing automated methodologies rely on intermediate segmentation and/or geometric modeling steps, with none directly regressing the three key coordinate points required for AoP computation.

Recent advancements in computational obstetrics have sought to address these segmentation challenges through alternative methodological paradigms. The IUGC 2025 Challenge (hosted at MICCAI) proposes a shift toward keypoint detection-based AoP measurement, departing from conventional segmentation-dependent approaches.This challenge focuses on keypoint detection , which directly identifies the three key coordinate points (PS1, PS2, and FH1) required for AoP measurement to calculate process parameters. The present study provides a comprehensive description of the baseline approach employed in IUGC2025 to enhance participants' comprehension of the methodological details.

2 Method

2.1 Overview of Network

This study employs a heatmap-based regression approach for landmark detection in intrapartum ultrasound images, enabling precise measurement of the AoP. Our methodology addresses the challenge of directly identifying the three critical landmarks (PS1, PS2, and FH1) required for AoP calculation through heatmap prediction and coordinate extraction.

Our approach employs a three-stage pipeline:

- **Label Preprocessing**: During training data loading, ground truth heatmaps are generated by encoding landmark coordinates as Gaussian distributions centered at each keypoint location.
- **Heatmap Prediction**: These heatmaps serve as training targets for a U-Net architecture, which learns to predict pixel-wise likelihood maps for keypoint presence.

– **Coordinate Extraction**: At inference time, predicted heatmaps undergo post-processing to extract precise landmark coordinates.

Fig. 1 illustrates the complete workflow of our proposed heatmap-based landmark detection system, which consists of the following key components: label preprocessing based on Gaussian distributions, heatmap prediction via U-Net, coordinate extraction and AoP calculation.

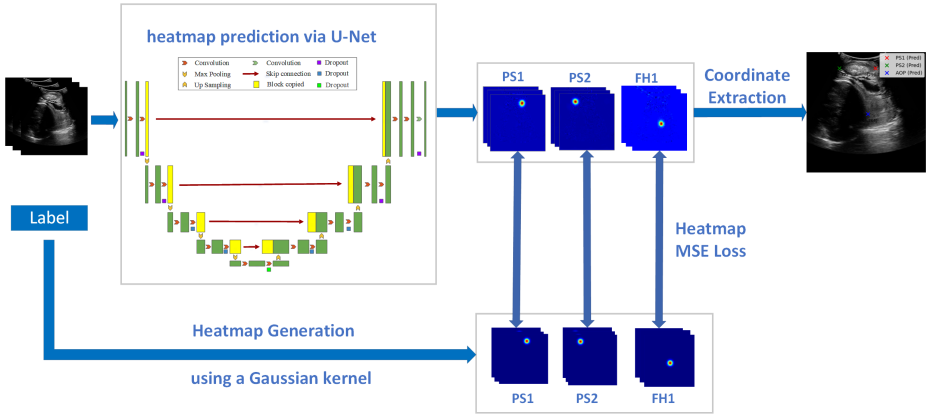


Fig. 1. Workflow of heatmap-based landmark detection system

2.2 Label Preprocessing

The network is trained to predict heatmaps that represent the spatial probability distribution of landmark locations. For each of the three landmarks (PS1, PS2, and FH1), a ground truth heatmap is generated using a Gaussian kernel centered at the annotated landmark coordinates[20].

The ground truth heatmap for each landmark is generated as follows:

$$H(x, y) = \exp\left(-\frac{(x - x_0)^2 + (y - y_0)^2}{2\sigma^2}\right) \tag{1}$$

where (x_0, y_0) represents the ground truth landmark coordinates scaled to the heatmap dimensions, and σ controls the spread of the Gaussian peak. We empirically set $\sigma = 2.0$ to balance between localization precision and network trainability.

The process of generating ground truth heatmaps involves:

1. Normalizing input images to dimensions of 512×512 pixels
2. Scaling annotated landmark coordinates (PS1, PS2, FH1) to match the heatmap dimensions (64×64)

3. Generating a separate Gaussian heatmap for each landmark using Equation 1
4. Ensuring that heatmap values range between 0 and 1, with the peak value of 1 at the landmark location

This representation transforms the discrete landmark detection problem into a continuous heatmap regression task, which proves advantageous in handling the inherent noise and ambiguity in ultrasound images. The Gaussian distribution accommodates slight annotation variations and provides a smoother optimization landscape during training.

