A Two-Stage Semi-Supervised Ensemble Framework for Automated Angle of Progression Measurement in Intrapartum Ultrasound

Bo Deng[1], Yu Chen[1], and Zilun Peng[1]

College of Information Science and Technology, Jinan University, Guangzhou 510632, China

Abstract. Accurate and reproducible measurement of the Angle of Progression (AoP) from intrapartum ultrasound is critical for modern labor management, yet manual annotation is hindered by significant intra- and inter-observer variability and workflow inefficiencies. To address this, we propose a fully automated, two-stage deep learning pipeline for precise landmark localization. The first stage employs a multi-model ensemble of U-Net architectures with diverse backbones (EfficientNet-B4 and -B7), trained under a Mean Teacher semi-supervised framework to leverage both labeled and unlabeled data. This stage generates a robust coarse prediction by performing a per-keypoint weighted average of the fused heatmaps. In the second stage, a dedicated Res-Net-18-based regression model refines the position of each landmark by predicting a precise offset from its coarse location on a localized image patch. Our integrated approach, trained on a combined dataset from the 2024 and 2025 IUGC challenges, demonstrates highly competitive performance, achieving a Mean Radial Error (MRE) of 12.7888 pixels and a mean Absolute Parameter Difference (APD) of 4.4581 degrees for the AoP on the test set. This automated framework promises to enhance diagnostic consistency and streamline clinical workflows, aligning with the WHO's vision for improved intrapartum care.

Keywords: Intrapartum Ultrasound · Angle of Progression (AoP) · Two-Stage Model · Semi-Supervised Learning.

1 Introduction

Effective intrapartum care, crucial for maternal and fetal well-being, relies on accurate labor monitoring. The World Health Organization (WHO) has recently advanced this effort with its Labour Care Guide (LCG) [10], which promotes standardized, evidence-based assessment. Within this framework, intrapartum ultrasound is an indispensable tool for evaluating fetal head progression, a recommendation strongly supported by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) [4]. A cornerstone of modern ultrasound-based labor assessment is the measurement of the Angle of Progression (AoP), a key biometric calculated from three anatomical landmarks: two points on the

pubic symphysis (PS1, PS2) and a point on the fetal head (FH1). The AoP provides critical, quantitative insight into the fetal head's descent through the birth canal, directly informing clinical decisions regarding the mode of delivery and the timing of interventions.

Despite its clinical utility, the manual annotation required to measure AoP presents a significant bottleneck in busy clinical settings. This process is not only time-consuming but also highly dependent on the operator's experience, suffering from substantial intra- and inter-observer variability that compromises its reliability and reproducibility [8]. This challenge underscores an urgent need for an automated, standardized solution that can provide objective and consistent AoP measurements.

To achieve such automation, deep learning (DL) has become the mainstream approach for landmark localization in medical imaging. While initial DL methods based on direct heatmap regression have shown promise [9,12], they often suffer from sensitivity to image quality variations and struggle with large coordinate ranges. In response, more sophisticated coarse-to-fine and cascaded regression strategies were developed to decompose the localization task into more manageable steps, thereby improving precision [3]. However, a key limitation persists across these advanced methods: they are typically supervised and thus fail to leverage the vast amounts of unlabeled data common in the medical domain. The framework proposed in this paper is designed specifically to address this limitation. The key contributions are as follows.

1. We introduce a robust two-stage, coarse-to-fine pipeline that synergistically combines the global context awareness of a first-stage model with the high-precision local analysis of specialized second-stage models.
2. We are the first to apply a multi-model, semi-supervised ensemble for the coarse localization stage. By integrating two U-Net models with diverse EfficientNet backbones [15] and training them within a Mean Teacher framework [16], our method effectively utilizes both labeled and thousands of unlabeled images to enhance generalization and robustness.
3. We demonstrate the efficacy of our complete pipeline through extensive experiments on a large-scale, combined dataset from the 2024 and 2025 IUGC challenges, achieving state-of-the-art performance.
4. The proposed automated framework offers a practical and powerful solution for standardizing AoP measurement, presenting a significant step towards the technical implementation of the WHO's LCG and the broader biomedical objective of safer intrapartum care.

The remainder of this paper is organized as follows. Section 2 reviews related work in landmark localization and semi-supervised learning. Section 3 details our proposed two-stage methodology. Section 4 presents our experimental setup and results, including comprehensive ablation studies. Finally, Section 5 concludes the paper and discusses future work.

