Adversarially Fine-tuned Self-Supervised Framework for Automated Landmark Detection in Intrapartum Ultrasound

Anirvan Krishna[1][0009−0000−3696−8572] and
Zaid Ahmed Khan[2][0009−0005−5718−3260]

[1] Department of Electrical Engineering, Indian Institute of Technology, Kharagpur
anirvankrishna@kgpian.iitkgp.ac.in
[2] Department of Chemical Engineering, Indian Institute of Technology, Kharagpur
ik241168@kgpian.iitkgp.ac.in

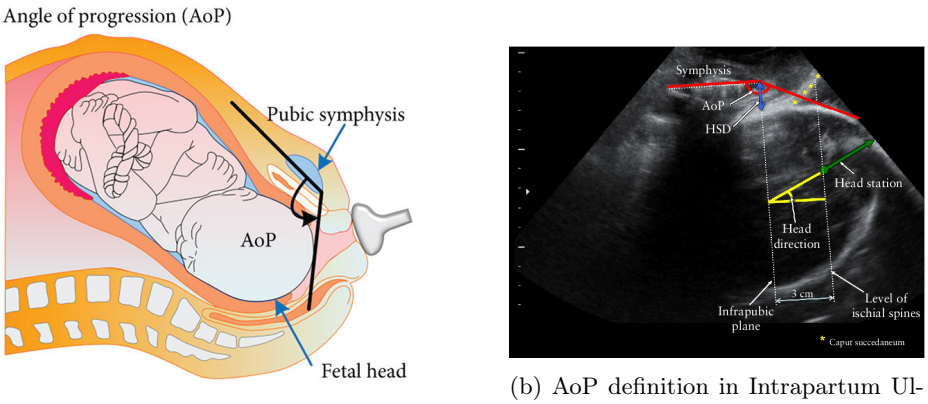
Abstract. Accurate assessment of fetal head progression during labor is essential for guiding timely clinical interventions and improving maternal-fetal outcomes. The World Health Organization's Labour Care Guide emphasizes standardized, evidence-based monitoring tools such as the Angle of Progression (AoP), derived from intrapartum ultrasound. However, current clinical practice relies on manual landmark annotation, which is labor-intensive and subject to variability. To address this limitation, we present a fully automated pipeline for anatomical landmark detection in intrapartum ultrasound as part of the Intrapartum Ultrasound Grand Challenge (IUGC) 2025. Our method combines (i) self-supervised pretraining on unlabeled standard plane ultrasound images to establish strong anatomical priors, (ii) an attention-enhanced decoder architecture for effective spatial localization, and (iii) adversarial fine-tuning using a PatchGAN-style discriminator to ensure anatomical plausibility and spatial precision. The model detects three key landmarks—two on the pubic symphysis and one on the fetal head—enabling robust AoP estimation. Our approach achieves a Mean Radial Error (MRE) of 25.66 pixels and an AoP Mean Absolute Error (MAE) of 8.54 degrees. These results highlight the potential of self-supervised learning and adversarially guided strategies to reduce observer variability, standardize labor monitoring, and support global initiatives for safer, more equitable intrapartum care. Source code is available at `https://github.com/VectorPoint-Analytics/IUGC2025`.

Keywords: Adversarial Learning · Intrapartum Ultrasound · Landmark Detection · Self-Supervised Learning

1 Introduction

Effective intrapartum monitoring is critical for maternal and fetal outcomes. In 2018, the World Health Organization (WHO) issued 56 evidence-based recommendations for labor management, emphasizing timely and respectful care [1].

To implement these, the WHO introduced the Labour Care Guide (LCG) in 2020 to support real-time clinical decision-making [2]. A key element of the LCG is assessing fetal head progression, which influences decisions on operative delivery. Intrapartum ultrasound (US) has emerged as a valuable tool for this task, offering an objective and reproducible alternative to digital examinations. The International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) now recommends ultrasound for labor monitoring [3, 4]. Among quantitative ultrasound metrics, the *Angle of Progression* (AoP)—defined by three anatomical landmarks on standard plane (SP) intrapartum US (Fig. 1), namely the anterior endpoint of the pubic symphysis (PS1), the posterior endpoint of the pubic symphysis (PS2), and the fetal head (FH) [5]—is a widely adopted measure for assessing fetal descent. AoP correlates with delivery outcomes [6, 7], but manual landmark annotation requires expert knowledge and suffers from inter-observer variability.



(a) Diagram of the fetus [8]

(b) AoP definition in Intrapartum Ultrasound [9]

Fig. 1: Description of Angle of Progression (AoP) defined using landmarks in transperineal intrapartum ultrasound on Pubic Symphysis (PS) and Fetal Head (FH)

Deep learning has demonstrated substantial potential in medical image analysis, including fetal biometry, placental assessment, and multi-organ segmentation [10–17]. While numerous studies have used ultrasound for fetal evaluation and anatomical measurements, its application to intrapartum ultrasound remains relatively underexplored. Existing works on Angle of Progression (AoP) estimation [12–17] predominantly rely on segmentation-based pipelines, which demand dense pixel-level annotations. Such annotations are costly, time-consuming, and challenging to scale, thereby limiting their utility in real-world labor ward settings. This gap highlights the need for scalable, label-efficient, and robust methods for AoP measurement.

