Hierarchical Ensembles of Complementary Landmark Detectors for Multi-Task Ultrasound Biometry

Ilia Semenkov[1,2][0000−0003−1515−7062]

[1] AXXX, Moscow, Russia
[2] HSE University, Moscow, Russia
semenkov@axxx.tech

Abstract. Ultrasound biometry combines landmark localization across anatomically and statistically heterogeneous tasks. The FoundUS challenge includes nine prenatal, intrapartum, and cardiac ultrasound tasks whose training-set sizes, landmark definitions, and derived measurements differ substantially. We propose a unified multi-task system using fully fine-tuned DINOv2 encoders, task-specific localization heads, and balanced task sampling. To address task-dependent errors, we construct a hierarchical ensemble combining DSNT, residual log-likelihood estimation, SimCC, and a dual RLE–SimCC head, together with diversity in training partitions, initialization, encoder capacity, and input resolution. The final system contains 17 checkpoints organized into three equally weighted blocks. On the official validation server, it achieved an Average MRE of 24.233 and an Average MAE of 29.345. Relative to the MRE-optimal two-block ensemble on the same validation set, the selected system changed Average MRE by only 0.015 pixels while improving MAE on seven of nine tasks. After the final system was fixed, the same implementation achieved an Average MRE of 21.956 and an Average MAE of 26.579 on the organizer-held hidden test set. These results support structured ensemble diversity for unified ultrasound landmark localization under substantial task heterogeneity.

Keywords: Ultrasound biometry · Landmark localization · Multi-task learning · DINOv2 · Deep ensembles

1 Introduction

Ultrasound biometry supports clinical assessment across prenatal, intrapartum, and adult care. Measurements are commonly obtained by locating anatomical landmarks and deriving distances, circumferences, or angles from their geometry. Manual landmark placement is time-consuming and operator-dependent, motivating automatic systems that can improve measurement consistency and clinical workflow. Previous methods have achieved accurate results for individual fetal or cardiac measurements [2, 5, 10], but models developed for a single anatomy, acquisition protocol, or measurement may not transfer reliably to other ultrasound domains.

The FoundUS 2026 Foundation Model Challenge for Ultrasound Biometry [3] addresses this limitation through a unified benchmark containing nine tasks from prenatal, intrapartum, and cardiac ultrasound. These tasks differ substantially in appearance, acquisition conditions, training-set size, landmark cardinality, and the geometric relationship between landmarks and clinical measurements. The smallest released task contains only 38 training images, whereas the largest contains 4000. A successful solution must therefore share useful representations across tasks without allowing data-rich domains to dominate training. It must also accommodate task-dependent landmark definitions and remain accurate under both point-localization and derived-parameter evaluation.

We build a unified system around fully fine-tuned DINOv2 encoders. A shared visual backbone is paired with task-specific localization layers, while temperature-balanced task sampling controls the severe imbalance between datasets. Since no single localization formulation was uniformly preferable across the heterogeneous tasks, we combine models based on DSNT, residual log-likelihood estimation, SimCC, and a dual RLE–SimCC head. The final hierarchical ensemble introduces complementary variation through localization objectives, training partitions and random initialization, encoder capacity, and input resolution.

Our main contributions are threefold. First, we present a unified multi-task framework for nine heterogeneous ultrasound-biometry tasks with explicit task balancing and task-specific landmark prediction. Second, we combine complementary coordinate-localization formulations within a common encoder and training framework. Third, we introduce a hierarchical ensemble that assigns equal weight to localization, partition/initialization, and capacity/scale diversity. On the official validation server, the selected system achieves an Average MRE of 24.233 and an Average MAE of 29.345. Adding the capacity and scale block changes Average MRE by only 0.015 pixels relative to the MRE-optimal two-block ensemble while improving MAE on seven of nine tasks.