2 Related Work

This section reviews the key areas of research that form the foundation of our work: deep learning-based landmark localization, semi-supervised learning in medical imaging, and advanced strategies for improving localization accuracy.

2.1 Deep Learning for Landmark Localization

Automated landmark localization is a fundamental task in medical image analysis. Traditional machine learning methods have largely been superseded by deep learning approaches, which have demonstrated superior performance. The dominant paradigm for this task is heatmap regression [9]. In this approach, instead of directly regressing the coordinates of a landmark, the network is trained to predict a 2D Gaussian-like heatmap for each keypoint, where the peak of the heatmap corresponds to the landmark's location. This method provides richer supervision and has been shown to be more robust to initialization and optimization challenges than direct coordinate regression. Seminal works, such as the Stacked Hourglass network [9] for human pose estimation, established the efficacy of this approach. Subsequently, architectures like U-Net [12], originally designed for segmentation, have been widely adapted for heatmap regression in medical imaging due to their powerful encoder-decoder structure and effective use of skip connections to preserve spatial details. While U-Net [12] remains a strong baseline, recent architectures such as TransUNet [1] have started incorporating Transformers to better capture long-range dependencies. To bridge the gap between heatmaps and coordinates, methods like Integral Pose Regression [14] have also proposed differentiable operations for end-to-end training. Nevertheless, these methods are primarily supervised and their performance is tied to the quantity of annotated data.

2.2 Semi-Supervised Learning in Medical Imaging

The acquisition of large, expertly annotated medical datasets is a significant bottleneck. Semi-supervised learning (SSL) offers a compelling solution by enabling models to learn from a small set of labeled data alongside a much larger set of unlabeled data. Consistency regularization has emerged as a leading SSL strategy. The core idea is that a model's prediction should remain stable (consistent) under different perturbations of its input or its own parameters. The Mean Teacher framework [16] is a state-of-the-art consistency-based method that has shown great success in medical imaging [17]. It maintains two models: a student model, which is trained via standard backpropagation, and a teacher model, whose weights are an exponential moving average (EMA) of the student's weights. The student is then encouraged to produce predictions consistent with those of the more stable teacher model on unlabeled data, typically by minimizing the Mean Squared Error (MSE) between their outputs. This EMA-based approach provides a more stable pseudo-labeling target than self-ensembling methods. Building upon this, recent approaches like FixMatch [13] have further

simplified the SSL framework. Moreover, the inherent robustness of consistency-based methods makes them well-suited for medical data, which often suffers from noisy labels [18], a challenge implicitly addressed by our Mean Teacher approach.

2.3 Advanced Strategies for High-Precision Localization

To push the performance boundaries of landmark localization, several advanced strategies have been proposed. One powerful paradigm is the coarse-to-fine, or two-stage, approach. This strategy decomposes the difficult task of global localization into two simpler steps: first, a coarse prediction to identify the general region of interest, and second, a refined prediction within that localized region. This cascaded approach, rooted in early works on pose regression [3], effectively manages large coordinate ranges and allows a specialized refinement model to focus on local details, leading to higher precision [15]. Another widely adopted technique for enhancing model robustness and accuracy is ensembling. By combining predictions from multiple diverse models, the variance of the individual models' errors can be reduced. A common and effective ensemble strategy involves training models with the same architecture but different backbones [2], such as EfficientNet-B4 and -B7 [15]. Fusing their predictions, for instance by averaging their output heatmaps, often yields a more reliable result than any single model could achieve alone. Our work integrates both of these advanced strategies into a unified, semi-supervised framework.

3 Methodology

To achieve accurate and fully automated measurement of the Angle of Progression (AoP), we propose a novel two-stage, semi-supervised ensemble framework, illustrated in Fig. 1. Our pipeline is divided into two main stages: Semi-Supervised Ensemble Coarse Localization and Local Offset Refinement. In the first stage, an ensemble of models processes the full ultrasound image to generate a Fused Heatmap, from which initial Coarse Coordinates (P) are extracted. Subsequently, in the second stage, specialized refinement networks analyze local image patches (C) centered at these coarse predictions to regress precise Coordinate Offsets (δ). These offsets are then added to the coarse coordinates to produce the final Refined Coordinates (R). The following subsections detail each component of this pipeline.

