CAMS: Coordinate-Aware Multi-task Semi-supervised Learning for Ultrasound Landmark Localization

Chenke Li[1†], Xiaoshuang Li[1†], Yuanhan Wang[1], Beining Wu[1], Feiwei Qin[1⋆], and Yifei Chen[2⋆]

[1] Hangzhou Dianzi University, Hangzhou, China
[2] Tsinghua University, Beijing, China
qinfeiwei@hdu.edu.cn, justlfc03@gmail.com

Abstract. Accurate anatomical landmark localization is fundamental to ultrasound biometry, yet manual identification is time-consuming and suffers from observer variability, while automated methods face substantial heterogeneity across tasks and devices and scarce annotations alongside abundant unlabeled data. We propose a Coordinate-Aware Multi-task Semi-supervised (CAMS) framework for unified ultrasound landmark detection. CAMS adopts a shared DINOv2 encoder to learn transferable representations, and equips each task with a coordinate-aware heatmap decoder and a task-specific DSNT module for continuous coordinate regression. To exploit unlabeled data, CAMS is optimized within a Mean Teacher paradigm using an epoch-dependent confidence-masked consistency strategy. On the MICCAI 2026 FU_Biometry dataset spanning nine tasks, CAMS outperforms representative semi-supervised baselines, reducing the average Mean Radial Error from 32.63 to 23.39 and the Mean Absolute Error from 33.16 to 31.01, demonstrating strong generalization across diverse clinical scenarios and imaging devices. Our source code will is available at https://github.com/lizizai0735/CAMS.

Keywords: Ultrasound biometry · Landmark detection · Semi-supervised learning · Mean Teacher · DINOv2

1 Introduction

Ultrasound biometry is widely used in prenatal examinations [18], intrapartum assessment [6], cervical ultrasound evaluation [3], and adult transthoracic echocardiography [14]. Many clinically important biometric measurements rely on accurate localization of anatomical landmarks, such as fetal cranial landmarks, femoral endpoints, pubic-symphysis landmarks, and cardiac anatomical reference points [8,9]. However, these landmarks are still predominantly identified manually in routine clinical practice, making the process time-consuming and susceptible to considerable intra- and inter-observer variability [1]. Therefore,

⋆ Corresponding authors. † Co-first authors.

developing a robust and reliable automatic landmark detection method that can generalize across different tasks, ultrasound devices, and clinical scenarios in real-world clinical practice is of great clinical significance.

To address the need for automated and standardized ultrasound biometry, the MICCAI 2026 Foundation Model for Ultrasound Biometry (FU_Biometry) Challenge was established as a unified benchmark for anatomical landmark localization across heterogeneous ultrasound applications. The challenge covers multiple prenatal, intrapartum, cervical, and echocardiographic tasks, involving substantial variations in anatomical structures, landmark configurations, image appearance, and acquisition devices. In addition to a limited set of landmark-annotated images, FU_Biometry provides a large collection of unlabeled ultrasound images. Although these unlabeled data contain valuable information, they cannot be directly exploited by conventional fully supervised learning methods. Recent ultrasound landmark-localization approaches have explored iterative pseudo-labeling [13], progressive pseudo-label refinement and model ensembling [22], Mean Teacher-based coarse-to-fine localization [4], Noisy Student self-training [11], and DSNT-based coordinate regression [20]. However, most of these methods were developed for a single task and often rely on task-specific architectures or training procedures [2,21], limiting their ability to learn transferable representations across the heterogeneous domains included in FU_Biometry.

To address these challenges, we propose a **Coordinate-Aware Multi-task Semi-supervised** (**CAMS**) learning framework. CAMS employs a pretrained DINOv2 model [16] as a shared feature encoder to learn transferable visual representations across different ultrasound tasks. Meanwhile, each task employs a coordinate-aware decoder combined with the Differentiable Spatial-to-Numerical Transform(DSNT) [15] to achieve continuous landmark coordinate prediction. Furthermore, CAMS is optimized within the Mean Teacher framework [19], together with an epoch-dependent confidence strategy for utilizing unlabeled data.