To address the limitations of segmentation-based AoP estimation, our approach predicts anatomical landmarks directly from intrapartum ultrasound images, eliminating the need for dense manual annotations. This shift greatly reduces the annotation burden, enables more scalable deployment in diverse clinical settings, and maintains the spatial precision necessary for reliable AoP measurement. By leveraging unlabeled data and adopting adversarial learning strategies t enhance generalization in low-data regimes, our method offers a label-efficient and clinically aligned solution that supports standardized, evidence-based intrapartum monitoring in line with WHO's vision for improving maternal and fetal outcomes.

2 Methodology

Our proposed framework for automated landmark detection is composed of two primary stages: (1) self-supervised contrastive pretraining of a high-resolution encoder using the Momentum Contrast v2 (MoCoV2) framework, and (2) supervised fine-tuning of the encoder for heatmap-based landmark localization using adversarial guidance. An overview of the complete methodology is illustrated in Fig. 2, highlighting both the unsupervised representation learning phase and the subsequent supervised adaptation for landmark detection.

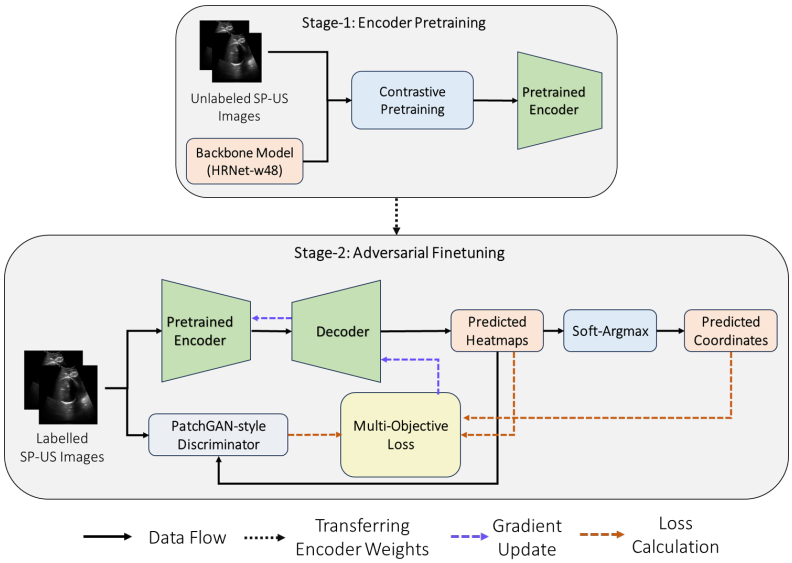


Fig. 2: Overview of the proposed framework. Stage 1: contrastive pretraining with MoCoV2 on unlabeled ultrasound images. Stage 2: supervised fine-tuning of the pretrained encoder for heatmap-based landmark detection with adversarial regularization.

2.1 Contrastive Pretraining for Encoder

We employ the MoCoV2 framework [18] for self-supervised pretraining of a high-resolution encoder on unlabeled standard plane ultrasound images. An overview of the pretraining pipeline is shown in Fig. 3. In our setup, the backbone encoder is the High-Resolution Network (HRNet-W48) [19], chosen for its ability to preserve spatial resolution throughout feature extraction. The encoder described here is instantiated twice within MoCoV2: as a *query encoder* f_q and as a momentum-updated *key encoder* f_k. Both share the same HRNet-W48 architecture and initialization, but differ in their parameter update mechanism: f_q is updated via gradient descent, whereas f_k is updated via momentum tracking from f_q.

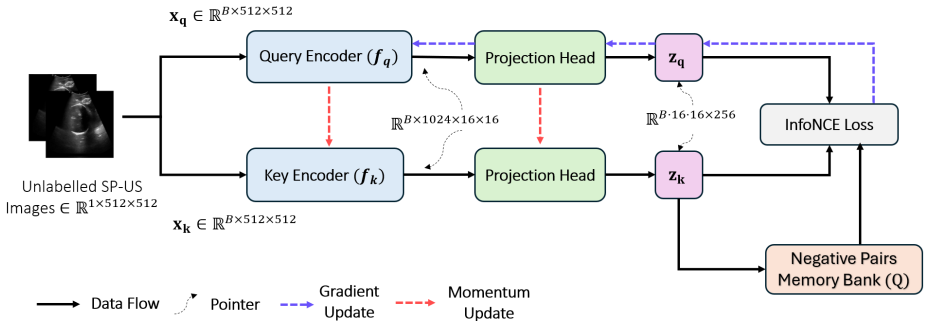


Fig. 3: Overview of the contrastive pretraining stage using MoCoV2. Each unlabeled ultrasound image is augmented into query ($\mathbf{x}_q$) and key ($\mathbf{x}_k$) views, encoded by f_q and a momentum-updated f_k with shared backbone. Feature maps are projected to embeddings ($\mathbf{z}_q$, $\mathbf{z}_k$) via a two-layer head. A queue of negatives with positive query–key pairs defines the InfoNCE loss 4. Only f_q receives gradients, while f_k is updated by momentum for stable training.

To generate positive pairs for contrastive learning, each grayscale ultrasound image $\mathbf{x} \in \mathbb{R}^{1\times512\times512}$ is transformed into two distinct but semantically consistent views, $\mathbf{x}_q$ and $\mathbf{x}_k$, through a stochastic augmentation pipeline (Table 1) designed to emulate variations in ultrasound acquisition while preserving anatomical content. The query view $\mathbf{x}_q$ is fed into f_q, and the key view $\mathbf{x}_k$ into f_k.

