High-Resolution DINOv3-DSNT Landmark Regression for Multi-Organ Ultrasound Biometry

Jiongpei Hou[1†], Meng Li[1†], Yushan Xing[1], and Hong Peng[1*]

Lanzhou University, Lanzhou, China
pengh@lzu.edu.cn

Abstract. Landmark-based ultrasound biometry reduces a broad spectrum of clinical measurements to a small set of anatomically defined points, yet the resulting learning problem is highly heterogeneous: the same coordinate representation is used for fetal, intrapartum, vascular, and echocardiographic structures whose image appearance, spatial scale, and point semantics differ substantially. This work investigates a high-resolution DINOv3-DSNT formulation for multi-organ ultrasound landmark regression. A DINOv3 ViT-base encoder provides dense self-supervised visual features, while a compact multi-scale decoder produces task-specific landmark heatmaps. Coordinates are obtained through a differentiable spatial-to-numerical transform (DSNT), allowing heatmap supervision and continuous coordinate regression to be optimized in a single end-to-end model. Rather than imposing one shared landmark vocabulary across unrelated anatomical measurements, each task retains its own output head, task-specific high-resolution training scale, and checkpoint. Internal validation across nine ultrasound biometry tasks yields an average mean radial error of 22.83 pixels, an average mean absolute coordinate error of 14.03 pixels, and an average diagonal-normalized radial error of 2.20% in the original image coordinate system. Per-task performance varies across anatomical views, with lower errors on geometrically stable tasks such as AOP and FUGC and higher errors on cardiac views such as A4C and PSAX, indicating that view regularity, sample size, and landmark ambiguity strongly affect localization accuracy. The findings indicate that dense self-supervised representations, when coupled with differentiable heatmap regression, offer a coherent technical route for high-resolution ultrasound biometric localization across diverse organs and measurement geometries.

Keywords: Ultrasound biometry · Landmark detection · DINOv3 · DSNT · Coordinate regression

1 Introduction

Quantitative ultrasound measurement underpins many decisions in prenatal care, labor monitoring, vascular assessment, and echocardiography. In these settings, clinicians frequently reduce a complex acoustic scene to a small number of

† These authors contributed equally to this work.
* Corresponding author.

landmarks and compute a distance, diameter, angle, or circumference from their geometry. The resulting measurements are compact, interpretable, and longitudinally comparable. Their reliability, however, remains constrained by acoustic shadowing, speckle, anatomical ambiguity, operator experience, and the fact that a displacement of only a few pixels may alter the derived measurement [9, 6, 1].

The breadth of automated ultrasound biometry has expanded rapidly. Deep learning systems have been reported for fetal head circumference, fetal standard-plane biometry, femur length estimation, intrapartum angle measurement, inferior vena cava assessment, and echocardiographic quantification [15, 11, 10, 8, 20, 5, 19, 18]. Many of these methods are optimized for a single anatomical target. Extending landmark regression to multiple organs introduces a less explored form of heterogeneity: the number of points changes, the spatial scale changes, and the meaning of a point index is determined by the measurement protocol rather than by a universal coordinate label. A fetal head endpoint, an intrapartum pubic symphysis point, and a cardiac chamber boundary point are all two-dimensional coordinates, but they are not semantically exchangeable.

This study examines whether a common high-resolution architecture can be reused across such heterogeneous landmark tasks while preserving task-specific point identity. The proposed solution combines DINOv3 dense visual features with DSNT-based heatmap regression. DINO-style pretraining has demonstrated strong transfer for spatial prediction because it yields robust patch-level representations [14, 16]. DSNT, in turn, preserves a heatmap representation while returning continuous coordinates without a non-differentiable argmax [13]. Related ultrasound foundation-model studies have also emphasized the value of unified representations and Transformer-based modules for high-precision ultrasound analysis [17, 12].

The resulting framework deliberately separates architectural reuse from semantic sharing. Our goal is not to learn a single universal landmark vocabulary shared across all organs and measurement protocols. The number of landmarks varies from two-point biometric measurements to dense cardiac landmark configurations, and the semantic meaning of each point index is defined by the corresponding clinical measurement protocol. Forcing these heterogeneous tasks into one shared prediction head would require artificial padding or remapping and could mix anatomically unrelated landmark definitions. Therefore, the encoder-decoder design, optimization objective, and validation procedure are reused across measurements, whereas output dimensionality, input resolution, and checkpoint selection remain task-specific. This design should be interpreted as a reusable architecture recipe for task-specific ultrasound biometry rather than a fully unified multi-task model. It allows a common DINOv3-DSNT formulation to accommodate nine ultrasound measurement tasks without collapsing their anatomical definitions into one shared landmark space.

