Noisy Student-Based Self-Training Enhances Landmark Detection in Intrapartum Ultrasound

Xiao Liu[1], Jiale Hu[1⋆], Yunda Li[1], Xufan Chen[1], and Yufeng Wang[2]

[1] School of Computer and Software, Nanyang Institute of Technology, Nanyang 473004, China
[2] Academy for Electronic Information Discipline Studies, Nanyang Institute of Technology, Nanyang 473004, China

Abstract. Accurate detection of fetal anatomical landmarks in ultrasound images during labor is crucial for clinical labor assessment. Despite significant progress in deep learning for medical image analysis, achieving high-precision and robust keypoint detection remains challenging under the realistic condition of scarce annotated data. Inspired by the Noisy Student paradigm in image classification, this paper proposes an improved semi-supervised method tailored for keypoint detection tasks.We construct a DenseUNet teacher-student framework to perform collaborative training using limited annotated data and a portion of unlabeled images. Specifically, the teacher model is trained on the labeled set to generate heatmap pseudo-labels for the unlabeled data; the student model, supervised by the pseudo-labels, leverages the dense connectivity of DenseNet to enhance feature reuse and gradient flow, and incorporates Dropout in the decoder to improve robustness. Furthermore, a linearly-decayed MixUp strategy is adopted for input perturbation, combined with heatmap supervision, to achieve a smooth transition from strong perturbation training to stable convergence. Experiments on the IUGC 2025 test set demonstrate that the proposed method significantly improves landmark detection performance, achieving an average distance error (Distance) of 13.1574 and an AOP_MAE score of 4.4244, which verifies the effectiveness of the method in scenarios with limited annotation resources. The project source code is available at: `https://github.com/apuomline/IUGC2025`.

Keywords: Anatomical landmark detection · Intrapartum ultrasound· Semi-supervised learning · Noisy Student framework.

1 Introduction

Accurate assessment of fetal head descent during labor is crucial for reducing the risks of dystocia and cesarean delivery. According to the World Health Organization's 2020 Guidelines on Intrapartum Care for a Positive Birth Experience, real-time monitoring of fetal head progression is one of the core intervention

⋆ Corresponding author

measures. Among these, the Angle of Progression (AoP) serves as a key indicator for quantifying this process. It is formed by two points at the distal end of the pubic symphysis and one point on the fetal head. However, the current measurement of AoP still relies on experienced clinicians to manually annotate each frame, which is time-consuming and subject to significant inter-observer variability.

To address the aforementioned limitations, the MICCAI 2025 Intrapartum Ultrasound Challenge (IUGC 2025) introduced[1] a semi-supervised end-to-end keypoint detection method for AoP estimation. Unlike the standard two-stage approach of segmentation[2] followed by measurement, this approach directly regresses the coordinates of the three anatomical landmarks from transperineal ultrasound images and outputs the AoP in real time, enabling zero-interaction and fully automatic measurement. This dataset covers nine descent stations of fetal head presentation ranging from -5 to +3 and includes 32,000 multicenter images, making it the largest and most comprehensive intrapartum ultrasound image repository to date. The 501 hidden test samples were independently annotated by multiple experts, ensuring clinical-grade reliability.

However, the IUGC 2025 dataset presents significant challenges in terms of fetal pose diversity, image quality variation, and label scarcity, limiting the effectiveness of traditional fully supervised methods in fully unlocking its clinical value. Therefore, there is an urgent need for a training strategy that can efficiently utilize the limited labeled samples along with the large amount of unlabeled images, in order to fully unlock the potential of this dataset in intrapartum assessment. Semi-Supervised Learning (SSL) provides a promising solution for such tasks, with major approaches including consistency regularization (e.g., FixMatch [3], MixMatch [4], ReMixMatch [5], UniMatch [6]) and pseudo-labeling strategies. Among these, pseudo-labeling methods such as Noisy Student [7] employ an iterative self-training paradigm, where a teacher model generates high-quality pseudo labels, which are then used to train a student model for further performance improvement. This approach introduces input noise (e.g., RandAugment [8]) and architectural noise (e.g., Stochastic Depth [9], Dropout [10]) into the student model, and gradually increases both model capacity and perturbation strength during training. It has achieved remarkable success in natural image classification tasks.