Each grayscale ultrasound input $\mathbf{x} \in \mathbb{R}^{1\times512\times512}$ is first processed by either f_q or f_k, yielding a spatial feature map $\mathbf{f} \in \mathbb{R}^{1024\times16\times16}$. This feature map is then passed through a projection head, denoted as `ProjHead(·)`, which is configured as follows:

$$\begin{aligned}\texttt{ProjHead}(\cdot) : \texttt{Conv2D(256c1w0p1s)} &\rightarrow \texttt{ReLU} \\ &\rightarrow \texttt{Conv2D(128c1w0p1s).}\end{aligned} \tag{1}$$

The resulting output is a projected representation

$$\mathbf{z} = \texttt{ProjHead}(\mathbf{f}) \in \mathbb{R}^{128\times16\times16}. \tag{2}$$

Here, the shorthand notation `XcYwZpWs` specifies convolutional layer parameters: `c` denotes the number of output channels (X), `w` the kernel size ($Y\times Y$), `p` the padding size (Z pixels), and `s` the stride length (W). For instance, `256c1w0p1s` indicates a convolutional layer with 256 output channels, a 1×1 kernel, zero padding, and stride equal to 1.

Applying this projection head to the query encoder features produces $\mathbf{z}_q$, while applying it to the key encoder features produces $\mathbf{z}_k$. Each is ℓ_2-normalized across the channel dimension, after which the spatial dimensions are flattened to obtain:

$$\mathbf{z}_{\text{flat}} \in \mathbb{R}^{N\times128}, \quad N = B \cdot 16 \cdot 16, \tag{3}$$

where B is the batch size.

MoCoV2 maintains a first-in-first-out (FIFO) memory bank $\mathcal{Q} \in \mathbb{R}^{128\times K}$ of size $K = 8192$, which stores ℓ_2-normalized key features from previous mini-batches. This provides a large and consistent set of negative examples for the InfoNCE loss [20] defined as:

$$\mathcal{L}_{\text{MoCo}} = -\log \frac{\exp\left(\mathbf{z}_q^\top \mathbf{z}_k/\tau\right)}{\exp\left(\mathbf{z}_q^\top \mathbf{z}_k/\tau\right) + \sum_{i=1}^{K} \exp\left(\mathbf{z}_q^\top \mathbf{z}_i^-/\tau\right)}, \tag{4}$$

where $\tau = 0.2$ is the temperature parameter and $\{\mathbf{z}_i^-\}_{i=1}^K$ are negative feature vectors from $\mathcal{Q}$.

The parameters θ_q and θ_k denote the weights of the query and key encoders, respectively. While θ_q is trainable via backpropagation, θ_k is updated by momentum-based weight tracking:

$$\theta_k \leftarrow m\,\theta_k + (1-m)\,\theta_q, \tag{5}$$

with momentum coefficient $m = 0.999$. This design stabilizes the contrastive objective by preventing rapid drift between the two encoders.

This spatially dense contrastive learning approach enables the encoder to learn rich local anatomical representations without requiring manual labels.

2.2 Heatmap Regression with Adversarial Supervision

Decoder Architecture. After pretraining, the HRNet-W48 encoder is fine-tuned for landmark detection by attaching a decoder head that transforms compact spatial features into high-resolution heatmaps. The decoder adopts a hierarchical design, wherein the encoded representation $\mathbf{f} \in \mathbb{R}^{1024\times16\times16}$ is progressively upsampled and refined through alternating convolutional and attention-enhanced blocks.

The architecture consists of composite attention-enhanced blocks (`ConvCBAM2D`). Specifically, each `ConvCBAM2D` unit comprises a 2D convolutional layer (`Conv2D`),

batch normalization, ReLU activation, and a Convolutional Block Attention Module (CBAM) [21], which jointly refines spatial and channel-wise feature representations. Formally:

$$\begin{aligned}\texttt{ConvCBAM2D}(\cdot) \mapsto \texttt{(1: Conv2D)} \rightarrow \texttt{(2: BatchNorm2D)} \\ \rightarrow \texttt{(3: ReLU)} \rightarrow \texttt{(4: CBAM)}\end{aligned} \tag{6}$$

The complete decoder $\texttt{net}_{\texttt{dec}}(\cdot) : \mathbb{R}^{1024\times16\times16} \mapsto \mathbb{R}^{3\times512\times512}$ is defined as:

$$\begin{aligned}\texttt{net}_{\texttt{dec}}(\cdot) \mapsto\ & \texttt{(1: ConvCBAM2D)512c3w1p1s} \rightarrow \texttt{(2: Upsample)} \\ \rightarrow\ & \texttt{(3: ConvCBAM2D)256c3w1p1s} \rightarrow \texttt{(4: Upsample)} \\ \rightarrow\ & \texttt{(5: ConvCBAM2D)128c3w1p1s} \rightarrow \texttt{(6: Upsample)} \\ \rightarrow\ & \texttt{(7: ConvCBAM2D)64c3w1p1s} \rightarrow \texttt{(8: Upsample)} \\ \rightarrow\ & \texttt{(9: ConvCBAM2D)32c3w1p1s} \rightarrow \texttt{(10: Upsample)} \\ \rightarrow\ & \texttt{(11: Conv2D)3c1w0p0s}\end{aligned} \tag{7}$$

Here, each `Upsample` operation performs bilinear interpolation with scale factor 2 to progressively recover spatial resolution. This attention-guided hierarchical decoding ensures preservation of fine anatomical details, resulting in spatially precise and anatomically consistent heatmaps essential for downstream localization.