2 Proposed Method

2.1 Model Architecture

As illustrated in Fig. 1, we propose a semi-supervised multi-task landmark detection framework built upon DINOv2 [16] and the Mean Teacher paradigm [19]. The framework follows a shared-encoder, task-specific-decoder architecture and consists of three key components. First, a shared DINOv2 backbone is employed to extract high-level semantic representations that are shared across different landmark detection tasks. Second, a set of task-specific heatmap heads are introduced to learn discriminative localization features for each individual task independently. Finally, each task is equipped with an individual DSNT module [15], which converts the predicted heatmaps into continuous landmark coordinates for coordinate localization. To leverage the abundance of unlabeled clinical ultrasound data, the entire framework is optimized end-to-end within the Mean Teacher semi-supervised learning framework [24,23].

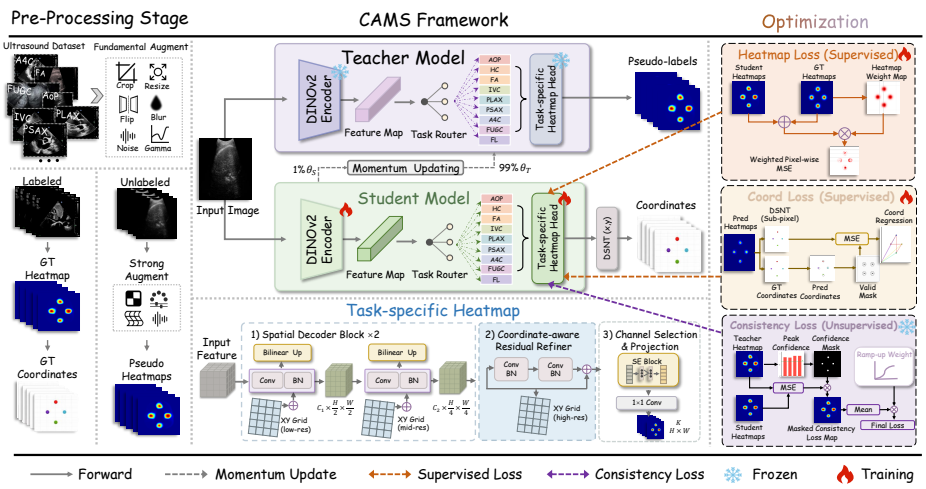

Fig. 1. Overview of **CAMS**. A shared DINOv2 encoder feeds task-specific coordinate-aware decoders, with DSNT producing continuous coordinates. The student learns via supervised and confidence-masked consistency losses; the teacher updates via EMA.

2.2 DINOv2-based Feature Encoder

Vision Transformers (ViTs) represent an image as a sequence of patch tokens and model long-range interactions using self-attention [5]. Instead of adopting a purely convolutional backbone, we employ a pretrained DINOv2 ViT-S/14 [16] as the shared feature extractor. Given an input ultrasound image, the image is first partitioned into a sequence of fixed-size patches through the Patch Embedding layer, where each patch is projected into a visual token. These tokens are then processed by multiple Transformer encoder layers to capture global contextual information while preserving local structural representations. Following the standard DINOv2 architecture, with the classification head removed, normalized patch tokens are extracted from the final layer. Finally, the one-dimensional token sequence is reshaped into a two-dimensional feature map, $F_{\text{in}} \in \mathbb{R}^{C \times H \times W}$, serving as input to the subsequent landmark heatmap decoder.

2.3 Coordinate-aware Heatmap Decoder

The decoder progressively restores spatial resolution using bilinear interpolation followed by convolutional feature refinement. Following the coordinate-channel principle introduced by CoordConv [10], a pair of normalized two-dimensional coordinate grids, X_{grid} and Y_{grid}, ranging within the continuous interval of $[-1, 1]$, are generated at each upsampling stage:

$$F_{\text{coord}} = \text{Concat}\left(F_{\text{up}}, X_{\text{grid}}, Y_{\text{grid}}\right), \tag{1}$$

where F_{up} denotes the upsampled feature map, X_{grid} and Y_{grid} represent the normalized coordinate grids along the horizontal and vertical directions, respectively, and F_{coord} denotes the coordinate-enhanced feature map obtained after performing a channel-wise concatenation of these components.

To enhance task-specific representations, the refined features are recalibrated by a residual Squeeze-and-Excitation (SE) module [7]. The recalibrated features are then projected into landmark heatmaps:

$$F_{\text{out}} = F_{\text{res}} \odot \sigma\left(W_2\, \delta\left(W_1\ \text{GAP}\left(F_{\text{res}}\right)\right)\right), \tag{2}$$

where F_{res} denotes the residual feature map enriched with high-frequency details, $\text{GAP}(\cdot)$ denotes global average pooling, $\sigma(\cdot)$ represents the sigmoid activation function, and $\delta(\cdot)$ denotes the GELU activation function. The fully connected layer W_1 reduces the channel dimension with a reduction ratio of $r = 16$, while W_2 restores its original dimension. This module suppresses background interference and adaptively recalibrates channel responses, guiding each task-specific decoder toward discriminative anatomical features.

