SSL-FetalBioNet: Self-Supervised Learning for Automated Angle of Progression Measurement in Intrapartum Ultrasound

Wang Kun[0009−0002−6896−3527], Li Lifei[0009−0007−1093−5167], Ma Yuzhang[0009−0003−8543−1407], Han Xiaoxin[0009−0002−6517−8531], Shao Haochen[0009−0009−1368−7423], and Wang Kun✉[0009−0002−6896−3527]

Gansu University of Chinese medicine, Lanzhou Gansu 730000, China
995956009@qq.com

Abstract. During childbirth, real-time assessment of fetal head position and progression is crucial for ensuring the safety of both mother and infant. Detecting key anatomical landmarks in intrapartum ultrasound images and calculating the Angle of progression (AoP) have become critical techniques in the next-generation childbirth monitoring protocol proposed by the World Health Organization (WHO). However, traditional manual analysis is time-consuming and prone to subjective bias, highlighting the urgent need for automated methods to achieve standardized and precise childbirth assessment. This paper presents a key point detection approach combining self-supervised pre-training with a U-Net architecture: first, the encoder is pre-trained using large-scale unlabeled images through self-supervision to uncover latent structural information; subsequently, this pre-trained encoder is transferred to the supervised learning stage to achieve precise localization of three key points (PS1, PS2, FH1). Our method achieved eighth place in the Intrapartum Ultrasound Grand Challenge 2025, demonstrating its effectiveness and generalization capability in the task of key point detection in intrapartum ultrasound. This work provides a practical and feasible pathway toward automated and scalable childbirth monitoring, with significant implications for global maternal and infant health.

Keywords: Self-supervised Learning · U-Net · Intrapartum Ultrasound · Angle of Progress.

1 Introduction

The dynamic nature of labor necessitates continuous monitoring of maternal and fetal health status in clinical practice. To standardize intrapartum monitoring and promote woman-centered childbirth experiences, the World Health Organization (WHO) [10]introduced the Labour Care Guide (LCG) in 2020, emphasizing the need for standardized measurement of key delivery parameters. The degree of fetal descent and rotation during delivery serves as crucial indicators for assessing labor progress. The Angle of Progression (AoP), as a

core parameter reflecting this process, is increasingly becoming a critical basis for clinical decisions regarding intervention timing and methods.AoP calculation relies on accurate identification of three key anatomical landmarks in intrapartum ultrasound images: the two most superior points of the pubic symphysis (PS1 and PS2) and the tangent point contacting the fetal head (FH1). However, current clinical practice predominantly depends on experienced sonographers for manual annotation of these landmarks, which is not only time-consuming and subjective but also susceptible to intra-/inter-observer variability, consequently compromising diagnostic consistency and reproducibility[8]. Therefore, developing an efficient, accurate, and automated landmark detection method is of significant importance for advancing intelligent labor assessment.In medical image processing, deep learning technologies—particularly convolutional neural network (CNN)-based models—have been widely applied to tasks such as image segmentation[1], object detection, and keypoint localization. Fully convolutional architectures like U-Net have demonstrated exceptional performance in medical image segmentation[11]. Nevertheless, acquiring annotated data remains challenging in practical applications, especially for high-quality medical imaging data, where annotation expertise and cost constraints hinder further development of supervised learning.

Self-Supervised Learning (SSL) has emerged as a vital approach to address the shortage of medical imaging data by learning useful image representations through designed pretext tasks without requiring human-generated labels. In the Intrapartum Ultrasound Grand Challenge (IUGC) 2025, the organizers provided a well-structured and comprehensive dataset comprising 300 labeled cases[2],31,421 unlabeled cases, and 2,045 reference standard plane images. The task required automatic localization of three key landmarks and precise calculation of AoP based on transperineal ultrasound images. To encourage exploration of model generalization capabilities, the use of additional pre-trained models was permitted. Two core evaluation metrics were adopted to assess algorithm performance: Mean Radial Error (MRE) for evaluating landmark localization accuracy and Absolute Parameter Difference (APD) for measuring AoP calculation accuracy.To address these challenges, this study designed a keypoint detection model integrating a self-supervised encoder with a U-Net architecture[7]. Specifically, extensive unlabeled images were utilized for self-supervised pre-training to enhance the encoder's perception of structural features. Subsequently, fine tuning was performed on labeled images to train the network to generate precise heatmaps for locating the three key points. The maximum activation positions in the heatmaps were accurately mapped to original image coordinates through normalization and interpolation, enabling automated AoP calculation.This method demonstrated outstanding performance on the IUGC challenge test set, achieving ninth place among global participants, validating its stability and adaptability on real clinical images. These results not only showcase the application potential of self-supervised strategies in medical image keypoint detection tasks but also establish a technical foundation for promoting the clinical translation of intelligent obstetric ultrasound analysis tools.