3.1 Semi-Supervised Ensemble Coarse Localization

Network Architecture. Our coarse localization models are built upon the U-Net architecture [12], a fully convolutional network renowned for its efficacy in biomedical image analysis. The U-Net consists of a contracting path (encoder) to capture context and a symmetric expanding path (decoder) to enable precise localization. To further enhance the feature extraction capabilities of our models, we employ powerful backbones from the EfficientNet family [15], which

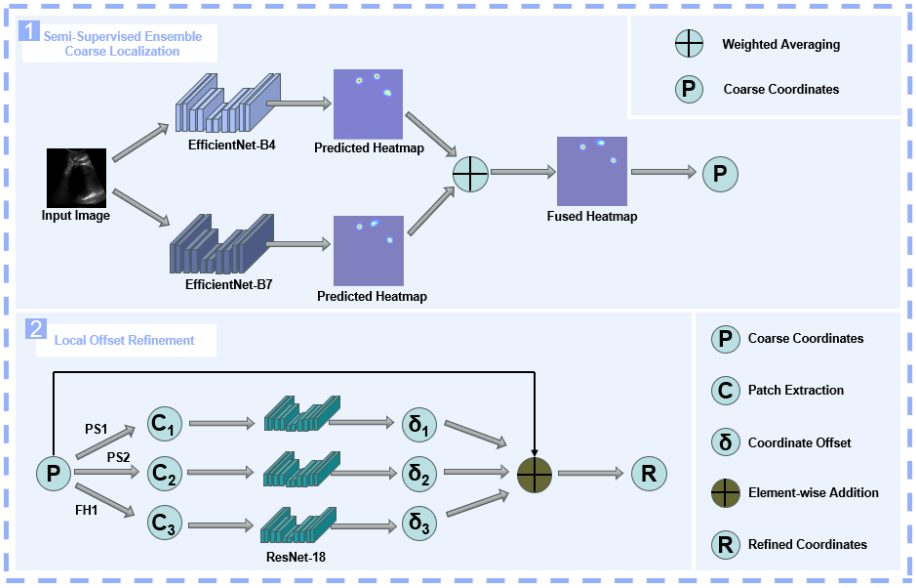

Fig. 1. Overview of our proposed two-stage framework for automated landmark localization.

are pre-trained on the ImageNet dataset. To foster model diversity, a critical component for effective ensembling, we trained two independent U-Net models utilizing EfficientNet-B4 and EfficientNet-B7 as their respective encoders. Each model outputs a 3-channel heatmap of size 128×128, where each channel corresponds to one of the three target landmarks (PS1, PS2, FH1).

Semi-Supervised Training with Mean Teacher. To effectively utilize the large volume of available unlabeled data, each U-Net model was trained within the Mean Teacher semi-supervised framework [16]. This framework consists of two identical models: a student model and a teacher model. The student model, with weights θ_t at training step t, is updated using standard backpropagation. The teacher model, with weights θ'_t, is not trained directly via backpropagation; instead, its weights are updated as an Exponential Moving Average (EMA) of the student's weights:

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t,$$

where α is a smoothing coefficient, or EMA decay rate. This EMA update makes the teacher model a more stable and reliable ensemble of the student's own past states.

The total loss function L for the student model is a combination of a supervised loss L_{sup} on labeled data and an unsupervised consistency loss L_{unsup} on unlabeled data:

$$L = L_{\text{sup}} + w(t) \cdot L_{\text{unsup}}.$$

where $w(t)$ is a time-dependent weighting factor that balances the two losses.

The supervised loss L_{sup} is a Weighted MSE applied to the labeled batch, designed to focus the model on the keypoint peaks. It is defined as:

$$L_{\text{sup}} = \frac{1}{N} \sum_{i=1}^{N} (H_{gl}^{(i)})^{\gamma} \cdot (H_{pred}^{(i)} - H_{gl}^{(i)})^2.$$

where H_{pred} and H_{gl} are the predicted and ground-truth heatmaps, respectively, and γ is a focusing parameter (set to 2 in our experiments).

The consistency L_{unsup} enforces that the student model's prediction on a strongly-augmented unlabeled image, denoted as $f(x'_{u,j}, \theta_t)$, should be consistent with the more stable teacher model's prediction on a weakly-augmented version of the same image, $f(x_{u,j}, \theta'_t)$. The loss is calculated as the MSE between the two outputs, but only for pseudo-labels where the teacher's confidence is high (i.e., the maximum value of the teacher's heatmap exceeds a threshold τ):

$$L_{\text{unsup}} = \frac{1}{M} \sum_{j=1}^{M} M_j \cdot \|f(x'_{u,j}, \theta_t) - f(x_{u,j}, \theta'_t)\|_2^2.$$

where M_j is a mask that is 1 if $\max(\mathcal{T}(f(x_{u,j}, \theta'_t))) \geq \tau$ and 0 otherwise.