2.4 Task-specific DSNT-based Coordinate Regression

Conventional landmark localization methods typically employ the Argmax operation to estimate landmark coordinates from the predicted heatmaps. However, the Argmax operation is non-differentiable, which prevents end-to-end gradient propagation and introduces significant quantization errors due to the limited spatial resolution of heatmaps. To address these limitations, we adopt the DSNT layer to achieve continuous coordinate regression. Considering that different anatomical landmark detection tasks exhibit distinct spatial distributions and scales, sharing a single coordinate transformation module across all tasks may lead to gradient interference during multi-task optimization. Therefore, instead of using a globally shared DSNT layer, we assign an individual DSNT module to each task, where each module is equipped with a task-specific learnable temperature parameter τ_t. Given the predicted heatmap of the c-th landmark for task t, the heatmap is first transformed into a normalized spatial probability distribution using a task-specific Spatial Softmax:

$$P_c^{(t)}(x, y) = \frac{\exp\left(\tau_t H_c^{(t)}(x, y)\right)}{\sum_{x', y'} \exp\left(\tau_t H_c^{(t)}(x', y')\right)}, \tag{3}$$

where $H_c^{(t)}(x, y)$ denotes the response value of the predicted heatmap for the c-th landmark of task t at each spatial location (x, y), $P_c^{(t)}(x, y)$ represents the corresponding normalized spatial probability value, and τ_t is the learnable temperature parameter associated with task t.

The final continuous landmark coordinates $(\hat{x}_c, \hat{y}_c)$ are obtained by computing the expectation of the normalized spatial probability distribution over the

coordinate grid defined in the range $[-1, 1]$:

$$\hat{x}_c^{(t)} = \sum_{x,y} x\, P_c^{(t)}(x, y), \qquad \hat{y}_c^{(t)} = \sum_{x,y} y\, P_c^{(t)}(x, y), \tag{4}$$

where $(\hat{x}_c^{(t)}, \hat{y}_c^{(t)})$ denote the predicted coordinates of the c-th landmark for task t, and (x, y) are the normalized spatial coordinates on the coordinate grid.

2.5 Mean Teacher-based Semi-supervised Learning

To fully exploit the large amount of unlabeled ultrasound data, the proposed model is optimized under the Mean Teacher semi-supervised learning framework [19]. The student model, denoted by θ_S, is updated through backpropagation based on the training loss, whereas the parameters of the teacher model, denoted by θ_T, are updated as the exponential moving average (EMA) of the historical parameters of the student model. The EMA decay rate is set to 0.999, which ensures smooth parameter updates and stable pseudo-label generation.

Supervised loss. For labeled samples, the supervised objective consists of a heatmap-based spatial loss and a DSNT-based coordinate regression loss. To address the varying difficulty levels across different ultrasound views, we introduce a task-specific difficulty multiplier, α_t, which assigns higher penalty weights to notoriously challenging tasks (e.g., Head Circumference and Short-Axis views). The supervised loss for task t is formulated as:

$$\mathcal{L}_{\text{sup}} = \alpha_t \cdot \left[\mathcal{L}_{\text{MSE\_Weighted}} \left(H_{\text{pred}}, H_{\text{gt}} \right) + 50.0\, \mathcal{L}_{\text{MSE}} \left(C_{\text{pred}}, C_{\text{gt}}^{\text{noisy}} \right) \right], \tag{5}$$

where H_{pred} and H_{gt} denote the final predicted and ground-truth heatmaps, respectively, while C_{pred} and $C_{\text{gt}}^{\text{noisy}}$ represent the predicted coordinates and the randomly perturbed ground-truth coordinates.