To address these challenges, we employ a DenseUNet model (as shown in Figure 2.) with heatmap regression and introduce the Noisy Student framework for semi-supervised training. The overall training pipeline is illustrated in Figure 1. Initially, a teacher model is trained exclusively on the labeled data by minimizing the mean squared error (MSE) loss. Once the teacher model has been trained, it is used to generate pseudo-labels for the unlabeled images. Subsequently, self-training is performed under both input and structural noise; this stage simultaneously receives labeled and unlabeled images as input. A student model that mirrors the architecture of the teacher model is optimized to minimize a joint mean squared error (MSE) loss computed on both the labeled data

and the pseudo-labeled data. After the student model is trained, it replaces the teacher model and the process proceeds to the next iteration.

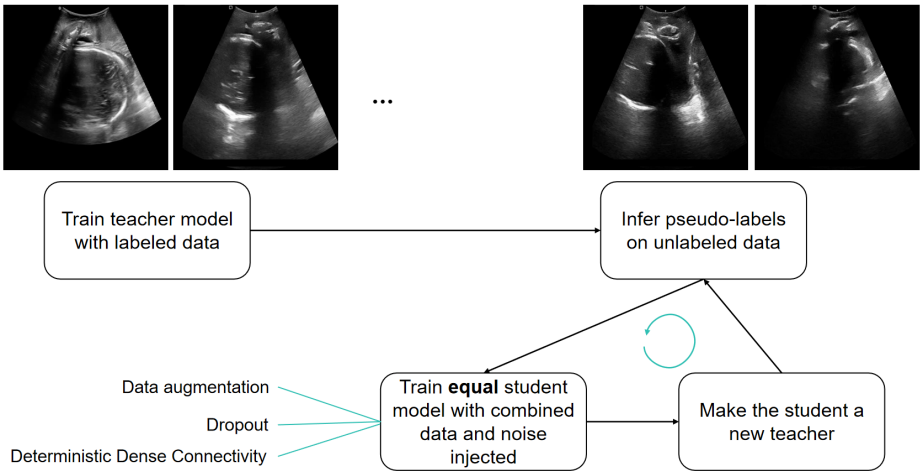

Fig. 1. The Noisy Student training pipeline used for landmark detection in intrapartum ultrasound.

To effectively enhance the generalization ability and training stability of the model in the fetal keypoint detection task, we introduce targeted improvements to the two noise mechanisms in the Noisy Student (NS) [7] framework. In terms of model noise design, we fully leverage the dense connectivity pattern of the Dense Block in DenseNet [11], whose multi-path structure inherently provides a regularization effect similar to Stochastic Depth [9]. Building upon this, we introduce Dropout [10] layers in the decoder as the primary regularization mechanism, enhancing model generalization while avoiding training instability that may arise from structural complexity. In terms of input noise design, we employ data augmentation methods tailored for medical ultrasound images and further propose a MixUp [12] augmentation strategy integrated with heatmap regression. The mixing weight coefficient in this strategy decays linearly throughout the training process, allowing the model to be exposed to stronger perturbations in the early training stages and gradually transition to a more stable phase, thereby enhancing model robustness.

Our main contributions can be summarized as follows:

1. We propose a semi-supervised learning framework based on the Noisy Student paradigm and successfully apply it to the task of fetal keypoint detection in labor ultrasound, providing a novel approach for medical image analysis in scenarios with limited annotations.

