Semi-supervised DINOv3 Adaptation for Multi-domain Ultrasound Biometry

Ping Cui[a 1], Zi Yang[b 1], Jinke Li[a], Zengxing Li[a], and Yu Fu[a ⋆]

[a] Lanzhou University, Lanzhou, China
fuyu@lzu.edu.cn
[b] The First Hospital of Lanzhou University, Lanzhou, China

Abstract. Ultrasound biometry underpins prenatal, intrapartum, and cardiac assessment, but many measurements remain time-consuming and operator-dependent. The MICCAI 2026 Foundation Model Challenge for Ultrasound Biometry asks a single model to localize task-specific landmarks across several ultrasound domains. We address this setting with a semi-supervised, multi-task detector. The method combines a DINOv3-pretrained encoder, CBAM heatmap heads, an exponential-moving-average teacher, task-aware measurement losses, and DARK sub-pixel decoding. The submitted model achieved an MRE of 24.96 and an MAE of 27.48 on the official challenge test set. On a fixed internal split, the best macro-validation MRE was 18.4827, 13.98% lower than the recorded DINOv2 supervised baseline. In a verified original-scale comparison, semi-supervised DINOv2 reduced MRE from 63.346 to 38.644 relative to supervised DINOv2. DARK provided only a small benefit over argmax for identical heatmaps. These results support the use of unlabelled data while placing the decoder contribution in proportion.

Keywords: Ultrasound biometry · Semi-supervised learning · Landmark detection · Multi-task learning

1 Introduction

Ultrasound biometry translates anatomical observations into clinically useful lengths, angles, diameters, and circumferences. Fetal head, abdominal, and femur measurements inform prenatal growth assessment [3]. The angle of progression (AOP) supports intrapartum monitoring [4], whereas chamber and vascular measurements are central to cardiac assessment [5, 6]. In routine practice, however, many of these measurements remain manual or semi-automated and vary with operator experience and image quality.

Artificial intelligence has automated individual ultrasound tasks, including fetal-plane detection and localization [7]. Most systems nevertheless remain tied to one anatomy, view, or clinical parameter. The MICCAI 2026 Foundation

⋆ Corresponding author.
[1] Ping Cui and Zi Yang contributed equally to this work.

Model Challenge for Ultrasound Biometry instead poses a multi-domain, multi-task landmark detection problem [1, 2]. A task identifier selects the relevant landmark template, and task-specific geometry converts the predicted landmarks into biometric parameters. Success therefore depends on transferable image features, precise localization, and effective use of sparsely labelled data.

We adapt a DINOv3-pretrained visual encoder to this heterogeneous benchmark [12, 13]. Task-specific CBAM heads predict landmark heatmaps [15], while an EMA teacher supplies soft targets for unlabelled images [16]. Task-aware losses connect landmark learning to the required measurements, and DARK refines the final coordinates [19]. The resulting framework accommodates different landmark templates without abandoning a shared representation. We assess the submitted system on the official challenge test set and use a fixed internal split for verified comparisons of semi-supervised training and heatmap decoding.

2 Related Work

Deep learning approaches to ultrasound biometry include segmentation, heatmap regression, landmark detection, and view classification [7–11]. These methods have reduced manual work for specific measurements, but most address one anatomical region or measurement family. Here, a single model spans prenatal, intrapartum, and cardiac tasks.

Self-supervised visual representations have become common starting points for medical image analysis [12–14]. For ultrasound, pretraining may aid transfer across devices, views, and variable speckle patterns. We use DINOv3 as a pretrained feature extractor; we do not train a new ultrasound foundation model.

Landmark annotation is costly, particularly when each task requires a different template. Mean Teacher methods use exponential moving averages to obtain more stable targets from unlabelled data [16–18]. In our detector, the teacher produces soft heatmap targets during training but is not part of the inference output.

Heatmap regression retains spatial information, although discrete heatmap coordinates introduce quantization error. DARK recovers sub-pixel coordinates from the predicted distribution [19]. This refinement is pertinent to biometry because small landmark offsets can propagate into length and angle errors.