The final decoder output $\mathbf{H} \in \mathbb{R}^{3\times512\times512}$ contains a predicted heatmap for each anatomical landmark. Here, the spatial coordinates are denoted by $x \in \{1, \ldots, W\}$ (horizontal axis) and $y \in \{1, \ldots, H\}$ (vertical axis), where W and H correspond to the heatmap width and height, respectively. Landmark coordinates are estimated by applying the differentiable soft-argmax function [22] over each heatmap:

$$\hat{\mathbf{y}}_i = \sum_{x=1}^{W} \sum_{y=1}^{H} \mathbf{H}_i(x, y) \cdot [x, y], \quad \text{for } i = 1, 2, 3, \tag{8}$$

yielding final landmark predictions $\hat{\mathbf{y}} \in \mathbb{R}^{3\times2}$.

Discriminator Architecture. To enforce anatomical plausibility in predicted landmark heatmaps, we incorporated a spectral-normalized PatchGAN discriminator [23], denoted $\mathcal{D}$. It receives a predicted heatmap tensor $\mathbf{H} \in \mathbb{R}^{3\times512\times512}$ and outputs a patch-wise realism score map $\mathbf{s} \in \mathbb{R}^{1\times62\times62}$.

Each convolutional block (`SpectralConv2D`) comprises a spectral-normalized 4×4 convolution followed by a leaky ReLU activation:

$$\begin{aligned}\texttt{SpectralConv2D}(\cdot) \mapsto \texttt{(1: Conv2D)} \rightarrow \texttt{(2: SpectralNorm)} \\ \rightarrow \texttt{(3: LeakyReLU)}\end{aligned} \tag{9}$$

The complete discriminator $\mathcal{D}(\cdot) : \mathbb{R}^{3\times512\times512} \mapsto \mathbb{R}^{1\times62\times62}$ is defined as:

$$\begin{aligned}\mathcal{D}(\cdot) \mapsto &\texttt{(1: SpectralConv2D)64c4w2p1s} \rightarrow \texttt{(2: SpectralConv2D)128c4w2p1s} \\ &\rightarrow \texttt{(3: SpectralConv2D)256c4w2p1s} \rightarrow \texttt{(4: SpectralConv2D)512c4w1p1s} \\ &\rightarrow \texttt{(5: Conv2D)1c4w1p1s}\end{aligned} \tag{10}$$

Spectral normalization is applied to layers (1)–(4) only, stabilizing adversarial training as proposed by Miyato et al. [24].

The adversarial component of the training objective is based on the least squares GAN (LSGAN) formulation [25], which stabilizes training and encourages sharper outputs. The discriminator is trained to minimize:

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2}\mathbb{E}_{\mathbf{H}_{\text{real}}}\left[(\mathcal{D}(\mathbf{H}_{\text{real}}) - 1)^2\right] + \frac{1}{2}\mathbb{E}_{\mathbf{H}_{\text{fake}}}\left[(\mathcal{D}(\mathbf{H}_{\text{fake}}))^2\right], \tag{11}$$

where $\mathbf{H}_{\text{real}}$ and $\mathbf{H}_{\text{fake}}$ denote the ground truth and predicted heatmaps respectively. The generator (i.e., the landmark detection network) is trained with the adversarial loss:

$$\mathcal{L}_{\text{GAN}} = \frac{1}{2}\mathbb{E}_{\mathbf{H}_{\text{fake}}}\left[(\mathcal{D}(\mathbf{H}_{\text{fake}}) - 1)^2\right]. \tag{12}$$

This architectural design — integrating spectral normalization, CBAM-based attention, and a PatchGAN structure — enables the discriminator to capture subtle anatomical inconsistencies, thereby providing fine-grained adversarial feedback that improves both the accuracy and realism of the predicted landmark heatmaps.

The proposed model is trained under a multi-objective framework, where the overall loss is a weighted sum of four complementary components: (1) a heatmap regression loss $\mathcal{L}_{\text{heatmap}}$, (2) a coordinate regression loss $\mathcal{L}_{\text{coord}}$, (3) an adversarial loss $\mathcal{L}_{\text{GAN}}$, and (4) an entropy penalty $\mathcal{L}_{\text{entropy}}$.

Heatmap Regression Loss. The heatmap regression term enforces spatial alignment between predicted heatmaps $\mathbf{H} \in \mathbb{R}^{K\times H\times W}$ and Gaussian ground-truth heatmaps $\mathbf{H}^* \in \mathbb{R}^{K\times H\times W}$ via mean squared error (MSE):

$$\mathcal{L}_{\text{heatmap}} = \frac{1}{KHW}\sum_{i=1}^{K}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(\mathbf{H}_i(x,y) - \mathbf{H}_i^*(x,y)\right)^2. \tag{13}$$

Where $\mathbf{H}_i$ and $\mathbf{H}_i^*$ are the ground truth and predicted heatmaps respectively. This loss guides the network to learn pixel-wise probability distributions centered on the correct landmark locations. By regressing to smooth Gaussian targets rather than binary masks, the model benefits from stable gradients and improved robustness to small spatial deviations.

Coordinate Regression Loss. While heatmap supervision captures spatial context, it may not fully penalize small but clinically significant displacements of

predicted peaks. Therefore, we introduce a coordinate-level MSE loss that directly measures Euclidean distance between predicted coordinates $\hat{\mathbf{y}}_i$ (obtained as shown in Eq. 8) and their ground-truth $\mathbf{y}_i^*$:

$$\mathcal{L}_{\text{coord}} = \frac{1}{K} \sum_{i=1}^{K} \|\hat{\mathbf{y}}_i - \mathbf{y}_i^*\|_2^2 . \tag{14}$$

This complementary term enforces precise localization at the point level, mitigating cases where heatmaps are visually plausible yet slightly shifted, which could impact clinical measurements.