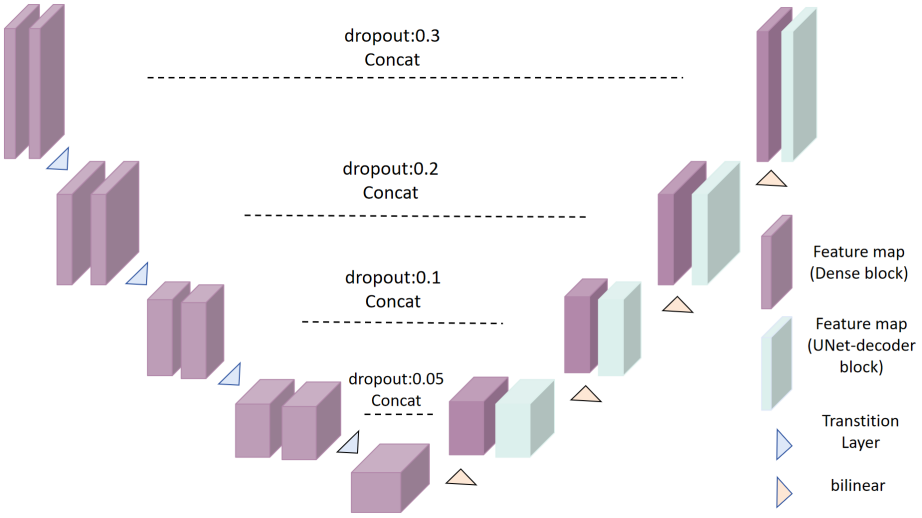

Fig. 2. Network architecture of DenseUNet.

2. We propose a collaborative noise injection mechanism that enhances the model's robustness to internal randomness by jointly leveraging the dense connectivity characteristics of DenseNet and Dropout in the decoder.
3. We introduce a dynamic MixUp augmentation strategy, whose perturbation strength linearly decays during training and is co-optimized with the heatmap regression objective, thereby enabling stable improvement in keypoint localization accuracy.

2 Methods

The overall framework of the proposed method is illustrated in Figure 1, following the Noisy Student paradigm for semi-supervised keypoint detection in intrapartum ultrasound images. The process involves iterative training of a teacher model and a student model.

2.1 Teacher Model Training and Pseudo-Label Generation

First, the teacher model is initialized and trained on the labeled dataset $\mathcal{D}_l$. A DenseUNet architecture is employed, and the model is optimized by minimizing the Mean Squared Error (MSE) loss between the predicted heatmap H_{pred} and the ground truth heatmap H_{gt}:

$$\mathcal{L}_{teacher} = \frac{1}{N_l} \sum_{i=1}^{N_l} \|H^i_{pred} - H^i_{gt}\|_2^2 \tag{1}$$

where N_l denotes the number of labeled samples. After the teacher model is fully trained, its weights are frozen and it is used to generate pseudo-labels $\hat{H}_{pseudo}$ for the unlabeled dataset $\mathcal{D}_u$. For each unlabeled image $x_u \in \mathcal{D}_u$, the corresponding pseudo-label is obtained via the teacher model's forward pass:

$$\hat{H}_{pseudo} = f_{teacher}(x_u) \tag{2}$$

where $f_{teacher}$ represents the inference process of the teacher model. The generated pseudo-labels are then used alongside the ground truth labels for the subsequent training of the student model.

2.2 Linearly-Decaying MixUp Augmentation for Heatmap Regression

The core idea of MixUp is to generate new training samples and corresponding soft labels by linearly interpolating between two randomly selected samples (images and their labels). In the context of keypoint detection, this is realized as follows. Let two randomly selected ultrasound frames be $\mathbf{I}_A, \mathbf{I}_B \in \mathbb{R}^{H\times W\times C}$, with corresponding keypoint annotations $\mathcal{K}_A = \{(x_i^A, y_i^A)\}_{i=1}^N$, $\mathcal{K}_B = \{(x_i^B, y_i^B)\}_{i=1}^N$, where N denotes the number of anatomical landmarks. A mixing coefficient is sampled from a Beta distribution $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$, with hyper-parameter $\alpha > 0$ controlling the interpolation strength. The mixed image and keypoint coordinates are then computed as

$$\mathbf{I}_{\mathrm{mix}} = \lambda\, \mathbf{I}_A + (1-\lambda)\, \mathbf{I}_B, \tag{3}$$

$$(x_i^{\mathrm{mix}}, y_i^{\mathrm{mix}}) = \lambda\, (x_i^A, y_i^A) + (1-\lambda)\, (x_i^B, y_i^B), \quad \forall i \in \{1, \ldots, N\}. \tag{4}$$

Consequently, the synthetic sample $(\mathbf{I}_{\mathrm{mix}}, \mathcal{K}_{\mathrm{mix}})$ exhibits smooth transitions in both appearance and geometry, effectively expanding the local neighborhood of the data distribution.