3 Method

3.1 Task Formulation and Benchmark

Given an ultrasound image x_i and task identifier t_i, the model predicts task-specific landmarks and converts them into biometric parameters:

$$\hat{P}_i = f_\theta(x_i, t_i) = \{(\hat{x}_k, \hat{y}_k)\}_{k=1}^{K_{t_i}}, \quad \hat{m}_i = g_{t_i}(\hat{P}_i), \tag{1}$$

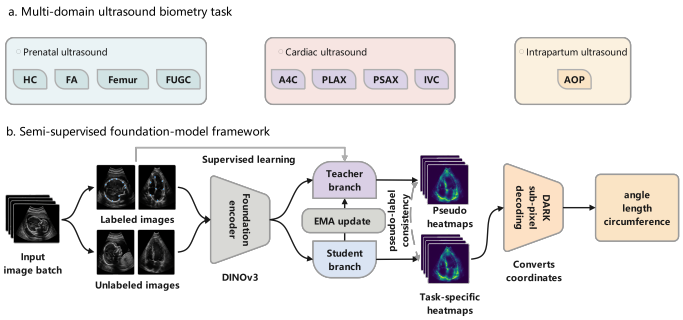

Fig. 1. Overview of the semi-supervised multi-task ultrasound biometry framework. A DINOv3-pretrained encoder feeds task-specific heatmap branches; the EMA teacher provides pseudo-heatmap supervision during training, and DARK decoding converts predicted heatmaps into biometric parameters at inference.

where K_{t_i} is the number of landmarks for task t_i. The function g_{t_i} represents a task-specific geometric computation, such as a distance, angle, diameter, or circumference.

Figure 1 summarizes the task setting. The benchmark includes prenatal tasks (HC, FA, FUGC, and fetal femur), intrapartum AOP, and cardiac tasks (A4C, PLAX, PSAX, and IVC). Their landmark counts, target scales, anatomical appearances, and measurement geometries differ substantially.

3.2 Model and Semi-supervised Training

The shared backbone is a DINOv3-pretrained visual encoder [13]. It maps images from the different ultrasound domains into a common feature representation. Each task-specific head then applies CBAM channel and spatial attention before predicting one heatmap for each required landmark [15].

Sharing the encoder allows information to transfer across domains, whereas separate heads preserve the output dimensionality and measurement geometry of each task. This arrangement avoids imposing one fixed landmark layout on anatomically distinct targets.

Training combines labelled and unlabelled images. Labelled samples supervise the student heatmaps directly. On unlabelled data, the student is matched to heatmaps from a teacher network. The teacher receives no gradients and is updated after each student step by an exponential moving average:

$$\theta_T \leftarrow \alpha\theta_T + (1-\alpha)\theta_S, \tag{2}$$

where θ_T and θ_S denote teacher and student parameters, and α is the EMA decay factor.

For an unlabelled image x_u, the teacher produces pseudo-heatmaps $H_T(x_u)$ and the student predicts $H_S(x_u)$. The consistency loss is:

$$\mathcal{L}_{cons} = \|H_S(x_u) - H_T(x_u)\|_2^2 . \tag{3}$$

The soft heatmaps retain spatial uncertainty instead of reducing the teacher output to coordinates. The verified experiments used a fixed confidence threshold of 0.0 and did not apply non-zero confidence filtering.

For labelled samples, the supervised heatmap term compares the student prediction with the Gaussian target heatmap,

$$\mathcal{L}_{heatmap} = \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} \|H_S(x_i) - H_i^*\|_2^2 , \tag{4}$$

where $\mathcal{B}_l$ is a labelled mini-batch and H_i^* is the corresponding target. For a task with an annotated biometric value m_i, differentiable coordinates $\widetilde{P}_i$ are obtained from the predicted heatmaps and converted by g_{t_i}. The measurement-aware terms are

$$\begin{aligned} \mathcal{L}_{angle} &= \frac{1}{|\mathcal{B}_{AOP}|} \sum_{i \in \mathcal{B}_{AOP}} \left| g_{AOP}(\widetilde{P}_i) - m_i \right| , \\ \mathcal{L}_{param} &= \frac{1}{|\mathcal{B}_{param}|} \sum_{i \in \mathcal{B}_{param}} \left| g_{t_i}(\widetilde{P}_i) - m_i \right| . \end{aligned} \tag{5}$$

The first term is restricted to AOP samples, whereas the second covers tasks for which a differentiable scalar biometric parameter is available. The training objective is

$$\mathcal{L} = \mathcal{L}_{heatmap} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{angle}\mathcal{L}_{angle} + \lambda_{param}\mathcal{L}_{param}. \tag{6}$$

The recorded values were $\lambda_{cons} = 0.3$, $\lambda_{angle} = 0.05$, $\lambda_{param} = 0.10$, and $\alpha = 0.99$; these weights were fixed throughout training rather than ramped. Geometric augmentation was applied jointly to each image and its landmark targets and was therefore part of data preparation, not an additional loss term.