The consistency weight $w(t)$ is gradually increased during training using a sigmoid ramp-up function to allow the model to learn from labeled data first before introducing the unsupervised signal.

Ensemble Strategy for Inference. During inference, we leverage the diversity of the two trained models to produce a single, highly robust prediction. For a given input image I, we obtain the predicted heatmaps H_{B4} from the EfficientNet-B4 model and H_{B7} from the EfficientNet-B7 model. The final fused heatmap, H_{fused}, is generated by a per-keypoint weighted average of these two heatmaps:

$$H_{\text{fused},k} = w_{B4,k} \cdot H_{B4,k} + w_{B7,k} \cdot H_{B7,k}.$$

where $k \in \{PS1, PS2, FH1\}$ denotes the keypoint channel, and the weights $w_{B4,k}$ and $w_{B7,k}$ are hyperparameters determined based on the individual performance of each model on the validation set for that specific keypoint. This strategy allows us to capitalize on the strengths of each model for each landmark, resulting in a superior coarse localization.

3.2 Local Offset Refinement

While the first stage provides a robust global localization, its output resolution may not be sufficient for achieving the highest possible precision, which is critical for accurate AoP calculation. To address this, we introduce a second refinement stage that operates on high-resolution local patches, a strategy proven effective in high-precision localization tasks [3,15]. This coarse-to-fine approach allows a specialized model to focus on fine-grained local details without being distracted by the complexity of the entire image.

Patch Extraction. For each of the three landmarks (PS1, PS2, and FH1), we use its coarse coordinate (x_c, y_c), predicted by the ensemble model in Stage 1, as a center point. A high-resolution image patch is then cropped from the original, full-resolution ultrasound image, centered at (x_c, y_c). To handle cases where the coarse prediction is near the image border, we employ zero-padding to ensure that all extracted patches have a consistent, predefined size. Based on our experiments, a patch size of 128×128 pixels was found to provide a robust balance between local detail and sufficient context for all three landmarks.

Refinement Network Architecture. Our refinement network is designed to be lightweight yet powerful enough for the local regression task. We employ a ResNet-18 architecture [5], pre-trained on ImageNet, as the feature extractor. We removed the final average pooling and fully-connected classification layers from the standard ResNet-18. In their place, we appended a custom Multi-Layer Perceptron (MLP) regression head. This head consists of a global average pooling layer, a fully-connected layer with 256 neurons and ReLU activation, a Dropout layer with a rate of 0.5 for regularization, and a final fully-connected layer that outputs a 2-dimensional vector representing the predicted offset.

Learning Objective and Loss Function. The objective of each refinement network is to learn a mapping from an input image patch P to a precise coordinate offset vector $(\Delta x, \Delta y)$. This offset represents the displacement from the coarse prediction (x_c, y_c) to the ground-truth landmark position $(x_g t, y_g t)$. To make the learning target independent of the patch size, the ground-truth offset is normalized by the patch dimension S_{patch}:

$$\text{label} = \left(\frac{x_{\text{gt}} - x_c}{S_{\text{patch}}}, \frac{y_{\text{gt}} - y_c}{S_{\text{patch}}} \right).$$

The network is trained to minimize the MSE between its predicted normalized offset and the ground-truth label. This loss function effectively penalizes deviations in the predicted offset, driving the model to learn a highly accurate local correction. We trained a separate, specialized refinement model for each of the three landmarks, allowing each model to learn the specific local features associated with its target.

4 Experiments and Results

To validate the efficacy of our proposed framework, we conducted a series of comprehensive experiments. This section is structured as follows: First, we describe the datasets, evaluation metrics, and our implementation details. Second, we present a thorough ablation study to dissect the individual contribution of each component within our pipeline—namely, semi-supervised learning, model ensembling, and the two-stage refinement. Finally, we report the performance of our complete, optimized model on the official IUGC 2025 test set and compare it against the provided baseline to demonstrate its state-of-the-art capabilities.

4.1 Dataset and Evaluation Metrics

Datasets. The datasets used in our experiments were constructed from the official data of the 2024 and 2025 IUGC challenges. The labeled dataset for our study was formed by combining the labeled sets from both challenges, resulting in a total of 2875 images with corresponding ground-truth coordinates for the three landmarks (PS1, PS2, and FH1). From this combined labeled set, we performed a fixed, stratified split, allocating 2500 images for training and reserving the remaining 375 images as our validation set for hyperparameter tuning and model selection. Additionally, we utilized 4787 unlabeled images from the 2025 IUGC challenge for consistency regularization within our semi-supervised framework.