2.3 Heatmap Prediction

Network Architecture A fully convolutional U-Net architecture is employed to predict heatmaps. This architecture is specifically designed for generating high-resolution feature maps while preserving spatial information crucial for precise landmark localization.

The network encompasses an encoder-decoder structure with skip connections, comprising four core components: an encoder pathway, a bottleneck, a decoder pathway, and an output layer. This structure facilitates gradient flow and feature reuse throughout the network.

Encoder Pathway: The encoder comprises four sequential blocks. Each block contains two 3×3 convolutional layers with batch normalization and ReLU activation, followed by 2×2 max pooling. The network begins with 64 channels, which double after each pooling operation, reaching 512 channels at the deepest encoder level. This progressive channel expansion enables the extraction of increasingly complex features while reducing spatial dimensions from 512×512 to 32×32.

Bottleneck: The bottleneck serves as a transition between encoder and decoder pathways, consisting of two convolutional layers with 1024 channels that capture the most abstract features with the broadest receptive field.

Decoder Pathway: The decoder mirrors the encoder with four sequential upsampling blocks. Each block begins with a 2×2 transposed convolution that doubles the spatial dimensions while halving the channel count. Features from the corresponding encoder level are concatenated via skip connections, followed by two 3×3 convolutional layers with batch normalization and ReLU activation. This architecture facilitates the gradual recovery of spatial detail while integrating multi-scale contextual information from the encoder.

Output Layer: The final layer consists of a 1×1 convolution that reduces the channel dimension to match the number of keypoints (three in our case), producing three separate heatmap channels corresponding to PS1, PS2, and FH1 landmarks.

The network architecture ensures that the output heatmaps maintain a consistent size of 64×64 pixels, achieving an optimal balance between computational efficiency and localization precision. The architectural parameters are summarized in Table 1.

Table 1. U-Net Architecture for Landmark Heatmap Regression

Layer	Output Size	Parameters
Input	3×512×512	-
Encoder Block 1	64×256×256	Conv(3→64), Conv(64→64), MaxPool
Encoder Block 2	128×128×128	Conv(64→128), Conv(128→128), MaxPool
Encoder Block 3	256×64×64	Conv(128→256), Conv(256→256), MaxPool
Encoder Block 4	512×32×32	Conv(256→512), Conv(512→512), MaxPool
Bottleneck	1024×32×32	Conv(512→1024), Conv(1024→1024)
Decoder Block 4	512×64×64	ConvTranspose(1024→512), Conv(1024→512), Conv(512→512)
Decoder Block 3	256×128×128	ConvTranspose(512→256), Conv(512→256), Conv(256→256)
Decoder Block 2	128×256×256	ConvTranspose(256→128), Conv(256→128), Conv(128→128)
Decoder Block 1	64×512×512	ConvTranspose(128→64), Conv(128→64), Conv(64→64)
Output	3×64×64	Conv(64→3), Resize

Loss Function We employ Mean Squared Error (MSE)[21] as the primary loss function for training our heatmap regression network. The MSE loss measures the pixel-wise difference between the predicted heatmaps and the ground truth Gaussian heatmaps:

$$\mathcal{L}_{MSE} = \frac{1}{NKP} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{p=1}^{P} (H_{n,k,p}^{pred} - H_{n,k,p}^{gt})^2 \tag{2}$$

where N represents the batch size, K denotes the number of landmarks (3 in our case), P is the number of pixels in each heatmap (64×64), $H_{n,k,p}^{pred}$ is the predicted heatmap value, and $H_{n,k,p}^{gt}$ is the ground truth heatmap value for the n-th sample, k-th landmark, and p-th pixel.

MSE loss is particularly suitable for heatmap regression as it penalizes large deviations more severely than small ones, encouraging the network to produce precise peaks at landmark locations. Additionally, it provides a stable gradient flow during training, facilitating convergence even with limited training data.