2 Method

2.1 Method Design

The adopted network model is based on the classic U-Net architecture[11], comprising symmetrical encoder and decoder modules. The encoder consists of four convolutional blocks, each containing two consecutive convolutional layers equipped with batch normalization and ReLU activation functions, all using 3×3 convolutional kernels[6]. The spatial dimensions of the feature maps are progressively reduced through max-pooling layers while increasing the number of feature channels, enabling multi-scale feature extraction[5]. A bottleneck layer is incorporated at the deepest part of the network to further extract high-level semantic information.The decoder section employs transposed convolution (ConvTranspose2d) for upsampling, combined with feature maps from the corresponding encoder layers to achieve feature fusion and spatial resolution recovery. The final output layer uses a 1×1 convolution to map to the number of heatmap channels corresponding to the keypoints, with outputs normalized through a Sigmoid activation function[13].The model accepts three-channel color images as input and generates two-dimensional heatmaps for each keypoint as output, with the heatmap size fixed at 64×64. This network design maintains consistency with the self-supervised pre-trained encoder architecture to facilitate loading of pre-trained weights, thereby accelerating training and enhancing model performance. The model supports loading pre-trained encoder weights, importing only layer parameters with matching key names and compatible shapes to ensure parameter compatibility and initialization quality.

2.2 Model architecture

A MoCo v2 framework was adopted to perform contrastive learning on large-scale unlabeled fetal ultrasound sequences. The encoder consisted of the customized U-Net encoder followed by a two-layer fully connected projection head. Input images underwent diverse augmentations, including random cropping and scaling, color jitter, grayscale conversion, Gaussian blur, and horizontal flipping. Positive pairs were generated from different augmented views of the same image, while negative samples were maintained in a feature queue updated by a momentum encoder. Training used a cross-entropy loss with the AdamW optimizer (learning rate = 1e-3, weight decay = 1e-3), a batch size of 64, and 200 epochs.

High-quality "gold standard" sequences[4] were identified by computing cosine similarity between candidate frames and a reference library using a ResNet50 feature extractor[9], with a threshold of 0.99. These sequences were then used to construct temporally adjacent frame pairs for further contrastive training under the MoCo v2 framework. Both query and key encoders were initialized from Stage 1, with the key encoder updated by momentum. A fixed-length negative sample queue of 16,384 was maintained. Training again used cross-entropy loss with AdamW (learning rate = 1e-4, weight decay = 1e-4, batch size = 64),

along with a ReduceLROnPlateau scheduler. This process produced the fully fine-tuned encoder weights.

The fine-tuned encoder from Stage 2 was integrated into a U-Net backbone for supervised keypoint localization. Input images were resized to 256×256, and the network produced 64×64 keypoint heatmaps. Training employed MSE loss with the Adam optimizer (learning rate = 1e-5, batch size = 8) for up to 1000 epochs, incorporating early stopping and dynamic learning rate adjustment. A custom collate function was designed to filter out invalid samples. Experimental results demonstrated that two-stage self-supervised pre-training substantially enhanced feature representation and training stability, achieving lower keypoint localization[3] error and AoP prediction error compared to models without pre-training[**?**]. Ablation studies further validated the positive contribution of high-quality sequence screening and the phased self-supervised learning strategy to final model performance.

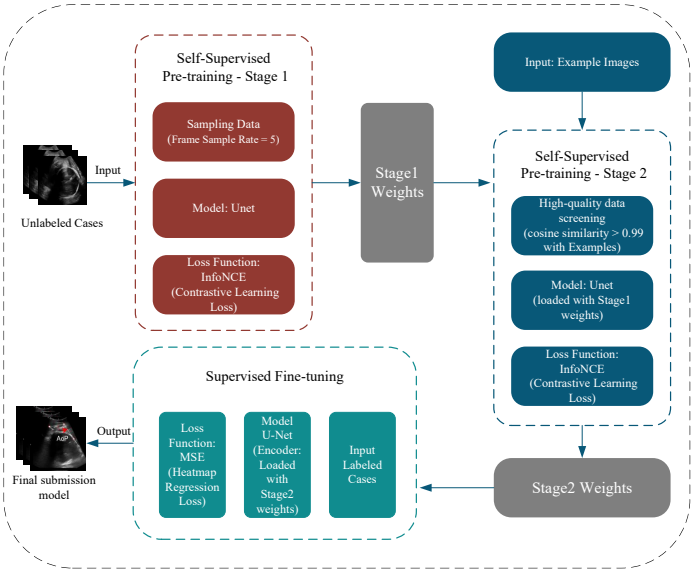

Fig. 1. This figure outlines the three-stage training pipeline of the proposed SSL-FetalBioNet model, which consists of two stages of self-supervised pre-training (MoCo v2) on unlabeled data to learn general feature representations, followed by a final supervised heatmap regression fine-tuning stage on labeled data for keypoint detection.