Adversarial Loss. To ensure anatomical plausibility of predicted heatmaps, we integrate an adversarial loss based on the least-squares GAN (LSGAN) formulation as mentioned in Eq. 12 This term constrains the spatial configuration of landmarks to resemble anatomically valid patterns observed in real data, reducing the risk of unrealistic arrangements even when pixel- or coordinate-level losses are minimized.

Entropy Penalty. Ambiguous or overly diffuse heatmaps hinder reliable landmark extraction. To promote confident and unimodal predictions, we compute the Shannon entropy [26] over the spatial softmax of each heatmap:

$$\mathbf{P}_i(x, y) = \frac{\exp(\mathbf{H}_i(x, y))}{\sum_{x', y'} \exp(\mathbf{H}_i(x', y'))}, \tag{15}$$

$$\mathcal{L}_{\text{entropy}} = -\frac{1}{K} \sum_{i=1}^{K} \sum_{x=1}^{W} \sum_{y=1}^{H} \mathbf{P}_i(x, y) \cdot \log \mathbf{P}_i(x, y). \tag{16}$$

Minimizing this term encourages sharp probability peaks at landmark locations while suppressing irrelevant responses, thereby improving confidence and reproducibility in landmark detection.

Each component addresses a distinct aspect of the landmark detection task, thereby ensuring that the learned representations are spatially accurate, anatomically consistent, and confidently localized. The total objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{h}} \cdot \mathcal{L}_{\text{heatmap}} + \lambda_{\text{c}} \cdot \mathcal{L}_{\text{coord}} + \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{GAN}} + \lambda_{\text{e}} \cdot \mathcal{L}_{\text{entropy}} \tag{17}$$

where $\lambda_{\text{h}}, \lambda_{\text{c}}, \lambda_{\text{adv}}$, and λ_{e} are scalar hyperparameters controlling the contribution of each term.

3 Experiments

3.1 Dataset Description

We employed the official dataset from the Intrapartum Ultrasound Grand Challenge (IUGC) 2025 [27], which is designed to support research on automated Angle of Progression (AoP) estimation from Transperineal Intrapartum Ultrasound

Table 1: Data augmentation parameters and application probabilities

Augmentation	**Parameters / Range**	**Probability**
Elastic Transform	$\alpha = 5$, $\sigma = 10$, $\alpha_{\text{affine}} = 10$	0.8
Random Brightness/Contrast	brightness limit: 0.2, contrast limit: 0.2	0.5
Gaussian Blur	Kernel size $\in [3, 5]$	0.5
Speckle Noise	$\mu = 0.0$, $\sigma = 0.01$	0.5
Normalization	$\mu = 0.0$, $\sigma = 255.0$	1.0

(TUS) images. The dataset consists of 31,421 training images, 100 validation images, and 501 test images. Among the training images, 300 are fully annotated standard-plane images containing landmark coordinates and AoP measurements, while the remainder are unlabeled. Within the unlabeled set, 2045 images are identified as standard-plane images. All images are stored in RGB format with a spatial resolution of 512×512 pixels. The validation and test sets are withheld from participants, with performance evaluation carried out exclusively via the challenge server.

For the self-supervised pretraining stage, we used the 2045 unlabeled standard-plane images to train the encoder backbone with the aim of capturing domain-specific features of TUS, in accordance with the ISUOG guidelines [28] which state that accurate AoP measurement requires simultaneous visualization of the longitudinal sagittal plane of the pubic symphysis and the fetal head. To enhance robustness to variations in acquisition conditions and to simulate realistic ultrasound noise patterns, each image underwent a series of spatial and intensity-based augmentations, carefully picked to mimic the natural variations in ultrasound images. Spatial augmentation included elastic deformation with parameters $\alpha = 5$, $\sigma = 10$, and $\alpha_{\text{affine}} = 10$. Intensity-based transformations included random brightness/contrast adjustment, Gaussian blurring with kernel sizes in the range $[3, 5]$, and multiplicative Gaussian speckle noise with mean 0.0 and standard deviation 0.01. Augmentations were applied with probabilities as detailed in Table 1. Following augmentation, images were normalized to zero mean and unit variance with respect to the original intensity range (mean = 0.0, standard deviation = 255.0) and converted into tensors.

In the supervised fine-tuning stage, we used the 300 labeled standard-plane images, each containing annotations for three anatomical landmarks: the distal edge of the pubic symphysis (PS1), the proximal edge of the pubic symphysis (PS2), and the most distal point of the fetal head (FH). Each image was also accompanied by a scalar AoP value. For heatmap-based regression, each landmark coordinate (x_i, y_i) was converted into a Gaussian heatmap centered at its location, with a fixed standard deviation $\sigma = 5$ pixels. The ground-truth heatmap for the i-th landmark is defined as

$$H_i(x, y) = \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right), \tag{18}$$

where (x, y) denotes a pixel coordinate in the image plane. The same augmentations listed in Table 1 were applied to both the images and the corresponding landmark coordinates to ensure spatial consistency between the inputs and the labels. This approach promotes generalization to unseen ultrasound acquisitions while preserving clinically relevant spatial relationships, which might be lost in the deeper layers of the network in an end-to-end regression pipeline.