2 Related Work

Ultrasound biometric measurement. Automated biometry seeks to reduce repeated manual interaction while retaining the clinical interpretability of geometric measurements. In fetal ultrasound, automatic systems have been proposed for head detection and circumference estimation [10], direct fetal biometric prediction [8], video-based measurement [15], fetal head biometry [11], and fetal femur length [20]. In intrapartum ultrasound, automatic angle of progression and head-perineum distance measurement have been studied as a way to provide objective labor-progress assessment [6, 1]. In cardiopulmonary ultrasound, automated echocardiographic quantification and inferior vena cava measurement demonstrate that landmark-style geometric quantification remains important outside obstetric imaging [19, 18, 5].

Recent challenge benchmarks have further highlighted the importance of robust ultrasound biometry across heterogeneous anatomical targets and clinical protocols. The Foundation Model Challenge for Ultrasound Biometry (FoundUS) emphasizes multi-organ and multi-protocol landmark localization for ultrasound biometric measurement [3]. Intrapartum ultrasound challenges and related benchmarks also show that fetal head, pubic symphysis, and angle-of-progression measurements remain difficult because landmark definitions are affected by view quality, acoustic shadowing, and clinical measurement conventions [2, 4]. These benchmarks motivate reusable technical frameworks that can preserve task-specific landmark semantics while sharing common visual representations.

Dense visual representations. Vision Transformers model images as patch sequences and have become a flexible substrate for dense visual prediction [7]. Self-supervised methods such as DINOv2 and DINOv3 further improve transfer by learning patch-level representations without task-specific labels [14, 16]. This property is well aligned with ultrasound, where annotated measurements are expensive and appearance can vary with scanner, probe angle, acoustic window, and patient anatomy. The present framework therefore treats DINOv3 as a general visual representation and learns only the lightweight components required for biometric localization.

Heatmap coordinate regression. Coordinate regression can be implemented either as direct numerical regression or as heatmap localization. Direct regression is parameter efficient but offers limited spatial interpretability. Heatmap argmax exposes the spatial response but introduces discretization and blocks gradient flow through the coordinate operation. DSNT represents coordinates as expectations over probability-normalized heatmaps, retaining spatial structure while producing continuous coordinates [13]. For ultrasound biometry, this combination is especially relevant because the clinically reported value is often a geometric function of a small number of points.

3 Method

3.1 Overview

The framework is summarized in Fig. 1. Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an ultrasound image and let $\mathbf{p} = \{(x_k, y_k)\}_{k=1}^{K}$ denote the annotated landmarks for one measurement task. The value of K is determined by the measurement protocol, ranging from two-point biometry tasks to multi-point cardiac views. Each task therefore keeps its own output dimensionality and training checkpoint, rather than being forced into a shared anatomical landmark vocabulary.

For all tasks, a resized image batch is passed to a DINOv3 ViT-base encoder. Features from blocks 2, 5, 8, and 11 capture texture, anatomical structure, and global layout. These multi-depth features are aligned on a stride-4 grid, fused with a shallow convolutional branch, and decoded into K heatmaps. DSNT then converts the normalized heatmaps into continuous landmark coordinates, which are mapped back to the original image space for evaluation. This organization keeps the architecture and validation protocol common across ultrasound domains while preserving task-specific point definitions.

This design should be interpreted as task-specific deployment of a reusable architecture rather than a fully unified multi-task model. The shared component is the representation-learning and heatmap-to-coordinate regression principle, whereas the anatomical coordinate system remains task-specific. This distinction is important because point indices are not semantically exchangeable across organs: a fetal femur endpoint, an intrapartum pubic-symphysis landmark, and a cardiac chamber boundary point are all two-dimensional coordinates but correspond to different clinical measurement definitions.

3.2 Task-Specific Resolution Selection

Mixed image resolution is a central practical issue in ultrasound biometry. Uniform downsampling to a small square input removes the fine boundary evidence that defines many landmarks, whereas training every task at the largest native frame size is computationally wasteful and unstable for small datasets. A fixed training size is therefore selected separately for each task. The target area is bounded by a maximum pixel budget, the aspect ratio is estimated from the median task aspect ratio, and both dimensions are rounded to multiples of 16. This preserves the ViT patch-grid alignment while retaining as much anatomical detail as possible.