3.3 Inference and Biometric Parameter Computation

At inference, the task identifier selects the relevant heatmap head. The teacher is discarded, and DARK converts the student heatmaps into sub-pixel landmark coordinates. The task-specific function g_{t_i} then computes the corresponding biometric parameter.

The output consists of task-specific landmarks and their derived biometric values rather than a generic segmentation mask or class label.

4 Experiments

4.1 Experimental Setting

We used a fixed internal split of the data available for the MICCAI 2026 Foundation Model Challenge for Ultrasound Biometry. It covers A4C, AOP, FA, FUGC, HC, IVC, PLAX, PSAX, and fetal femur, each with its own landmark template.

The submitted-model result is from the official challenge test set; all other results are internal development or controlled-evaluator results unless stated otherwise. The available records do not provide per-task sample or landmark counts.

Let P_{itk} and $\hat{P}_{itk}$ denote the reference and predicted coordinates of landmark k in image i from task t. The task-level radial error is the mean Euclidean distance $\|\hat{P}_{itk} - P_{itk}\|_2$ over valid landmarks and images. Macro-validation MRE averages these task-level errors with equal weight across the nine tasks. This metric was calculated in the common resized evaluation space and used for checkpoint selection. Original-scale MRE applies the same calculation after inverse resizing. The controlled evaluator additionally reports MAE as the mean absolute deviation over its valid scalar coordinate entries after inverse resizing; both controlled metrics are therefore reported in pixels. The official test MRE and MAE are the values returned directly by the challenge evaluator and were not recomputed locally. Because the official, resized internal, and controlled original-scale records use different evaluation spaces and aggregation pipelines, their absolute values should not be compared directly.

Images were resized to 512×512. A task-stratified 80:20 split with random seed 42 produced 5,414 labelled training images and 1,354 validation images. Pseudo-label training additionally used 191,170 unlabelled images. Eight tasks used semi-supervised training. Fetal femur had no unlabelled images and therefore used a fully supervised fallback in every semi-supervised configuration. Targets were 64×64 Gaussian heatmaps with $\sigma = 1.8$. The backbone was DINOv3 ViT-S/16 (`vit_small_patch16_dinov3.lvd1689m`). Teacher and student started from the same checkpoint. Gradient descent updated the student encoder and heads, while Eq. (2) updated the teacher.

We trained for 40 epochs with AdamW and cosine annealing ($T_{\max} = 40$, $\eta_{\min} = 10^{-6}$). The overall, backbone, and head learning rates were 10^{-4}, 2×10^{-5}, and 10^{-3}, respectively. Labelled and unlabelled batch sizes were both 4. Augmentation comprised shift-scale-rotation, brightness and contrast jitter, Gaussian noise, normalization, and tensor conversion. The consistency weight, confidence threshold, angle-loss weight, parameter-loss weight, and EMA decay were 0.3, 0.0, 0.05, 0.10, and 0.99, respectively. Thus, the verified experiments did not apply non-zero confidence filtering. Historical trials with thresholds of 0.1–0.3 were excluded because they lacked evaluation with the common pipeline. DARK was used at inference unless otherwise stated.

The checkpoint with the lowest validation MRE was retained and used for the official submission. Original-scale evaluation was performed separately to assess pixel-domain localization after mapping predictions to source-image coordinates.

4.2 Overall Performance

The official challenge test evaluator returned an MRE of 24.96 and an MAE of 27.48 for the retained submitted checkpoint (Table 1). These are official test-set outputs rather than public or internal validation scores. They are reported separately from the internal and controlled evaluations because their coordinate spaces and evaluation pipelines differ.

Table 1. Official challenge test result for the submitted model.

Submission	MRE ↓	MAE ↓
Proposed model	**24.96**	**27.48**

Table 2. Principal comparison on the fixed internal validation split.

Model / setting	Backbone	Best epoch	Best MRE ↓	Final MRE ↓
DINOv2 supervised baseline	DINOv2	–	21.4877	–
Proposed model	DINOv3	31	**18.4827**	**18.6927**

MRE denotes macro-validation mean radial error.