In our work, we adopt a heatmap regression-based keypoint detection approach, which is particularly well-suited for the soft-label formulation introduced by MixUp. Heatmaps inherently represent keypoint locations as dense, continuous 2D probability distributions. This characteristic aligns naturally with the interpolation mechanism of MixUp, enabling the mixed heatmap labels to be generated through the same linear interpolation process applied to the images and keypoint coordinates, thereby preserving spatial consistency. Moreover, heatmaps incorporate a certain degree of spatial uncertainty via a Gaussian kernel, which allows them to absorb and express the additional "blurring" effect introduced by MixUp—specifically, when keypoint positions lie between two true locations. Compared to direct coordinate regression, this heatmap-based approach demonstrates greater robustness to positional variations, thereby enhancing the model's generalization capability.

To align with the learning objectives at different training stages, we employ a linearly decaying MixUp augmentation strategy. In the early stages of training,

Algorithm 1: PyTorch-like pseudocode for our MixUp method

```
# Input:
# - images I: Batch of input image tensors (B, C, H, W).
# - heatmaps H: Corresponding heatmap tensors (B, K, H, W).
# - landmarks L: Corresponding landmark tensors (B, 2K).
# - mixup parameter alpha: Beta distribution parameter.
# Output:
# - mixed_images: Mixed images tensor (B, C, H, W).
# - mixed_heatmaps: Mixed heatmaps tensor (B, 2K, H, W).
# - mixed_landmarks: Mixed landmarks tensor (B, 2, 2K).
# - mixed_weights: Mixing weights tensor (B, 2).

def mixup_data(images, heatmaps, landmarks, alpha=0.2, device='cuda'):
    # Sample mixing coefficient from Beta distribution
    lam = np.random.beta(alpha, alpha) if alpha > 0 else 1.0
    # Randomly permute batch indices
    index = torch.randperm(images.size(0), device=device)
    # Mix images using linear interpolation
    mixed_images = lam * images + (1 - lam) * images[index]
    # Mix heatmaps
    mixed_heatmaps = torch.cat([lam * heatmaps, (1 - lam) * heatmaps[index]], dim=1)
    # Stack landmarks
    mixed_landmarks = torch.stack([landmarks, landmarks[index]], dim=1)
    # Mixing weights
    mix_weights = torch.tensor([lam, 1 - lam], device=device).repeat(images.size(0), 1)
    return mixed_images, mixed_heatmaps, mixed_landmarks, mix_weights
```

introducing moderate data augmentation helps the model learn more robust feature representations. This is particularly important for addressing the inherent speckle noise, low contrast, and significant variations in fetal pose commonly present in ultrasound images. The mixed samples during this stage also assist the model in developing an understanding of the regions and general shapes associated with anatomical keypoints. In the later stages of training, we gradually reduce the intensity of MixUp augmentation, enabling the model to focus on fine-tuning keypoint localization and achieving sub-pixel accuracy. Maintaining high localization precision at this stage is critical. Applying strong mixing strategies when the model is approaching convergence may introduce ambiguous labels or distort spatial structures, which can interfere with the learning of precise keypoint positions. The proposed decaying mechanism effectively mitigates this issue by adaptively adjusting the augmentation strength according to the

training progress. We provide the corresponding pseudo-algorithm description algorithm 1.