Task-specific preprocessing also protects point identity. Several measurements are represented by pairs of endpoints, but those endpoints are not necessarily exchangeable. Vertical and horizontal diameters may encode anatomical directions, and cardiac landmarks correspond to named structures. The original annotation order is therefore preserved throughout resizing, augmentation, and metric computation.

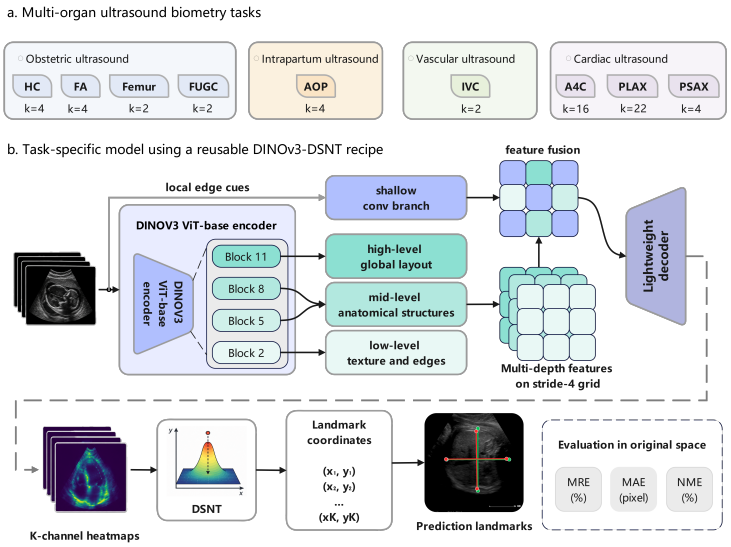

Fig. 1. Overview of the proposed multi-organ ultrasound landmark regression framework. Task-specific landmark heads are coupled with a shared DINOv3-DSNT architecture that predicts K heatmaps and converts them to continuous coordinates in the original image space.

3.3 DINOv3-DSNT Network

The encoder is a DINOv3 ViT-base model. Intermediate features are extracted from four transformer depths, corresponding to blocks 2, 5, 8, and 11. The selected token features are reshaped into spatial feature maps, and each feature map is projected to a common decoder dimension C_d using a 1×1 convolution followed by batch normalization and ReLU. The projected feature maps are then bilinearly resized to a common stride-4 grid. In parallel, a shallow two-layer convolutional branch extracts local edge cues from the input image and is aligned to the same spatial grid. The four DINOv3 feature maps and the shallow convolutional feature map are concatenated along the channel dimension, refined by convolutional fusion blocks, and finally passed to a task-specific 1×1 heatmap head with K output channels.

For landmark k, let $Z_k(u, v)$ be the heatmap logit at grid location (u, v). A spatial softmax produces a probability map:

$$H_k(u, v) = \frac{\exp\left(Z_k(u, v)/\tau\right)}{\sum_{u'=1}^{W_h} \sum_{v'=1}^{H_h} \exp\left(Z_k(u', v')/\tau\right)}. \tag{1}$$

DSNT then computes the expected coordinate on a normalized grid:

$$\hat{x}_k = \sum_{u,v} H_k(u, v) x_u, \qquad \hat{y}_k = \sum_{u,v} H_k(u, v) y_v, \tag{2}$$

Table 1. Decoder design used in the DINOv3-DSNT landmark regression framework. All task-specific models use the same decoder recipe, while the final heatmap head changes according to the number of landmarks K.

Component	Setting
DINOv3 feature layers	Transformer blocks 2, 5, 8, and 11
Feature projection	1×1 convolution + batch normalization + ReLU
Common spatial grid	Bilinear alignment to stride-4 resolution
Local branch	Two-layer shallow convolutional branch
Feature fusion	Channel concatenation followed by convolutional refinement
Prediction head	Task-specific 1×1 convolution with K heatmaps
Coordinate transform	DSNT expectation over normalized heatmaps

where $x_u, y_v \in [-1, 1]$. The heatmap remains a spatial confidence distribution, and the final landmark is no longer restricted to a discrete heatmap pixel.