For heatmap supervision, a spatial weighting mask, $W = 9.0\, H_{\text{gt}} + 1.0$, is employed to emphasize the Gaussian peak regions in the ground-truth heatmaps. Consequently, the Gaussian peak regions receive a penalty weight of 10, whereas the background regions retain a weight of 1. For coordinate supervision, we replace the conventional Smooth L1 loss with the Mean Squared Error (MSE) loss, $\mathcal{L}_{\text{MSE}}$, and empirically increase its scaling factor to 50.0. This modification is crucial because MSE applies a quadratically increasing penalty to severe spatial outliers (e.g., predictions deviating by hundreds of pixels in large-scale anatomical structures), thereby forcing the network to rapidly correct global structural misunderstandings. In addition, a dynamic label-perturbation strategy is introduced by injecting Gaussian noise, $\mathcal{N}(0, \sigma_e^2)$, into the ground-truth coordinates C_{gt}, where the standard deviation σ_e decreases linearly with the training epoch. This perturbation acts as an effective regularizer, encouraging the network to learn anatomical representations robust to small spatial variations, thereby improving generalization across ultrasound devices.

Unsupervised consistency loss. For unlabeled samples, the teacher model generates pseudo heatmaps, denoted by $\hat{H}_{\text{teacher}}$. To prevent the self-amplification of erroneous pseudo labels during early training, when the student model is insufficiently optimized, we design a progressive dynamic confidence mask M. Rather than utilizing a fixed threshold—which either discards too many valid pseudo-labels early on (if set too high) or introduces severe confirmation bias (if set too low)—we propose an epoch-dependent linear schedule. The confidence threshold $\theta_{\text{conf}}(e)$ is gradually increased with training progress according to:

$$\theta_{\text{conf}}(e) = 0.5 + 0.35 \left(\frac{e}{E_{\max}} \right), \tag{6}$$

where e denotes the current epoch and $E_{\max}$ represents the total training epochs. This linear curriculum starts at a relatively low threshold of 0.5 to encourage early participation of unlabeled data, and gradually scales up to a strict 0.85 to ensure high-quality pseudo-label filtering as the model matures.

Only the landmark channels whose maximum heatmap probability predicted by the teacher model exceeds $\theta_{\text{conf}}(e)$ are involved in the consistency optimization. Furthermore, while challenging tasks require stronger supervision (larger α_t), their corresponding pseudo-labels generated by the teacher model are inherently less reliable. To prevent the network from reinforcing incorrect predictions on hard tasks, we apply an inverse task weight $(1/\alpha_t)$ to the consistency loss:

$$\mathcal{L}_{\text{unsup}} = \frac{1}{\alpha_t} \frac{1}{|M|} \sum \left(\sigma(\hat{H}_{\text{student}}) - \sigma(\hat{H}_{\text{teacher}}) \right)^2 \odot M, \tag{7}$$

where $\hat{H}_{\text{student}}$ and $\hat{H}_{\text{teacher}}$ denote the output predicted spatial heatmaps of the student and teacher models, respectively, $\sigma(\cdot)$ represents the standard sigmoid activation function, M is the dynamic confidence mask, and $|M|$ denotes the total number of valid landmark channels selected by the mask.

Finally, the overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + w(e)\mathcal{L}_{\text{unsup}}, \tag{8}$$

where $w(e)$ is the consistency weight. A Gaussian ramp-up schedule is adopted to gradually increase $w(e)$ during the first 40% of the training process, with the maximum consistency weight set to 2.0.

3 Experiments and Results

3.1 Dataset and Experiments

Dataset. We conduct all experiments on the official FU_Biometry dataset released for the MICCAI 2026 Foundation Model for Ultrasound Biometry Challenge. The dataset is designed to establish a unified benchmark for ultrasound biometry across prenatal, intrapartum, and adult echocardiography scenarios. To capture the device heterogeneity encountered in routine clinical practice, the

ultrasound images were acquired using scanners from multiple vendors, including Philips CX50, Toshiba Aplio 300, GE Voluson P8, Esaote MyLab, Lian-med ObEye, and Youkey Q7. Image acquisition followed the relevant task-specific guidelines for intrapartum ultrasound [6], fetal biometry and growth [18], cervical ultrasound assessment [3], and adult transthoracic echocardiography [14]. The dataset contains nine ultrasound biometry tasks (A4C, AoP, FA, HC, IVC, PLAX, PSAX, FUGC, and fetal_femur), supporting more than 20 clinically relevant biometric measurements. The dataset consists of both labeled and unlabeled images. All labeled images are annotated with task-specific anatomical landmark coordinates, which serve as the primary supervision for landmark localization. For a subset of tasks, additional clinically derived biometric measurements are provided. In contrast, the unlabeled subset contains only ultrasound images without any annotation and is used exclusively for semi-supervised consistency learning. The detailed data statistics are shown in Table 1.