2 Related Work

Ultrasound biometry and landmark localization. Automatic ultrasound biometry has been approached through segmentation, direct parameter regression, and anatomical landmark detection. BiometryNet predicts measurement-specific landmarks directly from standard fetal ultrasound planes and enforces orientation consistency during training [2]. Domain shift between ultrasound devices has been addressed through adversarial adaptation for fetal brain biometry [5], while landmark detection and tracking have been combined to exploit sparsely annotated echocardiographic sequences [10]. In contrast to these anatomy- or measurement-specific settings, FoundUS requires one system to support multiple prenatal, intrapartum, and cardiac tasks.

Pretrained representations for ultrasound. Large-scale visual pretraining can reduce reliance on task-specific annotated data. DINOv2 learns transferable visual representations from natural images [12], while ultrasound-specific foundation

models such as USFM use large multi-organ ultrasound collections and domain-oriented self-supervised objectives [6]. UltraDINO further demonstrates that DINOv2-style pretraining can be effectively transferred to large fetal-ultrasound collections [1]. Our work does not introduce a new foundation model; instead, we fully fine-tune publicly available DINOv2 representations within a unified multi-task landmark-localization system.

Coordinate prediction and ensembles. Landmark localization may be formulated using spatial heatmaps, continuous coordinate regression, or coordinate classification. DSNT obtains continuous coordinates through differentiable spatial expectation [11]; RLE models regression residuals through a learned likelihood [8]; and SimCC treats the two coordinate axes as classification problems [9]. These formulations impose different spatial and statistical biases. Rather than selecting one formulation globally, our system combines them with variation in training data, encoder capacity, and inference resolution through a structured deep ensemble [7].

3 Method

3.1 Unified Multi-Task Architecture

The challenge comprises nine ultrasound biometry tasks with different anatomies, image characteristics, landmark definitions, and numbers of target points. Given an image $\mathbf{x}$ and its challenge-provided task identifier t, the model predicts

$$\hat{\mathbf{y}}^{(t)} = \left(\hat{\mathbf{p}}_k^{(t)}\right)_{k=1}^{K_t}, \qquad \hat{\mathbf{p}}_k^{(t)} \in [0,1]^2. \tag{1}$$

Here, K_t is the task-specific number of landmarks. The corresponding biometric parameters are computed from the predicted coordinates by the official challenge evaluation procedure.

All component models follow the same multi-task design. A DINOv2 ViT-B/14 or ViT-L/14 encoder with four register tokens [12, 4] is shared across the nine tasks, whereas the final localization layers are task-specific to accommodate different landmark definitions and output dimensionalities. The encoder is initialized from publicly available pretrained weights and fully fine-tuned. Every component model is jointly trained on all tasks; the ensemble does not contain independently trained task-specific expert models.

Images are converted to RGB and resized using aspect-ratio-preserving letterboxing. The resized image is aligned with the top-left corner and padded on the right and bottom. Unless stated otherwise, models are trained with 518×518 inputs. Predicted normalized coordinates are transformed back to the original image resolution before evaluation or ensembling.

3.2 Balanced Multi-Task Training

The number of training examples varies substantially across tasks. To prevent the largest datasets from dominating optimization, we use task-homogeneous

mini-batches and sample task t with probability

$$q(t) = \frac{n_t^\tau}{\sum_{j=1}^{T} n_j^\tau}, \tag{2}$$

where n_t is the number of training images for task t. Most component models use square-root balancing, $\tau = 0.5$, while one complementary ensemble member uses $\tau = 0.25$.

Training augmentation includes random rotation, translation, isotropic scaling, and photometric perturbations. Horizontal and vertical flips are excluded because they may change anatomical orientation or landmark identity. Different ensemble members use either lighter or stronger photometric augmentation to increase diversity.

3.3 Complementary Localization Heads

The component models use several established landmark-localization formulations with different inductive biases.

DSNT. The DSNT head predicts a spatial probability map for every landmark and obtains its coordinate through differentiable spatial expectation [11]. Training combines Smooth-L1 coordinate regression with Jensen–Shannon regularization toward a Gaussian distribution centered at the target. We use both a standard decoder and a higher-resolution variant with learned spatial upsampling before heatmap prediction.