3.4 Optimization Objective

Ground-truth coordinates are normalized to $[0, 1]^2$ using the original image width and height, and then mapped to the DSNT range $[-1, 1]^2$. The loss combines coordinate supervision and heatmap-shape supervision:

$$\mathcal{L} = \lambda_c \, \mathrm{SmoothL1}(\hat{\mathbf{p}}, \mathbf{p}) + \lambda_h D_{\mathrm{KL}}(G_\sigma(\mathbf{p}) \parallel H). \tag{3}$$

Here $G_\sigma(\mathbf{p})$ is a normalized Gaussian heatmap centered at the target coordinate. We use $\lambda_c = 1.0$, $\lambda_h = 0.1$, and $\sigma = 2.0$. The coordinate term penalizes localization error directly, whereas the KL term discourages diffuse probability mass and stabilizes the spatial response around the annotated landmark.

In the DSNT coordinate transform, the spatial softmax temperature was fixed to $\tau = 1.0$ for all tasks. We did not tune τ separately for individual measurements, in order to avoid introducing task-specific calibration into the coordinate transform. A smaller temperature would produce sharper heatmaps but may reduce gradient smoothness, whereas a larger temperature would generate more diffuse spatial responses. We therefore keep a fixed default temperature and leave systematic temperature calibration as future work.

3.5 Evaluation Metrics

Validation is first reported in the original image coordinate system. For N images and K landmarks, the mean radial error (MRE) is

$$\mathrm{MRE} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \|\hat{\mathbf{q}}_{ik} - \mathbf{q}_{ik}\|_2, \tag{4}$$

where $\mathbf{q}_{ik}$ and $\hat{\mathbf{q}}_{ik}$ are ground-truth and predicted pixel coordinates. We also report mean absolute coordinate error (MAE):

$$\mathrm{MAE} = \frac{1}{2NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(|\hat{x}_{ik} - x_{ik}| + |\hat{y}_{ik} - y_{ik}| \right). \tag{5}$$

MRE summarizes Euclidean displacement of the landmark, whereas MAE averages the horizontal and vertical coordinate deviations. Reporting both quantities makes it possible to distinguish radial localization error from axis-wise coordinate bias. Because pixel errors are not directly comparable across tasks with different image sizes, we additionally report a diagonal-normalized mean error:

$$\mathrm{NME} = \frac{100}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\|\hat{\mathbf{q}}_{ik} - \mathbf{q}_{ik}\|_2}{\sqrt{H_i^2 + W_i^2}}, \tag{6}$$

where H_i and W_i are the original image height and width. NME expresses the radial landmark error as a percentage of image diagonal length, giving a more comparable scale for multi-resolution and multi-organ evaluation.

4 Experiments

4.1 Dataset and Training Setup

Nine ultrasound landmark-regression tasks are used for evaluation, spanning obstetric, intrapartum, vascular, and echocardiographic measurements with different data volumes, point counts, resolutions, and anatomical definitions. For each task, labeled images are split into training and internal validation subsets.DINOv3 ViT-base is loaded through `timm` with the same pretrained checkpoint for all task-specific models.Training uses AdamW, mixed precision, gradient clipping, five warmup epochs, cosine learning-rate decay, and validation-based early stopping. The budget is capped at 150 epochs, and the reported checkpoint has the lowest validation coordinate error.

Table 2. Implementation settings used for all task-specific DINOv3-DSNT models.

Component	Setting
Backbone	DINOv3 ViT-base, patch size 16
Feature depths	Transformer blocks 2, 5, 8, 11
Decoder	Projection, stride-4 fusion, shallow image branch
Output	K landmark heatmaps followed by DSNT
Input channels	Three-channel ultrasound image
Optimizer	AdamW, learning rate 10^{-4}, weight decay 10^{-4}
Batching	Batch size 1, gradient accumulation 4
Regularization	Dropout 0.1, gradient clipping, heatmap KL term
Evaluation space	Original image pixel coordinates

For each task, the training and validation split follows the released task-specific annotation files. No validation image is used for parameter updates, and the validation set is only used for checkpoint selection through early stopping. All task-specific models are initialized from the same DINOv3 ViT-base pretrained checkpoint loaded through `timm`; no additional ultrasound-specific pretraining is used in this study.The same optimizer, loss weights, mixed-precision training strategy, gradient accumulation setting, and early-stopping criterion are used across tasks unless otherwise specified.