3 Experiments

3.1 Dataset and Implementation Details

Participants in the MICCAI IUGC 2025 challenge are provided with 300 annotated ultrasound images and 31,421 unannotated images. Additionally, 2,045 unlabeled standard AOP-plane images are released as exemplars; these exemplars are exclusively drawn from the unannotated pool and are intended to reflect the overall distribution of the unlabeled data.During the model training, all models are implemented using PyTorch and trained on a single NVIDIA Tesla 4090 GPU. The standard deviation of the Gaussian heatmap is set to $\sigma = 6$. During training, the following eleven data augmentation operations are applied: rotation, scaling, translation, brightness adjustment, Gaussian blur, gamma contrast, elastic transformation, image inversion, Cutout, and Coarse Dropout. During training, pixel-wise mean squared error (MSE) is adopted as the loss function for optimization, and the Euclidean distance between predicted and ground truth keypoints on the validation set is used as the evaluation metric for model performance, with only the best-performing parameters being saved. During inference, the DARK post-processing method is introduced as a debiasing strategy to further enhance the robustness and localization accuracy of the predictions.

Table 1. Data augmentation methods and their parameters used in the HeatmapLandmarkDataset

Augmentation Method	Parameters
Rotation	Range: -5 to 5 degrees
Scale	Range: $(1 - 0.125)$ to $(1 + 0.125)$
Translation	X-axis: $[-30, 30]$, Y-axis: $[-20, 20]$ pixels
Brightness Adjustment	Multiplication factor range: $(1 - 0.6)$ to $(1 + 0.6)$
Gaussian Blur	Applied with probability 0.1, Sigma range: $(0, 1.5)$
Gamma Contrast	Adjustment range: $(0.3, 2.0)$
Elastic Transformation	Alpha range: $(0, 400)$, Sigma fixed at 30
Image Inversion	Applied with probability 0.1
Cutout	Iterations: 0 or 1, Size range: 0.04 to 0.3 of image width/height, Non-square cutouts
Coarse Dropout	Rate: 0.02, Area size: 8% of image size

3.2 Model Selection and Hyperparameter Tuning

During the model selection and hyperparameter tuning phase, we conducted experiments based on a U-shaped network with an encoder-decoder architecture.

In the initial stage, we first adopted a baseline UNet model to perform module adjustments and hyperparameter optimization (as shown in Table 2). After determining the optimal hyperparameter configuration, we selected the best-performing backbone network based on this configuration (as shown in Table 3).

Table 2. Results of Module Adjustment and Hyperparameter Tuning

Method	Distance ↓	AOP_MAE ↓
b0 (lr: 0.0001; σ: 2; step_size: 15; γ: 0.9)	33.54	16.63
b1 (lr: 0.0001; σ: 3; step_size: 10; γ: 0.9)	28.88	11.43
b2 (lr: 0.0001; σ: 6; step_size: 10; γ: 0.9)	25.49	8.24
b3 (b2 + more aug data; step_size: 10; γ: 0.9)	22.51	8.62
b4 (b3 + decoder-dropout: [0.05, 0.1, 0.2, 0.3])	21.67	8.32
b5 (σ: 4)	25.14	8.67
b6 (γ: 0.75)	28.24	11.11
b7 (step_size: 8)	31.11	12.22
b8 (b2 + supervise the 3^{rd} and 4^{th} layer outputs)	24.83	8.54
b9 (b3 + ConvBlock: InstanceNorm + LeakyReLU)	22.47	8.77
b10 (b3 + ConvBlock: InstanceNorm + PReLU [13])	23.25	8.09
b11 (b3 + ConvBlock: BatchNorm + PReLU [13])	24.50	8.36
b12 (b3 + DARK [14])	20.45	7.34

As shown in Table 2, the following explains how to read the table. First, the table should be read from top to bottom. Our experimental exploration starts from the initial baseline b0, and new modules or hyperparameters are gradually introduced on this basis. Whenever a newly added module or adjusted hyperparameter configuration yields better performance than the previous baseline, it is established as the new baseline and highlighted in the table. If a row's description does not explicitly mention a reference baseline, it means the configuration is derived from the immediately preceding baseline with parameter modifications. For example, b5 is modified from b4, b6 is based on b5, and b7 further improves upon b6. Through this step-by-step optimization process, we ultimately determine b12 as the optimal configuration, which is then adopted as the hyperparameter setting for all subsequent experiments. In the table, `sigma` denotes the standard deviation σ used for generating Gaussian heatmaps, while `step_size` and γ are the step size and decay factor of the StepLR learning rate scheduler, respectively. In the table, ConvBlock refers to the ConvBlock used in the UNet decoder.