Residual log-likelihood estimation. Our RLE-style head adapts residual log-likelihood estimation [8] to spatial landmark prediction. A task-specific heatmap and differentiable spatial expectation produce the landmark mean, while globally pooled features predict one positive scale per landmark. A two-stage affine-coupling flow with a Laplace base distribution models the normalized two-dimensional residual. A higher-resolution variant introduces learned spatial upsampling before predicting the landmark mean and scale.

Spatial SimCC. SimCC represents two-dimensional localization as separate classification problems for the horizontal and vertical coordinates [9]. Our spatial variant first predicts a two-dimensional landmark logit map and then obtains axis-wise logits by log-sum-exp marginalization over the complementary spatial dimension. The logits are interpolated to 512 coordinate bins and trained using Gaussian soft targets together with a coordinate-regression term.

Dual RLE–SimCC head. To combine likelihood-based residual modeling and axis-wise coordinate classification, we train RLE and SimCC branches on the same encoder features. We set the agreement weight to $\lambda_{\text{agr}} = 0.25$ and average the two coordinate predictions at inference. The training objective is

$$\mathcal{L}_{\text{dual}} = \frac{\mathcal{L}_{\text{RLE}} + \mathcal{L}_{\text{SimCC}}}{2} + \lambda_{\text{agr}} \,\text{SmoothL1}\big(\hat{\mathbf{y}}_{\text{RLE}}, \hat{\mathbf{y}}_{\text{SimCC}}\big). \tag{3}$$

3.4 Hierarchical Ensemble

The final solution combines three blocks designed to capture complementary sources of variation. All predictions are first mapped to the original image coordinate system and then aggregated within and across blocks.

Block A: localization diversity. The first block contains eight multi-task models spanning DSNT, RLE, SimCC, and dual RLE–SimCC heads. The models also differ in encoder capacity, decoder resolution, task-sampling temperature, and augmentation strength. This block provides diversity across coordinate representations, decoder resolutions, encoder capacities, task-sampling distributions, and augmentation regimes.

Block B: partition and initialization diversity. The second block contains six RLE models. Five are trained using different partitions of the released training data and different random seeds, while one fixed-partition model is shared with Block A. Block B therefore provides joint training-partition and initialization diversity; the effects of these two factors are not separated.

Block C: capacity and scale diversity. The third block contains four RLE or dual-head models using DINOv2-B or DINOv2-L encoders. Each model is evaluated at input resolutions of 462, 518, and 574 pixels, corresponding to DINOv2 patch grids of 33×33, 37×37, and 41×41. Predictions are transformed to original-image coordinates, and the coordinate-wise median across the three scales is used as the model prediction. The four multiscale predictions are subsequently averaged.

Let $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ denote the mean predictions produced by the three blocks. The final prediction assigns equal weight to each block:

$$\hat{\mathbf{y}}_{\text{final}} = \frac{\mathbf{A} + \mathbf{B} + \mathbf{C}}{3}. \tag{4}$$

The complete ensemble contains 17 unique checkpoints and uses 25 distinct checkpoint–scale evaluations per image. We use block-level rather than uniform checkpoint-level weighting so that each of the three blocks contributes equally.

3.5 Anatomical Canonicalization

For A4C and PSAX, consecutive landmark pairs have a prescribed vertical ordering. A model can accurately localize both endpoints while reversing their identities. For these tasks, we exchange the two predicted points whenever the first endpoint lies below the second. This deterministic canonicalization is applied to each checkpoint-level prediction before cross-model averaging; for multi-scale models, it follows the within-checkpoint median aggregation. No analogous post-processing is used for the remaining tasks.

4 Experimental Setup

4.1 Data and Evaluation

The challenge comprises nine ultrasound landmark-localization tasks with strongly imbalanced training sets: A4C (108 training/20 validation images, 16 landmarks), AOP (4000/60, 4), FA (500/188, 4), fetal femur (702/62, 2), FUGC (260/20, 2), HC (999/215, 4), IVC (38/10, 2), PLAX (87/26, 22), and PSAX (49/18, 4). After excluding 25 fetal-femur images with organizer-identified orientation anomalies, 6743 annotated training images remained. The official validation set contained 619 images with hidden annotations. No additional ultrasound data were used, and DINOv2 initialization was the only source of external pretraining.