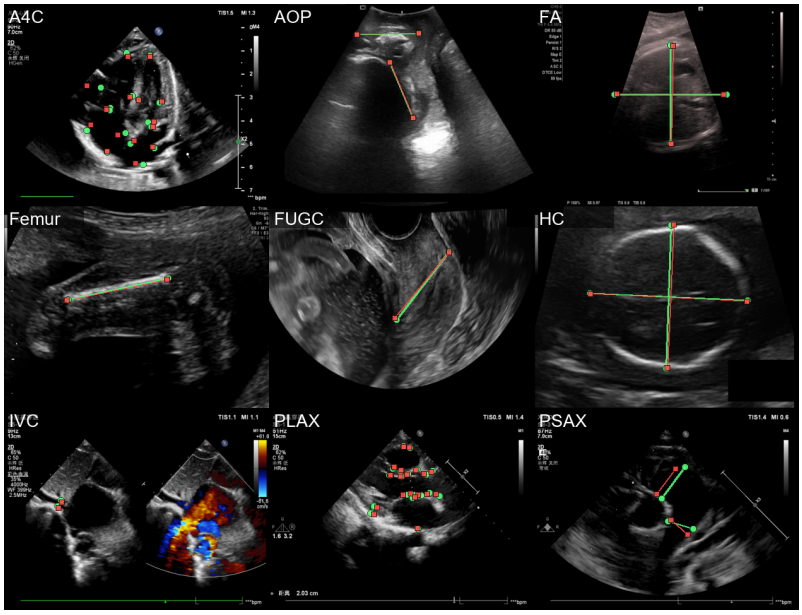

Fig. 2. Representative low-error validation predictions from the nine ultrasound regression tasks. Green circles denote ground truth, red squares denote predictions, and task labels are shown in each panel.

4.2 Internal Validation Results

Internal validation results are summarized in Table 3. Across the nine tasks, the average MRE is 22.83 pixels, the average MAE is 14.03 pixels, and the average NME is 2.20%. AOP and FUGC yield the smallest errors, consistent with their relatively stable image geometry and compact landmark configurations. PLAX remains competitive despite requiring 22 landmarks, indicating that the fused multi-depth representation can support dense cardiac endpoint localization when the anatomical view is sufficiently regular.

The highest errors occur in A4C and PSAX, with fetal femur and IVC also showing larger deviations than the simpler two-point tasks. These tasks either have limited validation samples, subtle endpoint definitions, or stronger dependence on image-plane selection. HC obtains moderate pointwise error, but its derived circumference is expected to be more sensitive to geometric consistency than to isolated coordinate accuracy. This pattern suggests that point-level supervision is an effective general training signal, while measurement-aware or shape-aware constraints may be necessary for structures governed by explicit geometric priors.

The averaged metrics should be interpreted together with the per-task breakdown. AOP and FUGC obtain the lowest errors, whereas A4C and PSAX show the largest localization deviations. This indicates that the mean performance is not uniformly distributed across tasks. In particular, cardiac views with limited

Table 3. Task-specific configuration and internal validation coordinate error. Image sizes are height × width. MRE and MAE are measured in original pixels; NME is the radial error normalized by the original image diagonal.

Task	K	Train	Val	Input	MRE	MAE	NME(%)
A4C	16	86	22	768×1024	50.12	28.67	4.93
AOP	4	3200	800	512×512	3.77	2.38	0.52
FA	4	400	100	768×1024	19.15	11.60	1.50
Femur	2	582	145	736×1056	29.43	18.10	2.79
FUGC	2	208	52	336×544	5.38	3.45	0.84
HC	4	799	200	544×816	16.74	10.36	1.73
IVC	2	30	8	768×1024	26.31	17.01	2.57
PLAX	22	70	17	768×1024	14.08	8.82	1.22
PSAX	4	39	10	768×1024	40.52	25.89	3.74
Mean	–	–	–	–	22.83	14.03	2.20

training data and compact or weak boundary targets remain substantially more challenging than geometrically stable two-point or four-point measurements.

4.3 Comparison with Existing Methods

Direct numerical comparison with prior ultrasound biometry methods is difficult because many existing studies focus on a single organ, use private datasets, or report measurement-specific errors under different annotation protocols [10, 11, 20, 6, 5, 19]. For example, fetal biometry, intrapartum ultrasound, inferior vena cava measurement, and echocardiographic quantification are usually evaluated with different landmark definitions, imaging views, and clinical target measurements. In contrast, this study evaluates one reusable DINOv3-DSNT recipe across nine heterogeneous ultrasound landmark-regression tasks.