Given that our network adopts a U-shaped encoder-decoder architecture, guided by baseline b12 in Table 2, we trained multiple different backbones to determine the optimal choices for the teacher and student models. As shown in Table 3, DenseNet121 [11] achieved the best performance and was therefore selected as the final backbone for our encoder.

Table 3. Performance comparison of different backbone networks on the IUGC dataset

Backbone	Distance ↓	AOP_MAE ↓
UNet [15]	23.70	9.04
PVT-v2-b1 [16]	20.99	5.87
PVT-v2-b2 [16]	19.06	5.60
ResNet18d [17]	26.16	7.59
ConvNeXtv2-Tiny [18]	15.51	6.63
ResNet34 [17]	25.11	8.01
seresnext26d_32x4d [19]	20.75	7.02
DenseNet121 [11]	18.08	5.99
DenseNet161 [11]	18.64	6.24

3.3 Pseudo-label filtering

We first perform 5-fold cross-validation on 300 annotated images (240 for training and 60 for validation) and construct a UNet model with DenseNet121 as the backbone to train the teacher model. Based on the two best-performing models on the validation set, we conduct inference on the 2,045 unlabeled images, initially setting the pseudo-label selection threshold to a maximum heatmap activation value of no less than 0.7.

However, due to artifacts, acoustic shadows, and low contrast in ultrasound images, the response magnitude becomes decoupled from localization accuracy: high-confidence false positives may arise at incorrect locations, while true positives at correct landmarks are erroneously filtered out due to weak signals (loss of low-confidence true positives). Moreover, variations in echogenicity across structures cause a uniform threshold to systematically favor "easy-to-detect" landmarks (e.g., the pubic symphysis), while neglecting clinically critical yet challenging structures, thereby exacerbating class imbalance. To this end, we explored an uncertainty-based pseudo-labeling paradigm, for example, an entropy-based correction strategy incorporating temperature scaling sharpness and per-landmark adaptive thresholds, to mitigate the confidence bias issue. However, due to its implementation complexity and difficulty in achieving stable convergence under the current data conditions, this approach was ultimately not successfully deployed.

Table 4. Comparison of different training-data compositions on the IUGC dataset

Training Data	Distance ↓	AOP_MAE ↓
pse825_1e260	18.35	6.85
pse825_se_165_le260	18.27	6.74
pse825_se_100_le260	17.96	6.57
pse825_se_100_1e220	18.24	5.58

Table 5. Comparison of different training-data with mixup compositions on the IUGC dataset.

Training Data	Distance ↓	AOP_MAE ↓
pse825_1e260	17.67	5.06
pse825_se_165_le260	17.43	5.10
pse825_se_100_le260	16.29	5.47
pse825_se_100_1e220	16.35	4.94

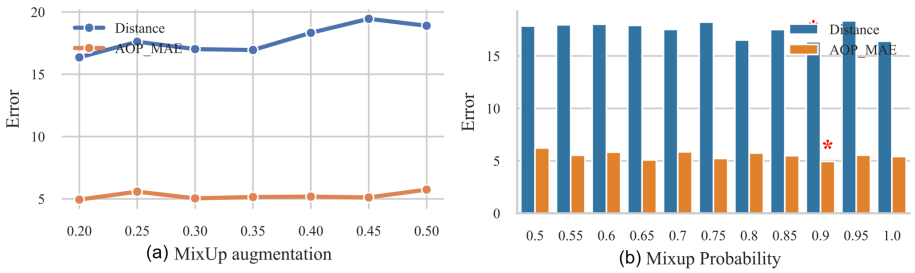

Fig. 3. Impact of MixUp application augmentation and probability on Distance and AOP-MAE.