2.4 Coordinate Extraction

Following heatmap prediction, we extract precise landmark coordinates through post-processing.

Maximum Response Location The simplest approach identifies the pixel with the maximum value in each heatmap channel:

$$(x^{pred}, y^{pred}) = \arg\max_{(x,y)} H(x, y) \tag{3}$$

where $H(x, y)$ represents the predicted heatmap value at location (x, y). The resulting integer coordinates are then normalized by dividing by the heatmap

dimensions to obtain values in the range [0,1], which can be mapped back to the original image coordinates.

The extracted coordinates for all three landmarks (PS1, PS2, and FH1) are then denormalized to the original image dimensions and used for calculating the AOP.

2.5 Training Details

The network was trained using an Adam optimizer with an initial learning rate of 10^{-4} and weight decay of 10^{-4} to prevent overfitting. We implemented a step learning rate scheduler that reduces the learning rate by a factor of 0.5 every 15 epochs, facilitating convergence in later training stages.

Training was conducted for 150 epochs with a batch size of 4 on standardized ultrasound images resized to 512×512 pixels. Data augmentation techniques were deliberately minimal to preserve the anatomical integrity of the ultrasound images, which is crucial for accurate landmark detection.

To monitor training progress and prevent overfitting, we tracked both the heatmap loss and the coordinate distance metrics. Model checkpoints were saved at regular intervals (every 50 epochs) as well as when achieving the best performance on either the training loss or coordinate distance metrics.

The training process leveraged TensorBoard for real-time visualization of loss curves, learning rates, and sample predictions, enabling continuous assessment of model convergence. Training was conducted on a NVIDIA RTX 2080Ti GPU.

3 Experiments and Results

3.1 Evaluation Metrics

We evaluated our model using several complementary metrics to assess landmark detection accuracy and clinical applicability:

Mean Radial Error (MRE) MRE is adopted as the primary metric, quantifying the average Euclidean distance (in pixels) between predicted landmarks (x_p, y_p) and ground truth landmarks (x_g, y_g) . For each landmark, the radial error R_i is computed as:

$$R_i = \sqrt{(x_p - x_g)^2 + (y_p - y_g)^2} \tag{4}$$

The MRE across N landmarks are then defined as:

$$MRE = \frac{1}{N} \sum_{i=1}^{N} R_i \tag{5}$$

All values are reported in pixel space (range 0–512). Lower MRE values indicate superior localization accuracy.

AOP Error Given the clinical significance of AoP, we directly calculate the absolute difference between the angles derived from predicted landmarks and ground truth landmarks:

$$\Delta AoP = |AoP^{pred} - AoP^{gt}| \tag{6}$$

The AoP is calculated using the law of cosines:

$$AoP = \cos^{-1}\left(\frac{a^2 + b^2 - c^2}{2ab}\right) \cdot \frac{180}{\pi} \tag{7}$$

where a is the distance between PS1 and PS2, b is the distance between PS1 and FH1, and c is the distance between PS2 and FH1.

Our evaluation protocol provides a comprehensive assessment of model performance, combining pixel-level accuracy with clinically relevant angular measurements to ensure that the proposed method meets both technical and practical requirements for automated AoP determination in clinical settings.

3.2 Datasets

The IUGC2025 challenge dataset comprises 31,421 intrapartum ultrasound images collected from multiple clinical centers, with the following official division:

- Training set: 31,421 cases (including 300 annotated samples with landmark coordinates)
- Validation set: 100 fully annotated cases
- Test set: 501 fully annotated cases

Our baseline method utilizes only the labeled portion of the training data (300 annotated samples) for supervised learning, intentionally excluding unlabeled data to establish a fundamental performance benchmark.

This experimental design intentionally limits the baseline model to supervised learning from scarce annotated data, demonstrating the fundamental feasibility of landmark detection while highlighting potential improvements through semi-supervised approaches that could leverage the substantial unlabeled data.