Unlike direct coordinate regression formulations, the proposed formulation preserves a spatial heatmap representation before producing continuous landmark coordinates.This is useful for ultrasound biometry because weak boundaries, speckle, acoustic shadowing, and ambiguous anatomical interfaces may produce multiple plausible local responses. Compared with task-specific single-organ pipelines, the present framework emphasizes architectural reuse while preserving the output dimension,input resolution, and landmark ordering required by each clinical measurement protocol. Therefore, the contribution is not a new single-task measurement pipeline, but a reusable high-resolution landmark regression recipe for heterogeneous ultrasound biometry.

5 Discussion

The experiments suggest that spatial scale must be preserved during ultrasound landmark regression. Many landmarks are defined by short echogenic interfaces or by intersections between anatomical contours and measurement axes. Once

these cues are lost through aggressive resizing, a stronger encoder cannot fully recover them. Task-level resolution selection therefore balances localization fidelity and computational cost.

The heatmap target also remains useful beyond coordinate output. It supplies a spatial confidence distribution that can be regularized and inspected, while DSNT converts this distribution into continuous coordinates without discarding differentiability. This is valuable when weak boundaries and speckle produce several plausible local maxima.

The separation between shared architecture and task-specific heads is also important. A fully shared landmark vocabulary would imply that point indices have comparable meanings across organs. The present design instead treats each measurement as a distinct anatomical coordinate system while reusing the same visual representation and decoder principle.

The current formulation remains point-centric, whereas some measurements also depend on parallelism, orthogonality, ellipse consistency, or angle construction. Future work should incorporate measurement-aware losses, shape priors, heatmap-entropy uncertainty estimates, and consistency checks for paired endpoints. Sharing the DINOv3 encoder across tasks while retaining lightweight task-specific decoders may further improve low-data measurements.

The task-level error profile indicates where such constraints are most needed. A4C and PSAX remain difficult because they combine small cardiac datasets with weak or compact boundary targets. Femur and IVC are two-point tasks, but their landmarks lie on narrow elongated structures affected by shadowing, dropout, or vessel-wall contrast. HC has lower pointwise NME, yet its downstream circumference remains sensitive to elliptical consistency. These observations motivate view-aware priors for cardiac tasks, boundary-enhanced supervision for femur and IVC, and explicit ellipse or paired-axis constraints for HC.

Several limitations remain. First, the evaluation is based on internal validation rather than an external multi-center test set, and some tasks contain small validation sets such as IVC and PSAX; the corresponding results should therefore be interpreted cautiously. Second, although the same architecture recipe is reused across measurements, the current implementation trains separate task-specific checkpoints and should not be considered a fully unified multi-task model. Third, the reported metrics are point-level localization errors, whereas clinical measurements may depend on derived geometry such as diameter, circumference, angle, parallelism, orthogonality, or ellipse consistency. Future work will incorporate measurement-level errors, shape-aware constraints, heatmap-entropy uncertainty estimates, and external validation.

6 Conclusion

High-resolution DINOv3 features with DSNT heatmap regression provide a compact and reusable formulation for multi-organ ultrasound landmark localization. The framework preserves anatomical point definitions by combining shared visual computation, task-specific heads, and native-scale inputs. Across nine inter-

nal validation tasks, it achieves 22.83-pixel MRE, 14.03-pixel MAE, and 2.20% NME in the original image coordinate system. The per-task error profile further shows that geometrically stable measurements are easier to localize, whereas cardiac views and small-data tasks remain more challenging. These findings motivate future work on measurement-level geometric constraints, uncertainty estimation, and external validation.