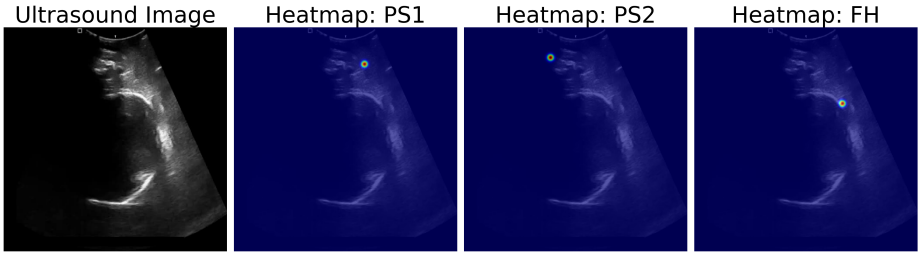


Fig. 4: Representation of the Gaussian heatmaps generated corresponding to the three landmarks present in standard plane intrapartum ultrasound image

3.2 Training Parameters

Self-supervised pretraining was performed for 25 epochs on unlabeled standard-plane ultrasound images with a batch size of 4. Feature projections were 128-dimensional, using a memory queue of 8,192, momentum coefficient 0.999, and temperature 0.2. Adam optimization [29] applied stage-wise learning rates to the HRNet-W48 backbone: 1×10^{-4} for the deepest stage, 0.5×10^{-4} for the intermediate stage, and 0.1×10^{-4} for the shallowest stage, with 1×10^{-3} for the projection head. Weight decay was 1×10^{-6}, and cosine annealing scheduling was applied over the total pretraining epochs. Mixed-precision training was employed.

Supervised fine-tuning was conducted for 400 epochs with a batch size of 4. AdamW optimization [30] was used for the generator (3×10^{-4}) and discriminator (5×10^{-5}), each with cosine annealing and weight decay 1×10^{-4}. The encoder was frozen initially and unfrozen in stages at epochs 150 and 250. The multi-objective loss (Eq. 17) is defined with weights $\lambda_h = 1$, $\lambda_c = 10$, $\lambda_{\text{adv}} = 0.01$, and $\lambda_e = 0.001$. The choice of $\lambda_c \gg \lambda_h$ reflects the emphasis on precise landmark localization, while the small λ_{adv} stabilizes adversarial training without overpowering regression terms.

3.3 Evaluation

We systematically evaluated a range of encoder–decoder architectures, pretraining strategies, attention modules, and output representations for automated

landmark detection and AoP estimation. Performance was assessed on the held-out test set using Mean Radial Error (MRE, in pixels) and AoP Mean Absolute Error (AOP MAE, in degrees) as evaluation metrics.

Mean Radial Error (MRE). The Mean Radial Error (MRE) is computed as the average Euclidean distance between the ground truth and predicted landmark coordinates:

$$\text{MRE} = \frac{1}{3}\sum_{i=1}^{3}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2}. \tag{19}$$

This metric directly quantifies the pixel-level localization accuracy of the predicted landmarks. Where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ denote the ground-truth and predicted coordinates for the ith landmark.

AoP Mean Absolute Error (AoP MAE). The Angle of Progression (AoP) is geometrically derived from the three landmark coordinates (two endpoints of the pubic symphysis and the fetal head point). Let $\theta(\mathbf{y}), \theta(\hat{\mathbf{y}})$ denote the AoP values computed from the ground truth and predicted landmarks, respectively. The AoP Mean Absolute Error is then defined as

$$\text{AoP MAE} = \big|\big|\, \theta(\mathbf{y}) - \theta(\hat{\mathbf{y}}) \,\big|\big|. \tag{20}$$

This metric evaluates the clinical reliability of the method by measuring the angular discrepancy in degrees between the ground truth and predicted AoP.

4 Results and Discussion

The quantitative results for all encoder–decoder configurations are presented in Table 2. We systematically explored multiple combinations of (i) encoder architectures (HRNet-w48, ResNet-50 [31], and Attention UNet [32]), (ii) fully supervised vs self-supervised pretraining strategies (MoCoV2 and Masked Autoencoding), (iii) decoder designs with or without Convolutional Block Attention Module (CBAM), (iv) the use of adversarial regularization, and (v) output formats (heatmap regression versus coordinate regression). All models were subsequently fine-tuned on our dataset for landmark localization. Fig. 5 illustrates the performance of the best model on representative training samples.

Baseline performance. A fully supervised HRNet-w48 trained directly on the landmark detection task yielded an MRE of 43.88 px and an AoP MAE of 11.24°, reflecting the difficulty of the problem given the limited training data and the complex echogenic patterns in intrapartum ultrasound.

Effect of self-supervised pretraining and attention. Replacing the randomly initialized encoder with a MoCoV2-pretrained HRNet-w48 and attaching a CBAM-enhanced decoder significantly improved localization accuracy (MRE