3.3 Results

The proposed methodology was evaluated on the official IUGC2025 challenge dataset comprising 601 annotated intrapartum ultrasound images, with 100 samples allocated for validation and 501 for testing. All experiments were conducted under standardized conditions using a single NVIDIA RTX 2080Ti GPU, ensuring consistency in computational resources and environmental parameters. We evaluated the proposed heatmap-based landmark detection model on the provided validation and test datasets.

Table 2. Comprehensive evaluation metrics on validation and test sets.

Metric	**Validation (N=100)**	**Test (N=501)**
PS1 MRE (pixels)	12.3408	10.6720
PS2 MRE (pixels)	21.5383	15.6234
FH1 MRE (pixels)	48.1807	39.1866
MRE of all landmarks (pixels)	27.35	21.83
coordinates MAE(pixels)	16.8517	14.0043
Mean AoP Error (°)	10.47	8.37

Quantitative Results Table 2 presents the comprehensive evaluation metrics for both validation and test sets. The model demonstrates robust performance across all metrics, with particularly notable results in clinical AoP measurement accuracy.

The performance was assessed using the metrics described in Section 3.2. The MRE for each individual landmark (PS1, PS2, FH1) are reported, demonstrating detailed localization accuracy across different anatomical points.Additionally, we introduce average Mean Absolute Error (MAE) computed for the coordinates of three landmarks to provide a more comprehensive assessment of localization precision.

The method achieved a mean AoP error of 8.37° on the test set, with average landmark localization(disantce) accuracy of 21.83 pixels.

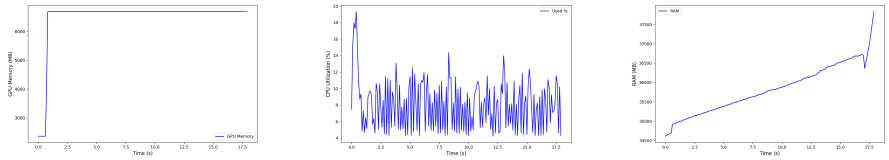
Computational Efficiency Analysis While accuracy is a primary metric for clinical applications, computational efficiency is important for real-time assessment in clinical environments. We conducted a comprehensive analysis of computational resource utilization[22] during inference on the test set, providing valuable benchmarks for future lightweight model development efforts. These efficiency metrics, though not part of the challenge evaluation criteria, establish important baselines for subsequent optimization research.

The area under curve (AUC) metrics provide an integrated view of resource consumption over time, offering a more comprehensive assessment than peak values alone.

Table 3. Computational efficiency metrics for the baseline model.

Metric	**Value**
Runtime (s)	17.66
Maximum GPU Memory (MB)	6,694
GPU Memory-Time AUC (MB·s)	117,188
Maximum CPU Utilization (%)	19.32
CPU Utilization-Time AUC (%·s)	144.96
Maximum RAM Usage (MB)	37,831.25
RAM-Time AUC (MB·s)	644,770

The resource utilization was monitored during model inference using the test set, capturing GPU memory usage, CPU utilization, and RAM consumption over time. Table 3 summarizes the key computational efficiency metrics.



(a) GPU Memory Usage (b) CPU Utilization (c) RAM Consumption

Fig. 2. Time-series visualization of computational resource utilization during model inference.

Figure 2 visualizes the resource usage patterns during model inference. The time-series plots demonstrate the evolving computational demands throughout the inference process, with notable observations including a rapid increase in GPU memory allocation during model initialization, followed by stable utilization during inference, and progressive RAM consumption over time.

Qualitative Analysis In addition to quantitative metrics, we generate visual overlays of predicted landmarks and heatmaps on test images to facilitate qualitative assessment. These visualizations enable expert evaluation of predicted landmark placements and identification of systematic errors.