The proposed model attained a best macro-val MRE of 18.4827 at epoch 31 and an MRE of 18.6927 at the final epoch (Table 2). Relative to the recorded DINOv2 supervised baseline, the best error decreased by 3.0050, or 13.98%. This internal comparison is not numerically comparable with the official test result because the evaluation spaces differ.

4.3 Effect of Semi-supervised Learning

Table 3. Verified original-scale comparison of backbone and training strategy using the common split and evaluator. Values are in pixels.

Configuration	MRE ↓	MAE ↓
DINOv2 supervised	63.346	110.225
DINOv2 semi-supervised[†]	**38.644**	**73.916**
DINOv3 supervised	57.616	112.650

[†]Eight tasks used semi-supervised learning; fetal femur used the fully supervised fallback because no unlabelled images were available.

Under the matched DINOv2 comparison, semi-supervised training reduced MRE by 24.702 pixels and MAE by 36.309 pixels (Table 3). Supervised DINOv3 reduced MRE by 5.730 pixels relative to supervised DINOv2, although its MAE was 2.425 pixels higher. This comparison indicates a backbone benefit in radial localization error, but not in MAE; it does not by itself separate the backbone contribution from that of semi-supervised training in the full system.

4.4 Decoder Comparison

For identical heatmaps, DARK marginally improved MRE by 0.255 pixels and MAE by 0.268 pixels relative to argmax (Table 4). Soft-argmax reduced MAE relative to both discrete decoders but increased MRE to 100.471 pixels. DARK therefore offered a small localization benefit for this checkpoint rather than accounting for the overall performance gain.

Table 4. Pixel-domain decoder comparison for the same DINOv3 supervised checkpoint and predicted heatmaps.

Decoder	MRE ↓	MAE ↓
Argmax	57.871	112.918
Soft-argmax	100.471	**92.984**
DARK	**57.616**	112.650

4.5 Task-level Results and Failure Cases

Table 5. Per-task MRE of the proposed model, sorted from lowest to highest.

Domain	Task	MRE ↓
Intrapartum	AOP	6.202
Prenatal	FUGC	7.339
Prenatal	FA	9.056
Cardiac	PLAX	15.299
Prenatal	HC	15.482
Prenatal	Fetal femur	23.357
Cardiac	PSAX	24.159
Cardiac	A4C	27.542
Cardiac	IVC	37.907

Task-level errors varied markedly (Table 5). AOP and FUGC had the lowest MREs, at 6.202 and 7.339, whereas IVC had the highest at 37.907. A4C, PSAX, and fetal femur were also among the more difficult tasks.

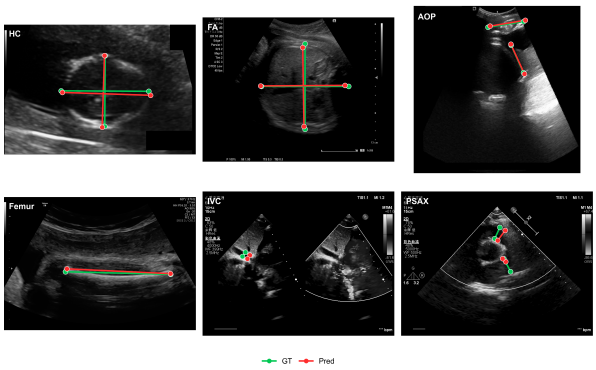

Fig. 2. Qualitative comparison between ground-truth landmarks (green) and predictions (red) on representative prenatal, intrapartum, and cardiac tasks.

Figure 2 illustrates predictions for HC, FA, AOP, femur, IVC, and PSAX. Residual offsets are most apparent in the IVC and PSAX examples, consistent with their higher task-level errors.

5 Discussion

The official test result provides evidence beyond the internal development split. In the verified original-scale comparison, semi-supervised DINOv2 substantially outperformed its supervised counterpart. Supervised DINOv3 also improved MRE relative to supervised DINOv2, although its MAE was slightly higher. These findings support contributions from both the backbone and unlabelled data, but the available comparisons do not quantify their independent effects in the full system. Matched full-model ablations were not available after filtering experiments by code version and evaluation protocol, so the independent effects of EMA, CBAM, and the measurement-aware losses remain unresolved.

The decoder comparison places DARK in proportion to the training strategy. DARK was slightly better than argmax on both metrics for the same checkpoint, whereas soft-argmax traded lower MAE for markedly higher MRE. DARK therefore provides a modest coordinate-refinement benefit rather than explaining the larger differences between supervised and semi-supervised training.