Evaluation Metrics. We use two primary metrics to assess model performance. The Mean Radial Error (MRE), also known as the mean point distance, calculates the average Euclidean distance in pixels between the predicted and ground-truth coordinates, providing a direct measure of localization accuracy. The APD measures the mean absolute error in degrees between the Angle of Progression (AoP) calculated from the predicted landmarks and that from the ground-truth landmarks. APD evaluates the clinical utility of the predictions by quantifying the accuracy of the derived geometric parameter.

4.2 Implementation Details

All models were implemented within the PyTorch framework and trained on NVIDIA A100 or RTX 4090 GPUs. All input images were preprocessed by resizing to 512×512, followed by Contrast Limited Adaptive Histogram Equalization (CLAHE) [11] to enhance local image contrast.

For the first-stage semi-supervised training, we utilized the AdamW optimizer [7], which decouples weight decay regularization from the adaptive learning rate update, often leading to better generalization. The initial learning rate was set to 1e-4. The teacher model's EMA decay rate, α, was set to 0.999. The maximum consistency weight, w_{max}, was set to 2.0 and was gradually increased over a ramp-up period of 40 epochs. The batch size was 8 for both labeled and unlabeled data.

For the second-stage refinement, we also employed the AdamW optimizer with an initial learning rate of 1e-4. The batch size was set to 64, and the patch size for all three landmarks was 128x128 pixels. For all training processes, we utilized a Cosine Annealing learning rate schedule [6] with a 10-epoch warm-up period to ensure smooth and stable convergence.

4.3 Quantitative Analysis

We now present the quantitative results of our experiments. First, we conduct a detailed ablation study to dissect the contribution of each component in our framework. Then, we compare the final performance of our full pipeline against the official challenge baseline on the test set.

Table 1. Ablation study of framework components on the IUGC 2025 validation set.

Method			Performance Metrics	
EfficientNet-B7	EfficientNet-B4	Refinement Stage	MRE (pixels) ↓	APD (degrees)↓
✓			9.8987	3.7632
	✓		8.5412	2.6852
✓	✓		8.2530	2.5839
✓	✓	✓	**7.9163**	**2.4559**

Table 2. Final performance comparison with the official baseline on the IUGC 2025 test set.

Method	MRE (pixels) ↓	APD (degrees)↓
Baseline	21.8273	8.3727
Ours	12.7888	4.4581

Ablation Studies. We performed a comprehensive ablation study on the IUGC 2025 validation set to validate each component of our framework. The results are summarized in Table 1.

Our analysis begins with the individual models. The U-Net with an EfficientNet-B4 backbone, trained under our full semi-supervised (SSL) framework, achieved a strong baseline performance with an MRE of 8.5412 pixels. In contrast, the larger EfficientNet-B7 model, trained only on supervised data, performed worse, as expected due to the smaller training set.

Intriguingly, ensembling these two diverse models yielded a result superior to either standalone model. The ensemble lowered the MRE to 8.2530 and the APD to 2.5839. This highlights a key benefit of ensembling heterogeneous models: the diversity in their training schemes (semi-supervised vs. supervised) and architectures created complementary error patterns. The supervised B7 model, though less accurate overall, acted as a regularizer, correcting specific failure modes of the more powerful but potentially biased SSL-trained B4 model.

Finally, the addition of our Refinement Stage provided the most significant performance gain, reducing the MRE to 7.9163 pixels and APD to 2.4559 degrees. This confirms that our full two-stage ensemble pipeline is highly effective, with each component providing a distinct and crucial contribution.