References

1. Angeli, L., Conversano, F., Dall'Asta, A., Volpe, N., Simone, M., Di Pasquo, E., Pignatelli, D., Schera, G.B.L.: New technique for automatic sonographic measurement of change in head-perineum distance and angle of progression during active phase of second stage of labor. Ultrasound in Obstetrics & Gynecology **56**(4), 597–602 (2020). https://doi.org/10.1002/uog.21963
2. 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 (2025). https://doi.org/10.5281/zenodo.15081529, mICCAI 2025 Challenge
3. Bai, J., Yaqub, M., Ma, J., Lekadir, K., Gan, J., Cai, W., Ni, D., Li, S.: Foundation model challenge for ultrasound biometry (2026). https://doi.org/10.5281/zenodo.19736827, mICCAI 2026 Challenge
4. Bai, J., Zhou, Z., Tang, Y., Gan, J., Liang, Z., Fan, J., Cai, W., Ni, D., Lu, Y., Lekadir, K., Li, S., et al.: Beyond benchmarks of iugc: Rethinking requirements of deep learning method for intrapartum ultrasound biometry from fetal ultrasound videos. Medical Image Analysis **111**, 104043 (2026). https://doi.org/10.1016/j.media.2026.104043
5. Chen, J., Li, J., Ding, X., Wei, G., Wang, X., Li, Q.: Automated ultrasound measurement of the inferior vena cava: An animal study. Ultrasonic Imaging **42**(3), 148–158 (2020). https://doi.org/10.1177/0161734620912345
6. Conversano, F., Peccarisi, M., Pisani, P., Di Paola, M., De Marco, T., Franchini, R., Greco, A., D'Ambrogio, G.: Automatic ultrasound technique to measure angle of progression during labor. Ultrasound in Obstetrics & Gynecology **50**(6), 766–775 (2017). https://doi.org/10.1002/uog.17441
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. Ghelich Oghli, M., Shabanzadeh, A., Moradi, S., Sirjani, N., Gerami, R., Ghaderi, P., Sanei Taheri, M., Shiri, I.: Automatic fetal biometry prediction using a novel deep convolutional network architecture. Physica Medica **88**, 127–137 (2021). https://doi.org/10.1016/j.ejmp.2021.06.020
9. Ghi, T., Eggebo, T., Lees, C., Kalache, K., Rozenberg, P., Youssef, A., Salomon, L.J., Tutschek, B.: ISUOG practice guidelines: intrapartum ultrasound. Ultrasound in Obstetrics & Gynecology **52**(1), 128–139 (2018). https://doi.org/10.1002/uog.19072
10. van den Heuvel, T.L.A., Petros, H., Santini, S., de Korte, C.L., van Ginneken, B.: Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries. Ultrasound in Medicine & Biology **45**(3), 773–785 (2019). https://doi.org/10.1016/j.ultrasmedbio.2018.09.015

11. Kim, H.P., Lee, S.M., Kwon, J.Y., Park, Y., Kim, K.C., Seo, J.K.: Automatic evaluation of fetal head biometry from ultrasound images using machine learning. Physiological Measurement **40**(6), 065009 (2019). https://doi.org/10.1088/1361-6579/ab21ac
12. Ma, C., Li, Y., Fu, J., Liang, S., Wang, Y., Guo, Y.: Unified ultrasound intelligence toward an end-to-end agentic system. arXiv preprint arXiv:2604.16914 (2026)
13. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. In: European Conference on Computer Vision Workshops. pp. 457–468. Springer (2018)
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
15. Plotka, S., Klasa, A., Lisowska, A., Seliga-Siwecka, J., Lipa, M., Trzcinski, T., Sitek, A.: Deep learning fetal ultrasound video model match human observers in biometric measurements. Physics in Medicine and Biology **67**(4) (2022). https://doi.org/10.1088/1361-6560/ac4d85
16. Simeoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D.: DINOv3 (2025)
17. Yang, Z., Li, Z., Liu, Q., Wang, K., Fu, Y.: UniUltra: A unified foundation model for universal ultrasound analysis. In: IEEE International Symposium on Biomedical Imaging (2026)
18. Zamzmi, G., Rajaraman, S., Hsu, L.Y., Sachdev, V., Antani, S.: Real-time echocardiography image analysis and quantification of cardiac indices. Medical Image Analysis **80**, 102438 (2022). https://doi.org/10.1016/j.media.2022.102438
19. Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E.: Fully automated echocardiogram interpretation in clinical practice. Circulation **138**(16), 1623–1635 (2018). https://doi.org/10.1161/CIRCULATIONAHA.118.034338
20. Zhu, F., Liu, M., Wang, F., Qiu, D., Li, R., Dai, C.: Automatic measurement of fetal femur length in ultrasound images: a comparison of random forest regression model and segnet. Mathematical Biosciences and Engineering **18**(6), 7790–7805 (2021). https://doi.org/10.3934/mbe.2021387