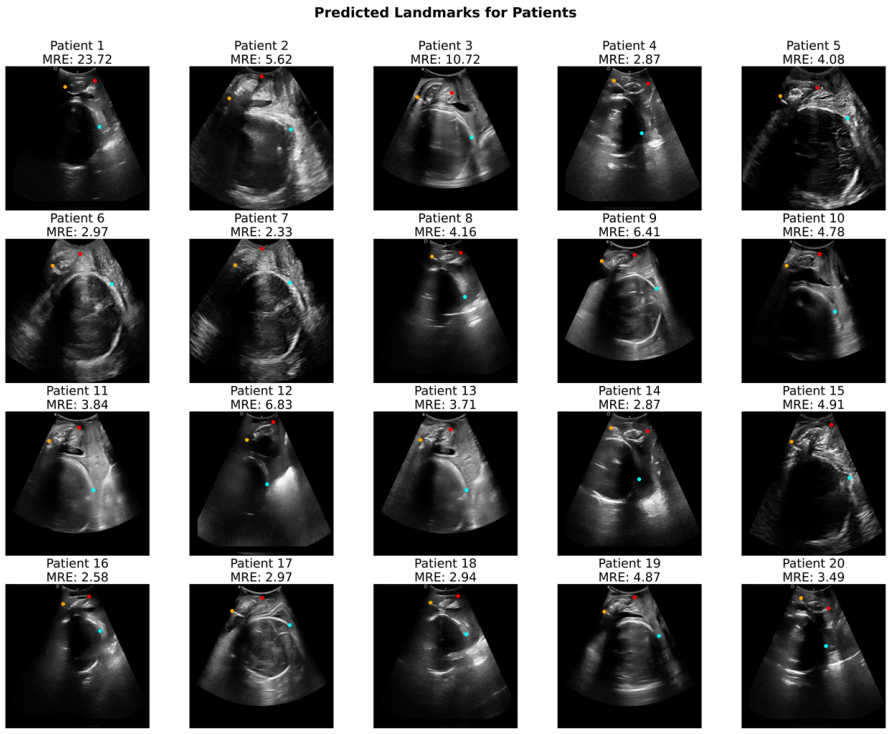


Fig. 5: Representation of landmark detection performance on 20 US image samples from the training set. The Mean Radial Error (MRE) is indicated for each of these images.

34.83 px), indicating that self-supervised pretraining effectively transfers domain-relevant spatial features and that CBAM helps focus on discriminative anatomical cues. Removing CBAM resulted in degraded accuracy (MRE 35.46 px), confirming the contribution of attention mechanisms.

Role of adversarial learning. Introducing a PatchGAN discriminator into the MoCoV2 + CBAM pipeline further reduced the error to 25.66 px (AoP MAE 8.54°), the best performance across all settings. This suggests that adversarial regularization encourages more anatomically consistent heatmaps, improving both localization and downstream AoP estimation.

Coordinate regression vs. heatmap regression. To assess the relative merits of coordinate prediction, we evaluated the same MoCoV2-pretrained HRNet-w48 + CBAM backbone with a coordinate regression head. This achieved an MRE of 29.64 px and AoP MAE of 10.68°, outperforming the fully supervised baseline and the heatmap-based counterpart.

Table 2: Performance of different encoder–decoder configurations, pretraining, attention, adversarial guidance, and output formats for landmark localization. Here, Attn denotes the presence of attention in decoder block, Adv. denotes the presence of adversarial guidance and AE stands for Autoencoder. The best scores are in **bold** and the second best are underlined. Downward arrow (↓) indicates lower values are better.

Encoder	Pretrain	Attn	Adv.	Output	MRE (↓)	AoP MAE (↓)
HRNet-w48	–	–	–	Heatmaps	43.88	11.24
HRNet-w48	MoCoV2	✓	✓	Heatmaps	**25.66**	**8.54**
HRNet-w48	MoCoV2	✓	–	Heatmaps	34.83	12.26
HRNet-w48	MoCoV2	–	–	Heatmaps	35.46	13.85
ResNet-50	MoCoV2	–	–	Heatmaps	30.40	19.29
HRNet-w48	MoCoV2	✓	–	Coordinates	29.64	10.68
Attention UNet	Masked AE	✓	–	Coordinates	31.81	12.14

These findings justify our *hybrid training objective*: combining heatmap regression with a differentiable coordinate extraction (soft-argmax) and an explicit coordinate MRE term. Heatmaps provide dense spatial supervision, capture uncertainty, and preserve contextual structure, while the coordinate term enforces geometric precision. The synergy of both was critical to achieving the optimal balance observed in our best-performing configuration in the adversarial setting.

Backbone variation. Substituting the HRNet-w48 encoder with a ResNet-50 [31] (MoCoV2-pretrained) degraded AoP accuracy (MRE 30.40 px, AoP MAE 19.29°), highlighting the advantage of HRNet's high-resolution feature representations for fine-grained anatomical localization.

Masked autoencoder pretraining. A Masked Autoencoder [33] (MAE)-pretrained Attention U-Net [32] with a coordinate regression head achieved an MRE of 31.81 px and AoP MAE of 12.14°, outperforming the fully supervised baseline but not matching the MoCoV2-pretrained HRNet configurations. This suggests that instance-discrimination-based self-supervised learning may transfer more directly useful features for this task than masked image modeling in its current form.

5 Conclusion

In this work, we introduced a robust pipeline for automated detection of key anatomical landmarks—PS1, PS2, and FH—in intrapartum transperineal ultrasound images, enabling precise estimation of the Angle of Progression (AoP). Our approach integrates carefully designed models and training strategies to ensure strong generalization across diverse imaging conditions. By reducing reliance on manual measurement, the proposed method holds potential to enhance clinical efficiency, support objective decision-making during labor, and improve

maternal–fetal outcomes. This work lays the foundation for future advancements in AI-driven intrapartum ultrasound analysis.