The configuration `pse825_le260` trains the model using 825 pseudo-labeled images generated via the pseudo-labeling technique and 260 fully annotated images. Building upon this, `pse825_se_165_le260` incorporates an additional 165 pseudo-labeled images derived from the original unlabeled pool, while keeping the number of annotated images fixed at 260. Similarly, `pse825_se_100_le260` and `pse825_se_100_le220` both introduce pseudo-labels from 100 selected unlabeled images; however, the former uses 260 annotated images, whereas the latter reduces the number of annotated images to 220, aiming to evaluate the compensatory capability of pseudo-labels when labeled data is reduced., aiming to evaluate the compensatory capability of pseudo-labeling when labeled data is reduced.

As a compromise, we designed a multi-threshold filtering strategy.: we first identified high-confidence samples using higher thresholds 0.7, then excluded these from the pseudo-labels generated under a lower threshold (0.6), retaining moderate-confidence candidates with broader spatial coverage. This yielded 825 pseudo-labeled images with more comprehensive representation. Further analysis revealed a significant number of redundant samples with similar imaging angles, which could lead to overfitting. Therefore, we randomly divided these 825 images into five non-overlapping subsets, to be used for training. Furthermore, to reduce redundancy in the annotated data, we removed 40 highly similar samples from the original 260 annotated images and transferred them to the validation set. Fi-

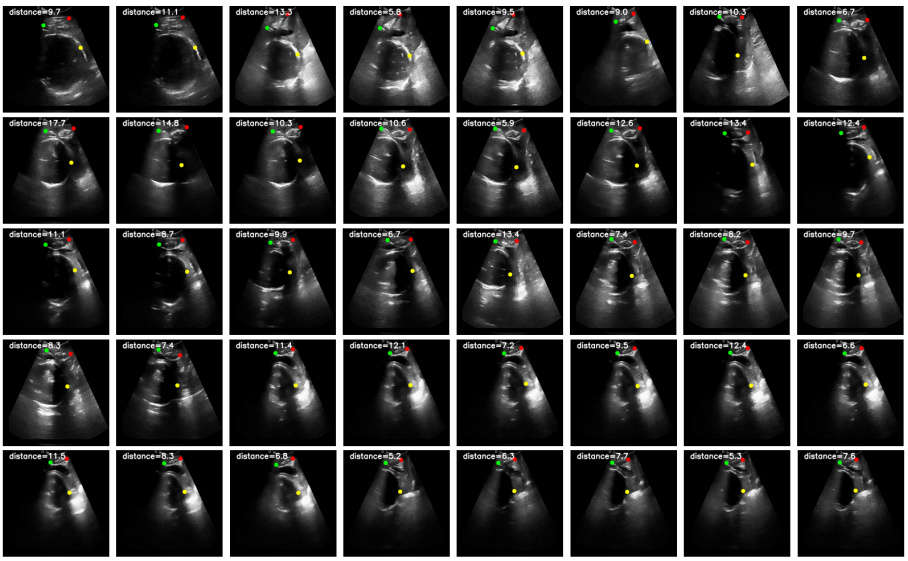

Fig. 4. Visualization of the final ensemble model's predictions on validation set samples. The predicted keypoint locations on the input images are shown, where red dots indicate PS1, green dots indicate PS2, and yellow dots indicate FH1. The distance difference (Distance) between the predicted results and the ground truth labels is annotated in the top-left corner of each image.

nally, the student model's training set consists of 220 annotated images and 100 high-quality pseudo-labeled images. This strategy led to a modest improvement in model performance, as shown in Table 4. Based on the aforementioned dataset and with the Mixup augmentation strength set to 0.2, we further evaluate the impact of different training data combinations on the performance of the student model, with specific results presented in Table 5. Notably, the models under the configurations `pse825_se100_le220` and `pse825_se100_le260` exhibit particularly outstanding performance. We therefore choose to ensemble the models corresponding to the best weights obtained under these two configurations and visualize the predictions of this ensemble model on the validation set used during local training, as shown in Figure 4. Subsequently, this ensemble model is submitted as the final test model for the IUGC 2025 challenge. On the IUGC 2025 official validation set, the ensemble model achieves Distance and AOP_MAE scores of 15.85 and 5.14, respectively, and further improves to 13.16 and 4.42 on the independent test set.