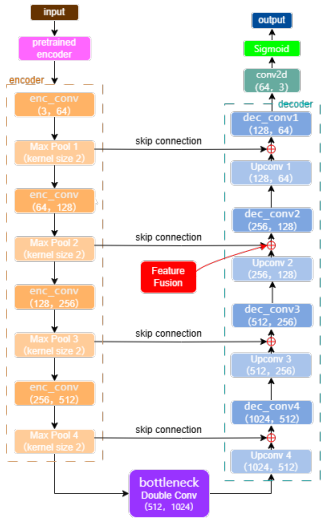

Fig. 2. This figure illustrates the core algorithmic architecture employed in our study. The model is based on the classic U-Net structure, where the left-side encoder pathway utilizes pre-trained weights for multi-scale feature extraction, while the right-side decoder pathway progressively restores spatial details through upsampling and feature fusion.

3 Experiments

3.1 Evaluation Metrics

This experiment selected four evaluation metrics to comprehensively assess the performance of the model in keypoint detection and AoP prediction, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Average Point Distance, and the Mean Absolute Error for AoP (AOP_MAE). The specific definitions are as follows:

Mean Squared Error (MSE) is used to measure the squared difference between pixel values of the predicted heatmap and the ground truth heatmap. The calculation formula is:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

where y_i and $\hat{y}_i$ represent the true value and predicted value of the i-th pixel, respectively, and N is the total number of pixels.

Mean Absolute Error (MAE) measures the absolute difference between predicted values and true values. The calculation formula is:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2}$$

Average Point Distance is used to evaluate the accuracy of keypoint localization. It is defined as the average Euclidean distance between predicted keypoint coordinates and ground truth coordinates. The calculation formula is:

$$AveragePointDistance = \frac{1}{K}\sum_{k=1}^{K}\sqrt{(x_k - \hat{x}_k)^2 + (y_k - \hat{y}_k)^2} \tag{3}$$

where K is the number of keypoints, and (x_k, y_k) and $(\hat{x}_k, \hat{y}_k)$ are the true and predicted coordinates of the k-th keypoint, respectively.

The Mean Absolute Error for AoP (AOP_MAE) quantifies the error in predicting the key angle parameter. The formula is:

$$\text{AOP\_MAE} = \frac{1}{M}\sum_{m=1}^{M}|a_m - \hat{a}_m| \tag{4}$$

where M is the number of predicted angle parameters, and a_m and $\hat{a}_m$ are the true and predicted values of the m-th angle, respectively.

3.2 Loss Function Analysis

In this task, the keypoint detection framework is based on a **heatmap regression** approach, which can be formalized as follows:

- **Target Heatmap Generation**: The ground-truth heatmap for each keypoint is constructed by centering a two-dimensional isotropic Gaussian distribution at the annotated coordinate location:

$$H_k(x, y) = \exp\left(-\frac{(x - x_k)^2 + (y - y_k)^2}{2\sigma^2}\right), \tag{1}$$

  where (x_k, y_k) represents the spatial coordinate of the k-th keypoint on the heatmap, and the standard deviation σ is set to 2.
- **Model Output**: The U-Net architecture predicts a corresponding heatmap $\hat{H}_k(x, y)$ for each keypoint. Each predicted heatmap is normalized to the range $[0, 1]$ via a sigmoid activation function.

– **Loss Function**: The discrepancy between the predicted heatmaps $\hat{H}_k$ and the target heatmaps H_k is quantified using the Mean Squared Error (MSE) loss, computed over all keypoints K and all spatial positions (H, W) within the heatmaps:

$$\mathcal{L}_{MSE} = \frac{1}{K \cdot H \cdot W} \sum_{k=1}^{K} \sum_{y=1}^{H} \sum_{x=1}^{W} \left(\hat{H}_k(x, y) - H_k(x, y) \right)^2. \tag{2}$$

Here, K denotes the total number of keypoints, and $H \times W$ specifies the spatial dimensions of the heatmap (64×64 in this implementation).

This loss function enforces pixel-wise consistency between the predicted and target heatmaps, guiding the network to learn the underlying spatial probability distribution for each keypoint. The final keypoint coordinates are subsequently deduced by identifying the pixel locations associated with the maximum values (peaks) in the predicted heatmaps $\hat{H}_k(x, y)$.