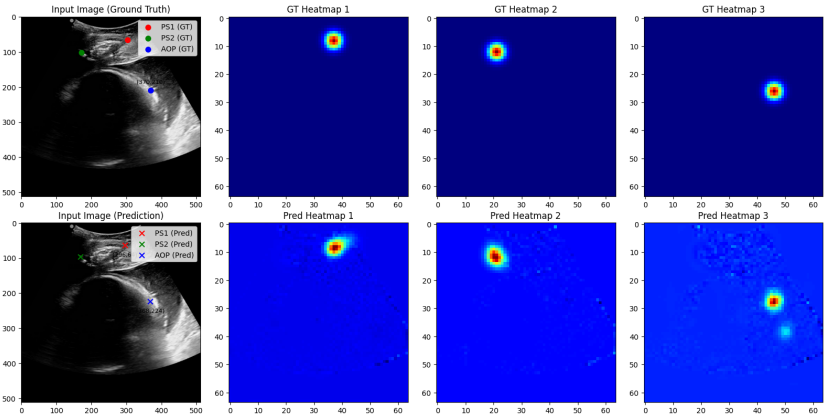


Fig. 3. Comparison of Ground Truth annotations and model predictions

In Figure 3 ,we illustrate comparison of Ground Truth annotations (top row) and model predictions (bottom row) for ultrasound image feature point detection. The Ground Truth group (left: ultrasound image and three heatmaps) marks red (PS1), green (PS2), and blue (FH1) dots; the Prediction group (left: ultrasound image and three heatmaps) indicates predicted points with "×" symbols in corresponding colors. Heatmaps transition from blue (low probability) to red/yellow (high probability), showing differences in localization confidence.

4 Discussion

The baseline method presented in this paper demonstrates the effectiveness of heatmap-based keypoint detection for AoP measurement. As a baseline approach, it has limitations that provide participants with room for improvement.

The main limitations of the baseline method include using only 300 annotated samples for training, employing a standard U-Net[23] architecture, implementing simple post-processing methods, and inadequate consideration of clinical anomalies.

Participants can explore improvement directions including: leveraging the large amount of unlabeled data through semi-supervised learning, adopting more advanced network architectures , integrating anatomical prior knowledge, and optimizing post-processing techniques.

Additionally, our computational efficiency analysis (runtime: 17.66s, GPU memory: 6,694MB, RAM: 37,831MB) suggests potential for optimization through techniques like pruning and quantization. Future work should explore lightweight architectures while maintaining clinical accuracy.

Technical challenges primarily include domain generalization capability, adaptation to anatomical variability, achieving real-time performance, and improving model explainability. We expect participants to focus not only on improving technical metrics but also on the clinical applicability and robustness of their solutions.

5 Conclusion

This study presents a baseline method for AoP measurement in intrapartum ultrasound images based on heatmap regression for keypoint detection. Despite using limited annotated data, the method achieves an average AoP error of 8.37° and a MRE of 21.83 pixels, demonstrating the feasibility of automated AoP measurement.

The IUGC2025 challenge aims to advance artificial intelligence technologies in obstetric ultrasound imaging, particularly through keypoint detection for more accurate and objective labor assessment. This baseline method provides participants with a starting point, showcasing both the potential and limitations of foundational approaches.

We encourage participants to explore innovative methodologies, leverage the large amount of unlabeled data, integrate clinical knowledge, and consider diverse clinical scenarios. Successful solutions will directly contribute to the advancement of obstetric clinical practice by providing more precise decision support tools, ultimately improving maternal and neonatal health outcomes.

Acknowledgements. This work was supported by the Natural Science Foundation of Guangdong Province (2023A1515012833 to J.B.; 2024A1515011886 to J.B), the National Natural Science Foundation of China (61901192 to J.B.), High-end Foreign Experts Recruitment Plan of China (H20240205 to J.B.), China Scholarship Council (202206785002 to J.B.) and the Guangdong Health Economics Association (2025-WJHX-17) .