On the pse825_se100_le220 training dataset, we conducted experiments on the strength of Mixup augmentation, with results shown in subplot (a) of Figure 3. The results indicate that the model achieves optimal performance when the Mixup strength is set to 0.2. This is primarily because this strength value strikes a good balance between preserving keypoint spatial accuracy and enhancing model generalization: under this setting, the mixed samples generated by Mixup can better retain the clear structure of the original images, effectively avoiding blurring or distortion of keypoints caused by excessive interpolation; at the same time, the moderate introduction of data diversity helps alleviate overfitting in small-sample scenarios, thereby improving model robustness. Building upon the determination that the optimal Mixup strength is 0.2, we further adjusted its application probability (as shown subplot (b) of Figure 3, where the best result is marked with a red star). All experiments set the minimum application probability of Mixup to 0.1 to ensure that the model can converge more stably as training approaches the end.

4 Conclusion

This study proposes a semi-supervised learning method based on the Noisy Student framework to address the issue of insufficient accuracy in fetal anatomical landmark detection from labor ultrasound images. We construct a Dense-UNet model with DenseNet121 as the encoder and introduce a linearly decaying MixUp data augmentation strategy specifically designed for heatmap regression during training, effectively enhancing the model's robustness and generalization capability. On the competition dataset containing 34,421 images, our method achieved significant performance improvement by using only 220 labeled images and an additional 100 carefully selected unlabeled images for training. Experimental results demonstrate that the Noisy Student training framework can effectively leverage the value of unlabeled data even under extremely limited labeling conditions, revealing its potential for fetal keypoint detection in labor ultrasound.

Acknowledgements This work was supported in part by the Research and Practice Project of Research Teaching Reform in Henan Undergraduate University under Grant 2022SYJXLX114, in part by the Key Research Programs of Higher Education Institutions in Henan Province under Grant 24B520026 and 25A520041, and in part by the Special Research Project for the Construction of Provincial Demonstration Schools at Nanyang Institute of Technology under Grant SFX202314, and in part by the Interdisciplinary Sciences Project, Nanyang Institute of Technology.

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

References

1. Bai, J., Khobo, I., Lu, Y., Ni, D., Yaqub, M., Lekadir, K., Ma, J., Li, S.: Landmark detection challenge for intrapartum ultrasound measurement meeting the actual clinical assessment of labor progress (Apr 2025). https://doi.org/10.5281/zenodo.15172238, https://doi.org/10.5281/zenodo.15172238
2. Bai, J., Lekadir, K., Ni, D., Slimani, S., Campello, V.M., Ohene-Botwe, B., Lu, Y., Chen, G., Hou, H., Qiu, D., Zhou, Z.: Intrapartum ultrasound grand challenge 2024 (Apr 2024). https://doi.org/10.5281/zenodo.10979813, https://doi.org/10.5281/zenodo.10979813
3. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence (2020), https://arxiv.org/abs/2001.07685
4. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning (2019), https://arxiv.org/abs/1905.02249
5. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=HklkeR4KPB
6. Li, R., Li, M., Liu, W., Zhou, Y., Zhou, X., Yao, Y., Zhang, Q., Chen, H.: Unimatch: Universal matching from atom to task for few-shot drug discovery. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=v9EjwMM55Y
7. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification (2020), https://arxiv.org/abs/1911.04252
8. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 18613–18624. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf
9. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.: Deep networks with stochastic depth (2016), https://arxiv.org/abs/1603.09382
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning (2016), https://arxiv.org/abs/1506.02142

11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
12. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=r1Ddp1-Rb
13. Pinto, R.C., Tavares, A.R.: Prelu: Yet another single-layer solution to the xor problem (2024), https://arxiv.org/abs/2409.10821
14. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation (2019), https://arxiv.org/abs/1910.06278
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597
16. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022). https://doi.org/10.1007/s41095-022-0274-8
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385
18. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders (2023), https://arxiv.org/abs/2301.00808
19. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks (2017), https://arxiv.org/abs/1611.05431