Table 1. Summary of datasets used in this study. The * indicates that the dataset contains more than three distinct resolutions.

Dataset	**Labeled**	**Unlabeled**	**Resolution**
A4C	108	63,390	256×256, 800×600, 1024×768
AoP	4,000	38,666	512×512, 1080×852, 720×564
FA	500	27,188	744×562, 1024×768
fetal_femur	727	0	959×661, 640×392, 961×663*
FUGC	260	23,230	1136×864, 544×336
HC	999	33,480	720×564, 800×540, 800×542*
IVC	38	1,554	800×600, 1024×768
PLAX	87	1,281	800×600, 1024×768
PSAX	49	2,381	800×600, 1024×768
Total	**6,768**	**191,170**	

All labeled and unlabeled images were enhanced by applying Contrast-Limited Adaptive Histogram Equalization (CLAHE) [17] to the luminance channel in the LAB color space and were subsequently converted to RGB. During training, images were resized to 518×518 and augmented using ShiftScaleRotate ($p = 0.5$), color jittering ($p = 0.5$), Gaussian blur ($p = 0.2$), Gaussian noise ($p = 0.1$), and Coarse Dropout ($p = 0.05$). For labeled samples, geometric transformations were synchronously applied to the corresponding landmark coordinates. All images were finally normalized using the ImageNet mean and standard deviation values. Validation images underwent the exact same CLAHE preprocessing, resizing, and normalization, without any stochastic augmentation.

Experimental Setup and Training Details. All models were implemented in PyTorch and trained on a single NVIDIA A40 GPU, with the random seed fixed to 42 for reproducibility. Parameters were optimized using AdamW [12] with a weight decay of 1×10^{-4}, using module-specific learning rates: 2×10^{-5}

for fine-tuning the pretrained DINOv2 encoder and 1×10^{-3} for the task-specific heatmap heads and DSNT modules. A OneCycle policy with cosine annealing was applied over 50 epochs, with a 30% warm-up phase. Both labeled and unlabeled batch sizes were set to 4. For semi-supervised learning, we adopted the Mean Teacher framework [19], where the teacher was updated by an EMA of the student parameters with decay $\alpha = 0.999$. The consistency-loss weight followed a Gaussian ramp-up over the first 40% of training, capped at 2.0. Pseudo-label filtering used an epoch-dependent confidence threshold θ_{conf} that increased linearly from 0.5 to 0.85 to exclude low-confidence landmark channels.

Evaluation metrics. Following the ChallengeR-style rank-then-aggregate evaluation protocol, the final score is computed from two equally weighted metrics. Mean Radial Error (MRE) quantifies landmark localization accuracy as the average Euclidean distance, measured in pixels, between the predicted landmarks p_k and the reference landmarks g_k:

$$\text{MRE} = \frac{1}{K} \sum_{k=1}^{K} \|p_k - g_k\|_2, \tag{9}$$

Mean Absolute Error (MAE) measures the accuracy of clinical parameters via the mean absolute error $|\hat{y}_j - y_j|$. To prevent parameters with large dynamic ranges from dominating the aggregated score, each parameter error is standardized using its clinically defined tolerance range or interquartile range (IQR).

Table 2. Performance comparison of representative semi-supervised methods on the FU_Biometry 2026 validation set. The 1st and 2nd best results are highlighted, and the results of our proposed CAMS method are marked in **bold**.

Method	Prenatal Ultrasound						Intrapartum Ultrasound				Cardiac Ultrasound								Average	
	HC		FA		fetal_femur		AoP		FUGC		IVC		A4C		PLAX		PSAX		Overall	
	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE
1st-place	51.80	61.93	27.59	92.56	31.85	26.77	16.87	8.04	161.07	301.25	79.90	12.89	32.78	29.23	22.59	11.43	44.63	24.71	52.12	63.20
2nd-place	47.69	65.45	22.01	76.31	43.50	36.23	21.97	12.88	29.43	19.74	57.64	70.12	35.21	23.67	26.80	25.37	54.63	37.21	37.65	40.78
3rd-place	273.85	1075.72	209.83	1235.31	202.11	363.05	147.68	61.37	182.71	308.49	80.50	22.66	129.09	122.53	95.33	72.18	104.18	70.68	158.36	370.22
4th-place	50.82	75.36	26.76	98.26	29.56	20.74	15.94	12.53	12.79	8.42	55.73	15.84	37.01	31.61	24.38	12.10	40.65	23.57	32.63	33.16
CAMS (Ours)	**44.76**	**74.81**	**23.11**	**110.75**	**21.59**	**19.50**	**13.87**	**7.79**	**11.56**	**12.81**	**17.55**	**9.60**	**25.42**	**20.09**	**16.68**	**8.10**	**35.93**	**15.64**	**23.39**	**31.01**