Comparison with Official Baseline. To provide a final, unbiased evaluation of our complete framework, we submitted our best-performing model—the full two-stage ensemble with refinement—to the official challenge evaluation server for assessment on the test set. We compare our final results against the official baseline provided by the IUGC 2025 challenge organizers. As shown in Table 2, our method achieves a dramatic improvement over the baseline across both key metrics. Our final model obtained an MRE of 12.7888 pixels and an APD of 4.4581 degrees on the test set. Compared to the official baseline's performance of 21.83 pixels MRE and 8.37 degrees APD, our framework achieved a

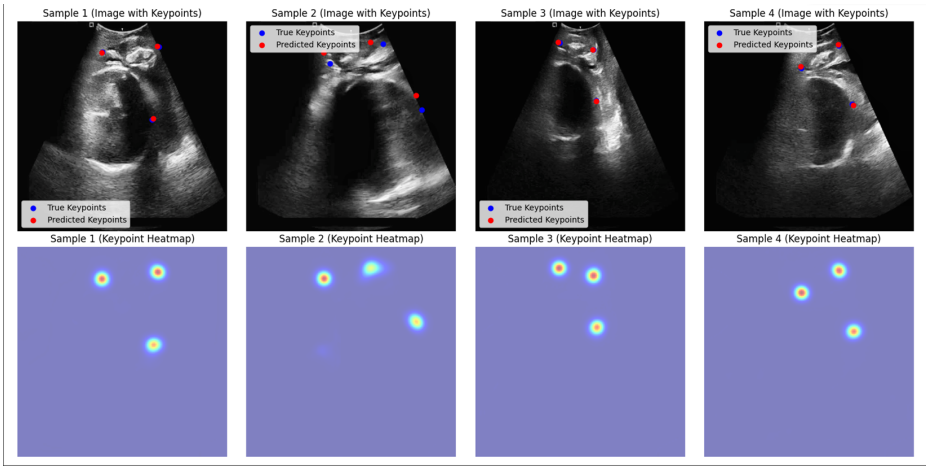

Fig. 2. Qualitative visualization of our model's predictions on four randomly selected samples from the training set.

remarkable 41.4% reduction in MRE and a 46.8% reduction in APD. This substantial improvement in a challenging, unseen dataset underscores the real-world efficacy and strong generalization capability of our integrated semi-supervised, ensemble, and two-stage refinement approach.

4.4 Qualitative Analysis

In addition to the quantitative metrics, we provide qualitative visualizations to offer further insight into our model's behavior. Fig. 2 displays the prediction results on four randomly selected samples from our combined training set. In most cases (Samples 1, 3, and 4), our full pipeline demonstrates excellent performance, accurately localizing all three landmarks with high precision.

Sample 2 illustrates a more challenging scenario. While the pubic symphysis landmarks (PS1 and PS2) are still accurately identified, the prediction for the Fetal Head (FH1) landmark shows a noticeable deviation from the ground truth. The corresponding heatmap for FH1 appears more diffuse and less confident compared to the other landmarks. This type of failure case typically occurs in images with low contrast or ambiguous anatomical features for the fetal head, highlighting a potential area for future improvement, such as incorporating more advanced context-aware mechanisms. Overall, the visualizations confirm the strong performance of our method on the majority of samples.

5 Conclusion

In this work, we proposed a novel two-stage, semi-supervised ensemble framework to address the critical challenge of automated Angle of Progression (AoP) measurement in intrapartum ultrasound. By synergistically combining a powerful

semi-supervised, multi-model ensemble for coarse localization with specialized, high-resolution refinement models, our pipeline robustly handles the complexities of clinical ultrasound data. Extensive ablation studies validated the significant contribution of each component. Our final model achieves state-of-the-art performance on the IUGC 2025 test set, reaching an Average Point Distance of 12.7888 pixels and an APD of 4.4581 degrees. This work presents a highly effective and practical solution for standardizing intrapartum assessment, representing a tangible step towards enhancing the quality and efficiency of modern labor care.

References

1. 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. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer (2000)
3. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 1078–1085. IEEE (2010)
4. 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)
5. 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)
6. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
8. Nassr, A.A., Hessami, K., Berghella, V., Bibbo, C., Shamshirsaz, A.A., Shirdel Abdolmaleki, A., Marsoosi, V., Clark, S.L., Belfort, M.A., Shamshirsaz, A.A.: Angle of progression measured using transperineal ultrasound for prediction of uncomplicated operative vaginal delivery: systematic review and meta-analysis. Ultrasound in Obstetrics & Gynecology **60**(3), 338–345 (2022). `https://doi.org/https://doi.org/10.1002/uog.24886`, `https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.24886`
9. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
10. Organization, W.H., et al.: Who labour care guide: user's manual. In: WHO labour care guide: user's manual (2020)
11. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
12. 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)

13. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
14. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European conference on computer vision (ECCV). pp. 529–545 (2018)
15. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
16. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
17. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine learning **109**(2), 373–440 (2020)
18. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-src: A contrastive approach for combating noisy labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5192–5201 (2021)