References

1. World Health Organization. Who recommendations: Intrapartum care for a positive childbirth experience, 2018.
2. World Health Organization. Labour care guide: User manual, 2020.
3. T. Ghi, T. M. Eggebø, C. Lees, K. Kalache, P. Rozenberg, A. Youssef, B. Tutschek, and G. H. A. Visser. Intrapartum ultrasound: A review of current practice and challenges. *Ultrasound in Obstetrics & Gynecology*, 52(4):437–444, 2018.
4. T. Ghi, T. M. Eggebø, C. Lees, A. M. Marconi, A. Youssef, B. Tutschek, K. Kalache, I. Mappa, G. Rizzo, G. H. A. Visser, and et al. Isuog practice guidelines: intrapartum ultrasound. *Ultrasound in Obstetrics & Gynecology*, 44(4):464–473, 2014.
5. A. F. Barbera, X. Pombar, G. Perugino, D. C. Lezotte, and J. C. Hobbins. A new method to assess fetal head descent in labor with transperineal ultrasound. *Ultrasound in Obstetrics & Gynecology*, 33(3):313–319, 2009.
6. T. Ghi, A. Farina, A. Pedrazzi, M. Perugini, L. Savelli, B. Viscolani, and G. Pilu. Sonographic pattern of fetal head descent: relationship with duration of active second stage of labor and mode of delivery. *Ultrasound in Obstetrics & Gynecology*, 44(1):82–89, 2014.
7. A. F. Barbera, X. Pombar, J. C. Hobbins, J. C. Melchor, C. Roque, and F. D'Antonio. Angle of progression: a valuable ultrasound parameter to predict operative delivery. *American Journal of Obstetrics and Gynecology*, 213(2):229.e1–229.e5, 2015.
8. Yaosheng Lu, Dengjiang Zhi, Minghong Zhou, Fan Lai, Gaowen Chen, Zhanhong Ou, Rongdan Zeng, Shun Long, Ruiyu Qiu, Mengqiang Zhou, Xiaosong Jiang, Huijin Wang, and Jieyun Bai. Multitask deep neural network for the fully automatic measurement of the angle of progression. *Computational and Mathematical Methods in Medicine*, 2022:1–14, 09 2022.
9. Tullio Ghi, Torbjorn Eggebo, C. Lees, Karim Kalache, P. Rozenberg, Aly Youssef, L. Salomon, and Boris Tutschek. Isuog practice guidelines: intrapartum ultrasound. *Ultrasound in Obstetrics Gynecology*, 52:128–139, 07 2018.
10. R. S. Barros, M. O'Connor, and J. A. Noble. Towards real-time angle of progression estimation from transperineal ultrasound using deep learning. In *Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4, 2020.
11. J. J. Cerrolaza, A. Mukhopadhyay, N. Cerrolaza, M. P. Heinrich, and J. A. Noble. Weakly supervised learning of the angle of progression from transperineal ultrasound. *IEEE Transactions on Medical Imaging*, 41(4):802–814, 2022.
12. Y. Zhu, P. Singh, S. Vasudevan, H. Esfandiari, and J. A. Noble. Attention-based deep learning for landmark detection in intrapartum ultrasound. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14225 of *Lecture Notes in Computer Science*, pages 457–466. Springer, 2023.
13. Long S Campello VM Bai J Lekadir K. Chen Z, Lu Y. Fetal head and pubic symphysis segmentation in intrapartum ultrasound image using a dual-path boundary-guided residual network. *IEEE Journal of Biomedical and Health Informatics*, 28(8):4648–4659, 2024.

14. Zhou M Lai F Chen G Ou Z Zeng R Long S Qiu R Zhou M Jiang X Wang H Bai J. Lu Y, Zhi D. Multitask deep neural network for the fully automatic measurement of the angle of progression. *Computational and Mathematical Methods in Medicine*, 2022:5192338, 2022.
15. Yu S Lu Y Long S Wang H Qiu R Ou Z Zhou M Zhi D Zhou M Jiang X Chen G. A Bai J, Sun Z. A framework for computing angle of progression from transperineal ultrasound images for evaluating fetal head descent using a novel double branch network. *Frontiers in Physiology*, 13:940150, 2022.
16. Pengzhou Cai, Lu Jiang, Yanxin Li, and Libin Lan. Pubic symphysis-fetal head segmentation using pure transformer with bi-level routing attention, 2024.
17. Zihao Zhou, Yaosheng Lu, Jieyun Bai, Víctor M. Campello, Fan Feng, and Karim Lekadir. Segment anything model for fetal head-pubic symphysis segmentation in intrapartum ultrasound image analysis. *Expert Systems with Applications*, 263:125699, 2025.
18. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.
19. Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5406–5415, 2019.
20. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.
21. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.
22. Diogo C Luvizon, David Picard, and Hedi Tabia. Human pose regression by combining indirect part detection and contextual information. In *Computer Vision and Image Understanding*, volume 192, page 102897. Elsevier, 2020.
23. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.
24. Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.
25. Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017.
26. Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
27. Jieyun Bai, Isaac Khobo, Yaosheng Lu, Dong Ni, Mohammad Yaqub, Karim Lekadir, Jun Ma, and Shuo Li. Landmark detection challenge for intrapartum ultrasound measurement: Meeting the actual clinical assessment of labor progress. Dataset, Version v1, Zenodo, March 2025.
28. T. Ghi, T. Eggebø, C. Lees, K. Kalache, P. Rozenberg, A. Youssef, L. J. Salomon, and B. Tutschek. ISUOG Practice Guidelines: intrapartum ultrasound. *Ultrasound in Obstetrics & Gynecology*, 52(1):128–139, 2018.
29. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

30. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
31. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
32. Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
33. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.