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

References

1. Rosenberg, K.R., Trevathan, W.R.: Evolutionary perspectives on cesarean section. Evolution, Medicine, and Public Health **2018**(1), 67–81 (2018)
2. Gregory, K.D., Jackson, S., Korst, L., Fridman, M.: Cesarean versus vaginal delivery: whose risks? whose benefits? American journal of perinatology **29**(01), 07–18 (2012)
3. Sandall, J., Tribe, R.M., Avery, L., Mola, G., Visser, G.H., Homer, C.S., Gibbons, D., Kelly, N.M., Kennedy, H.P., Kidanto, H., et al.: Short-term and long-term effects of caesarean section on the health of women and children. The Lancet **392**(10155), 1349–1357 (2018)
4. Seval, M.M., Yuce, T., Kalafat, E., Duman, B., Aker, S., Kumbasar, H., Koc, A.: Comparison of effects of digital vaginal examination with transperineal ultrasound during labor on pain and anxiety levels: a randomized controlled trial (2016)
5. Tutschek, B., Torkildsen, E., Eggebø, T.: Comparison between ultrasound parameters and clinical examination to assess fetal head station in labor. Ultrasound in Obstetrics & Gynecology **41**(4), 425–429 (2013)
6. Tutschek, B., Braun, T., Chantraine, F., Henrich, W.: A study of progress of labour using intrapartum translabial ultrasound, assessing head station, direction, and angle of descent. BJOG: An International Journal of Obstetrics & Gynaecology **118**(1), 62–69 (2011)
7. Bellussi, F., Ghi, T., Youssef, A., Cataneo, I., Salsi, G., Simonazzi, G., Pilu, G.: Intrapartum ultrasound to differentiate flexion and deflexion in occipitoposterior rotation. Fetal Diagnosis and Therapy **42**(4), 249–256 (2017)
8. Malvasi, A., Tinelli, A., Barbera, A., Eggebø, T., Mynbaev, O., Bochicchio, M., Pacella, E., Di Renzo, G.: Occiput posterior position diagnosis: vaginal examination or intrapartum sonography? a clinical review. The Journal of Maternal-Fetal & Neonatal Medicine **27**(5), 520–526 (2014)
9. Youssef, A., Salsi, G., Montaguti, E., Bellussi, F., Pacella, G., Azzarone, C., Farina, A., Rizzo, N., Pilu, G.: Automated measurement of the angle of progression in labor: a feasibility and reliability study. Fetal Diagnosis and Therapy **41**(4), 293–299 (2017)

10. Youssef, A., Brunelli, E., Azzarone, C., Di Donna, G., Casadio, P., Pilu, G.: Fetal head progression and regression on maternal pushing at term and labor outcome. Ultrasound in Obstetrics & Gynecology **58**(1), 105–110 (2021)
11. Ghi, T., Eggebø, T., Lees, C., Kalache, K., Rozenberg, P., Youssef, A., Salomon, L., Tutschek, B.: Isuog practice guidelines: intrapartum ultrasound. Ultrasound in Obstetrics & Gynecology **52**(1), 128–139 (2018)
12. 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)
13. Rueda, S., Fathima, S., Knight, C.L., Yaqub, M., Papageorghiou, A.T., Rahmatullah, B., Foi, A., Maggioni, M., Pepe, A., Tohka, J., et al.: Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge. IEEE Transactions on medical imaging **33**(4), 797–813 (2013)
14. Dietz, H.P.: Ultrasound imaging of the pelvic floor. part i: two-dimensional aspects. Ultrasound in Obstetrics and Gynecology **23**(1), 80–92 (2004)
15. 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)
16. 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)
17. Zhou, M., Yuan, C., Chen, Z., Wang, C., Lu, Y.: Automatic angle of progress measurement of intrapartum transperineal ultrasound image with deep learning. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 406–414. Springer International Publishing, Cham (2020)
18. Lu, Y., Zhi, D., Zhou, M., Lai, F., Chen, G., Ou, Z., Zeng, R., Long, S., Qiu, R., Zhou, M., Jiang, X., Wang, H., Bai, J.: Multitask deep neural network for the fully automatic measurement of the angle of progression. Computational and Mathematical Methods in Medicine **2022**(1), 5192338 (2022)
19. 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)
20. Thaler, F., Payer, C., Urschler, M., Stern, D.: Modeling annotation uncertainty with gaussian heatmaps in landmark localization. arXiv preprint arXiv:2109.09533 (2021)
21. Marmolin, H.: Subjective mse measures. IEEE transactions on systems, man, and cybernetics **16**(3), 486–489 (1986)
22. Ma, J., Zhang, Y., Gu, S.: Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis **82**, 102616 (2022)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)