The official server reports mean radial error (MRE) for landmark localization and mean absolute error (MAE) for biometric parameters derived from the predicted landmarks. Average MRE and Average MAE are unweighted means over the nine tasks. Since the biometric parameters have task-dependent scales and units, raw Average MAE is interpreted as a challenge-level aggregate rather than a single physical error. The official validation server was used during model and ensemble development. After the final system had been fixed, the same implementation was evaluated on the organizer-held hidden test set, whose annotations were not available to the author.

4.2 Implementation Details

All component models were trained for 50 epochs using AdamW, mixed-precision training, and cosine learning-rate annealing. The encoder and localization decoder used learning rates of 10^{-5} and 10^{-3}, respectively, with weight decay 10^{-4} and batch size four.

A held-out subset of the released training data was used for checkpoint selection. Most models used one fixed internal split, whereas the five Block B models used different splits and random seeds. For each run, we retained the checkpoint with the lowest unweighted mean MRE across the nine internal validation tasks. Final predictions were generated from these selected checkpoints and combined as described in Sect. 3.

5 Results and Discussion

5.1 Progressive Ensemble Construction

Table 1 reports official validation results for the three ensemble blocks and their combinations. Block A, which combines localization formulations and model configurations, outperformed the RLE split/seed ensemble in Block B when evaluated independently. Nevertheless, averaging Blocks A and B reduced Average MRE to 24.218, compared with 24.350 for Block A and 24.531 for Block B. This

Table 1. Progressive ensemble results on the official validation server. Lower values are better.

System	Primary diversity	Avg. MRE	Avg. MAE
Block A	Localization and architecture	24.350	29.401
Block B	Partition and initialization	24.531	29.738
Blocks A+B	Two-block ensemble	**24.218**	29.446
Blocks A+B+C	Final ensemble	24.233	**29.345**

Table 2. Task-level results on official validation and the organizer-held hidden test set. Bold indicates the better result between A+B and the final ensemble on validation. Validation and hidden-test results are evaluated on different images and are not directly comparable. Lower values are better.

	MRE			MAE		
Task	A+B val.	Final val.	Final test	A+B val.	Final val.	Final test
A4C	**23.188**	23.649	23.928	**22.989**	23.439	23.437
AOP	13.095	**12.875**	18.845	8.622	**8.416**	6.379
FA	**19.986**	19.992	24.763	96.943	**96.851**	100.650
Fetal femur	**21.612**	21.750	12.518	15.891	**15.872**	12.117
FUGC	9.218	**9.212**	14.470	9.237	**8.909**	9.426
HC	49.543	**48.823**	26.352	74.258	**73.746**	49.256
IVC	**28.041**	28.641	19.800	**7.635**	7.902	5.857
PLAX	**15.882**	15.913	16.702	7.662	**7.636**	8.143
PSAX	37.401	**37.242**	40.224	21.777	**21.332**	23.947
Average	**24.218**	24.233	21.956	29.446	**29.345**	26.579

improvement over either constituent block indicates that partition and initialization diversity remained complementary to the stronger localization-diverse ensemble.

Adding the capacity and multiscale models in Block C produced an Average MRE of 24.233, only 0.015 pixels above the MRE-optimal two-block system, while reducing Average MAE from 29.446 to 29.345. Because the challenge evaluates both landmark localization and the accuracy of derived biometric parameters, we selected the three-block system as our final submission.

5.2 Task-Level Performance

Table 2 compares the MRE-optimal two-block ensemble with the selected three-block system. Adding Block C improved MAE on seven of nine tasks and improved both metrics for AOP, FUGC, HC, and PSAX. Fetal femur, FA, and PLAX showed slightly higher MRE but lower MAE, whereas A4C and IVC favored the two-block system on both metrics. The largest MRE gain occurred for HC, while the largest degradations occurred for A4C and IVC; the latter contains only ten validation images and is therefore particularly sensitive to individual cases.