3.3 Data processing and experimental environment

In this study, the dataset provided by the IUGC Challenge was used, containing delivery ultrasound images with annotated keypoints and AoP parameters. To prepare the data for training, several preprocessing steps were applied. First, all images were uniformly resized from their original resolution to 256×256, ensuring consistency with the U-Net input requirements. The corresponding keypoint labels were then mapped to this resized coordinate system, and target heatmaps of 64×64 were generated using a Gaussian kernel to provide smooth supervision signals. In addition, pixel intensities were normalized to stabilize network training and improve convergence. The dataset was divided into training and validation subsets in an 8:2 ratio, with a batch size of 8. To handle potential issues such as missing images or corrupted labels, a custom filtering mechanism was integrated into the data loader to exclude invalid samples dynamically. These preprocessing operations—resizing, normalization, label transformation, heatmap generation, and data integrity checks—together established a robust and standardized input pipeline, providing a reliable foundation for supervised fine-tuning[12] and ensuring that the pre-trained encoder could be effectively leveraged for accurate keypoint localization and AoP prediction.

3.4 Experimental Results

The experimental results indicate that when using ResNet as the baseline, all error metrics were relatively high. After introducing the U-Net architecture, which leverages skip connections to restore spatial details, metrics such as MSE and APD showed noticeable improvement. By further applying the complete training pipeline implemented in this study—including Gaussian heatmap supervision, MSE loss, dynamic learning rate scheduling, early stopping, and invalid sample

Table 1. Experimental environment configuration.

System	Windows 11
CPU	Intel(R) Core(TM) i7-14650HX (2.20 GHz)
RAM	16 × 4GB/s
GPU (number and type)	NVIDIA GeForce RTX 4060 16G
CUDA version	11.7
Programming language	Python 3.9
Deep learning framework	PyTorch (Torch 2.0.1)

filtering—significant performance gains were achieved even without loading self-supervised weights, reducing MSE to 343.8 and APD to 19.65. Building on this, the incorporation of two-stage self-supervised pre-training (combining MoCo v2 with golden sequence selection) further enhanced geometric localization performance, bringing APD and AOPMAE down to 18.90 and 6.97, respectively. This demonstrates that self-supervised learning effectively strengthens feature representation and spatial relationship modeling. Overall, the trend in results is highly consistent with the code implementation: architectural improvements lead to foundational gains, a stable training pipeline substantially reduces errors, and self-supervised pre-training further refines the accuracy of both keypoint localization and angle prediction.

Table 2. Performance comparison of different methods.

Model	**MSE**	**MAE**	**APD**	**AOP_MAE**
Resnet	626.9202	17.9708	28.7208	8.1679
Unet	571.4269	17.1102	27.6144	9.7752
U-Net w/o SSL	343.7685	**11.6983**	19.6536	7.5685
Ours (SSL-FetalBioNet)	**313.5583**	12.1833	**18.8999**	**6.9679**

4 Conclusion

In this keypoint detection task, our SSL-FetalBioNet model delivered strong performance on the validation set, achieving perfect detection (Missing Rate: 0.0000) across all 100 samples. The model attained a mean absolute error of 12.06 pixels and an average point distance of 18.73 pixels in localization tasks, while angular prediction achieved a mean absolute error of 7.00 degrees. Detection accuracy varied across anatomical structures, with errors of 11.37 pixels for PS1, 15.95 pixels for PS2, and 28.86 pixels for the tangency point. With an average inference time of 24.34 milliseconds per image, the proposed self-supervised U-Net framework demonstrates both efficiency and reliability in fetal ultrasound keypoint detection, showing particular strength in angle estimation and offering promising support for clinical biometric applications.

References

1. Aghasizade, M., Kiyoumarsioskouei, A., Hashemi, S., Torabinia, M., Caprio, A., Rashid, M., Xiang, Y., Rangwala, H., Ma, T., Lee, B., Wang, A., Sabuncu, M., Wong, S.C., Mosadegh, B.: A coordinate-regression-based deep learning model for catheter detection during structural heart interventions. Applied Sciences **13**(13) (2023). https://doi.org/10.3390/app13137778, https://www.mdpi.com/2076-3417/13/13/7778 2
2. Bai, J., K.I.L.Y.N.D.Y.M.L.K.M.J..L.S.: Landmark detection challenge for intrapartum ultrasound measurement meeting the actual clinical assessment of labor progress. Zenodo (2025). https://doi.org/10.5281/zenodo.15172238 2
3. Cano-Espinosa, C., González, G., Washko, G.R., Cazorla, M., Estépar, R.S.J.: Biomarker localization from deep learning regression networks. IEEE Transactions on Medical Imaging **39**(6), 2121–2132 (2020). https://doi.org/10.1109/TMI.2020.2965486 4
4. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 3
5. Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R.: Deep learning-enabled medical computer vision. NPJ digital medicine **4**(1), 5 (2021) 3
6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022) 3
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 2
8. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017) 2
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 3
10. Organization, W.H., et al.: WHO recommendations on maternal and newborn care for a positive postnatal experience. World Health Organization (2022) 1
11. 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) 2, 3
12. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical image analysis **63**, 101693 (2020) 7
13. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018) 3