3.2 Experiment Results

Since CAMS targets semi-supervised ultrasound landmark localization under limited annotations, we compare it against representative semi-supervised methods ranked within the top 5 of the IUGC 2025 Challenge, as summarized in the official benchmark paper [1].The fifth-ranked fully supervised approach is excluded from comparisons, since it cannot exploit large-scale unlabeled data and fails to match our semi-supervised experimental setup. As shown in Table 2, CAMS achieves the lowest average MRE and MAE among all reproduced baselines under the unified multi-task evaluation protocol. Compared with the strongest reproduced baseline, CAMS reduces the average MRE from 32.63 to

23.39 and the average MAE from 33.16 to 31.01, demonstrating improved aggregate performance across heterogeneous ultrasound tasks.

Compared with perturbation-based semi-supervised methods such as Noisy Student, **CAMS** always achieves lower localization errors. This improvement can be attributed to three components of the proposed framework. First, the shared DINOv2 encoder learns transferable visual representations across heterogeneous ultrasound tasks. Second, the task-specific coordinate-aware decoder preserves task-dependent localization characteristics while enabling multi-task learning. Third, the confidence-masked Mean Teacher strategy gradually increases consistency weight and pseudo-label confidence threshold during training, enabling the model to exploit unlabeled data while suppressing unreliable pseudo labels.

4 Conclusion

In this work, we proposed **CAMS**, a coordinate-aware multi-task semi-supervised framework for ultrasound landmark localization. By combining a shared DINOv2 encoder, task-specific coordinate-aware heatmap decoders and DSNT modules, and a confidence-masked Mean Teacher strategy, **CAMS** unifies limited annotations and abundant unlabeled data while preserving task-dependent characteristics. Experiments on the FU_Biometry dataset across nine tasks show that CAMS achieves the lowest average MRE and MAE among the reproduced baselines. These results confirm its strong generalization across diverse anatomies and imaging devices for automated ultrasound biometry.

Acknowledgements This work was supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK259909299001-006).

Disclosure of Interests The authors declare no competing interests.

References

1. Bai, J., Tang, Y., Liu, X., Hu, J., Li, Y., Chen, X., et al.: Iugc: A benchmark of landmark detection in end-to-end intrapartum ultrasound biometry. Medical Image Analysis **110**, 103960 (2026). https://doi.org/10.1016/j.media.2026.103960
2. Chen, Y., Zhu, Z., Zhu, S., Qiu, L., Zou, B., Jia, F., Zhu, Y., Zhang, C., Fang, Z., Qin, F., et al.: Sckansformer: Fine-grained classification of bone marrow cells via kansformer backbone and hierarchical attention mechanisms. IEEE Journal of Biomedical and Health Informatics (2024)
3. Coutinho, C.M., Sotiriadis, A., Odibo, A., Khalil, A., D'Antonio, F., Feltovich, H., et al.: Isuog practice guidelines: role of ultrasound in the prediction of spontaneous preterm birth. Ultrasound in Obstetrics & Gynecology **60**(3), 435–456 (2022). https://doi.org/10.1002/uog.26020
4. Deng, B., Chen, Y., Peng, Z.: A two-stage semi-supervised ensemble framework for automated angle of progression measurement in intrapartum ultrasound. In: Bai, J., Huang, Y., Khobo, I., Yaqub, M., Lekadir, K., Ni, D., et al. (eds.) Intrapartum Ultrasound. pp. 88–99. Springer Nature Switzerland, Cham (2026)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2020), https://api.semanticscholar.org/CorpusID:225039882
6. Ghi, T., Eggebø, T., Lees, C., Kalache, K., Rozenberg, P., Youssef, A., et al.: Isuog practice guidelines: intrapartum ultrasound. Ultrasound in Obstetrics & Gynecology **52**(1), 128–139 (2018). https://doi.org/10.1002/uog.19072
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
8. Hu, X., Huang, J., Liu, M., Anmahapong, K., Chen, Y., Luo, Y., et al.: Fetalagents: A multi-agent system for fetal ultrasound image and video analysis. ArXiv **abs/2603.09733** (2026), https://api.semanticscholar.org/CorpusID:286429078
9. Hu, X., Liu, M., Huang, J., Anmahapong, K., Chen, Y., Huang, Y., et al.: Towards reliable fetal ultrasound interpretation with multi-agent collaboration. ArXiv **abs/2605.25357** (2026), https://api.semanticscholar.org/CorpusID:288670430
10. Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., et al.: An intriguing failing of convolutional neural networks and the coordconv solution. Advances in neural information processing systems **31** (2018)
11. Liu, X., Hu, J., Li, Y., Chen, X., Wang, Y.: Noisy student-based self-training enhances landmark detection in intrapartum ultrasound. In: Bai, J., Huang, Y., Khobo, I., Yaqub, M., Lekadir, K., Ni, D., et al. (eds.) Intrapartum Ultrasound. pp. 1–13. Springer Nature Switzerland, Cham (2026)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Ma, C., Li, Y., Guo, B., Jiao, J., Huang, Y., Wang, Y., et al.: Unlabeled data-driven fetal landmark detection in intrapartum ultrasound. In: Bai, J., Huang, Y., Khobo, I., Yaqub, M., Lekadir, K., Ni, D., et al. (eds.) Intrapartum Ultrasound. pp. 14–23. Springer Nature Switzerland, Cham (2026)
14. Mitchell, C.C., Rahko, P.S., Blauwet, L.A., Canaday, B., Finstuen, J.A., Foster, M., et al.: Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: Recommendations from the american society of