After the final system had been fixed, the same implementation achieved an Average MRE of 21.956 and an Average MAE of 26.579 on the hidden test set.

5.3 Landmark and Biometric Errors

The results demonstrate that landmark MRE and downstream parameter error provide related but non-equivalent views of performance. MRE penalizes the Euclidean displacement of each landmark independently, whereas a biometric parameter depends on the geometry of a particular landmark subset. For example, a similar displacement of both endpoints may increase landmark error while largely preserving their distance. Conversely, smaller coordinate errors can produce a substantial parameter error when they perturb an angle or act in opposing directions.

This distinction is visible for fetal femur and PLAX, where the final ensemble produced slightly higher MRE but lower MAE. It also explains why optimizing the ensemble exclusively for Average MRE need not produce the strongest solution under a challenge protocol that additionally evaluates clinically derived measurements. The three-block ensemble provides a more balanced operating point: it preserves the best Average MRE within 0.015 pixels while improving parameter accuracy across most tasks.

The main limitation of the proposed system is computational cost. The final ensemble uses 17 trained models and requires 25 forward evaluations per image, compared with 13 models and 13 evaluations for the MRE-optimal A+B ensemble. This additional cost was accepted because the final ensemble improved MAE on seven of nine tasks while preserving essentially the same Average MRE. A smaller distilled or selectively routed ensemble would be preferable for clinical deployment.

6 Conclusion

We presented a unified solution for nine heterogeneous ultrasound-biometry tasks based on fully fine-tuned DINOv2 encoders, temperature-balanced multi-task training, and task-specific landmark heads. A hierarchical ensemble combined diversity in localization formulation, training partition, initialization, encoder capacity, and input scale. On the official validation server, the final system achieved an Average MRE of 24.233 and an Average MAE of 29.345. Compared with the MRE-optimal two-block ensemble, it preserved landmark accuracy within 0.015 pixels while improving MAE on seven of nine tasks. On the organizer-held hidden test set, the fixed final system achieved an Average MRE of 21.956 and an Average MAE of 26.579.

Acknowledgments. This research was supported in part through computational resources of HPC facilities at HSE University. Large language models were used to assist with manuscript preparation and research code development.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Ambsdorf, J., Munk, A., Llambias, S., Christensen, A.N., Mikolaj, K., Balestriero, R., Tolsgaard, M.G., Feragen, A., Nielsen, M.: General Methods Make Great Domain-Specific Foundation Models: A Case-Study on Fetal Ultrasound. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 271–281. Springer Nature Switzerland, Cham (2025)
2. Avisdris, N., Joskowicz, L., Dromey, B., David, A.L., Peebles, D.M., Stoyanov, D., Ben Bashat, D., Bano, S.: BiometryNet: Landmark-based Fetal Biometry Estimation from Standard Ultrasound Planes. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 279–289. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-16440-8_27
3. Bai, J., Yaqub, M., Ma, J., Lekadir, K., Gan, J., Cai, W., Ni, D., Li, S.: Foundation model challenge for ultrasound biometry (Apr 2026). https://doi.org/10.5281/zenodo.19736827
4. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers (2024), https://arxiv.org/abs/2309.16588
5. Gao, Y., Lee, L., Droste, R., Craik, R., Beriwal, S., Papageorghiou, A., Noble, A.: A dual adversarial calibration framework for automatic fetal brain biometry. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3246–3254 (October 2021)
6. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., Guo, Y.: USFM: A Universal Ultrasound Foundation Model Generalized to Tasks and Organs towards Label Efficient Image Analysis (2024), https://arxiv.org/abs/2401.00153
7. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6405–6416. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
8. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11025–11034 (October 2021)
9. Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., Xia, S.T.: SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. pp. 89–106. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20068-7_6
10. Lin, J., Sahebzamani, G., Luong, C., Dezaki, F.T., Jafari, M., Abolmaesumi, P., Tsang, T.: Reciprocal landmark detection and tracking with extremely few annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15170–15179 (June 2021)
11. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks (2018), https://arxiv.org/abs/1801.07372
12. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma,

V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (2024), https://arxiv.org/abs/2304.07193