echocardiography. Journal of the American Society of Echocardiography : official publication of the American Society of Echocardiography **32 1**, 1–64 (2019), https://api.semanticscholar.org/CorpusID:52917786
15. Nibali, A., He, Z., Morgan, S., Prendergast, L.A.: Numerical coordinate regression with convolutional neural networks. ArXiv **abs/1801.07372** (2018), https://api.semanticscholar.org/CorpusID:8683193
16. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., et al.: DINOv2: Learning Robust Visual Features Without Supervision. Transactions on Machine Learning Research (2024), https://mlanthology.org/tmlr/2024/oquab2024tmlr-dinov2/
17. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., et al.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing **39**(3), 355–368 (1987). https://doi.org/10.1016/S0734-189X(87)80186-X
18. Salomon, L., Alfirevic, Z., Da Silva Costa, F., Deter, R., Figueras, F., Ghi, T., Glanc, P., et al.: Isuog practice guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound in Obstetrics & Gynecology **53**(6), 715–723 (2019). https://doi.org/10.1002/uog.20272
19. 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)
20. Yang, Z., Liu, Q., Hu, Y., Kuang, J., Song, S., Wang, J.: Dsnt-deepunet: A coordinate prediction method for intrapartum ultrasound. In: Bai, J., Huang, Y., Khobo, I., Yaqub, M., Lekadir, K., Ni, D., et al. (eds.) Intrapartum Ultrasound. pp. 33–46. Springer Nature Switzerland, Cham (2026)
21. Zhang, C., Chen, Y., Fan, Z., Huang, Y., Weng, W., Ge, R., Zeng, D., Wang, C.: Tc-diffrecon: texture coordination mri reconstruction method based on diffusion model and modified mf-unet method. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
22. Zhang, X., Chen, X., Yan, H., Tong, L., Du, B.: Pseudo-label enhanced transunet for robust landmark localization in intrapartum ultrasound. In: Bai, J., Huang, Y., Khobo, I., Yaqub, M., Lekadir, K., Ni, D., et al. (eds.) Intrapartum Ultrasound. pp. 77–87. Springer Nature Switzerland, Cham (2026)
23. Zhu, S., Chen, Y., Chen, W., Jiang, S., Zhou, G., Wang, Y., et al.: No modality left behind: Adapting to missing modalities via knowledge distillation for brain tumor segmentation. Medical Image Analysis **112**, 104108 (2026). https://doi.org/10.1016/j.media.2026.104108
24. Zhu, S., Chen, Y., Chen, W., Wang, Y., Liu, C., Jiang, S., et al.: Bridging the Gap in Missing Modalities: Leveraging Knowledge Distillation and Style Matching for Brain Tumor Segmentation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15967. Springer Nature Switzerland (September 2025)