The largest errors occurred in IVC and several cardiac tasks, where short structures make the measurements sensitive to small endpoint displacements. These results require cautious interpretation. Although the submitted model was evaluated on the official challenge test set, model development and checkpoint selection used a single internal split. Per-task test results, per-task baseline and biometric-MAE comparisons, and an explicit uncertainty model are unavailable. The verified experiments used a confidence threshold of 0.0; historical non-zero filtering trials were not evaluated through the same pipeline and therefore do not support a quantitative conclusion.

Future work should evaluate pseudo-label confidence, repeated internal splits, and task-level official-test outputs. Uncertainty-aware endpoint refinement for cardiac and vascular tasks also requires independent testing.

6 Conclusion

We adapted a pretrained DINOv3 encoder to landmark-based biometry across prenatal, intrapartum, and cardiac ultrasound. The submitted model achieved an official challenge test MRE of 24.96 and MAE of 27.48. On a fixed internal split, the semi-supervised system reached a best macro-val MRE of 18.4827. A matched DINOv2 comparison showed lower error with semi-supervised than supervised training, whereas DARK provided only a small improvement over argmax for identical heatmaps. The evidence supports the use of unlabelled data, but matched component ablations, repeated internal splits, and per-task test metrics remain unavailable.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Grant No. 82502508), in part by the Explorer Program of Shanghai (Grant No. 26TS1403100), the Central Government Guides Local Science and Technology Development Fund Projects (25ZYJA017), and in part by the Youth Science and Technology Talent Innovation Project of Lanzhou City (Grant No. 2025-QN-037).

References

1. MICCAI 2026 Foundation Model Challenge for Ultrasound Biometry organizers: Challenge design document: Foundation Model Challenge for Ultrasound Biometry. Zenodo, https://doi.org/10.5281/zenodo.19736827 (2026)
2. MICCAI 2026 Foundation Model Challenge for Ultrasound Biometry organizers: Foundation Model Challenge for Ultrasound Biometry Codabench evaluation platform. `https://www.codabench.org/competitions/15590/`, last accessed 2026/06/20
3. Salomon, L.J., Alfirevic, Z., Da Silva Costa, F., Deter, R.L., Figueras, F., Ghi, T., Glanc, P., Khalil, A., Lee, W., Napolitano, R., Papageorghiou, A., Sotiriadis, A., Stirnemann, J., Toi, A., Yeo, G.: ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound in Obstetrics & Gynecology **53**(6), 715–723 (2019)
4. Ghi, T., Eggebø, 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)
5. Lang, R.M., Badano, L.P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F.A., Foster, E., Goldstein, S.A., Kuznetsova, T., Lancellotti, P., Muraru, D., Picard, M.H., Rietzschel, E.R., Rudski, L., Spencer, K.T., Tsang, W., Voigt, J.U.: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. Journal of the American Society of Echocardiography **28**(1), 1–39.e14 (2015)
6. Rudski, L.G., Lai, W.W., Afilalo, J., Hua, L., Handschumacher, M.D., Chandrasekaran, K., Solomon, S.D., Louie, E.K., Schiller, N.B.: Guidelines for the echocardiographic assessment of the right heart in adults: a report from the American Society of Echocardiography. Journal of the American Society of Echocardiography **23**(7), 685–713 (2010)
7. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Transactions on Medical Imaging **36**(11), 2204–2215 (2017)
8. van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., Ginneken, B.: Automated measurement of fetal head circumference using 2D ultrasound images. PLoS ONE **13**(8), e0200412 (2018)
9. Zhang, L., Dudley, N.J., Lambrou, T., Allinson, N., Ye, X.: Direct estimation of fetal head circumference from ultrasound images based on regression CNN. In: Proceedings of the Third Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research, vol. 121, pp. 914–922 (2020)
10. Bai, J., Zhou, Z., Tang, Y., Gan, J., Liang, Z., Fan, J., McGuire, L.B., Clarke, J.L., Cai, W., Spurway, J., Tan, Y., Wang, S., Shen, W., Yu, W., Li, Y., Zhang, P., Jiang,

W., Li, Y., Al Nasi, S.M.A.B., et al.: IUGC: a benchmark of landmark detection in end-to-end intrapartum ultrasound measurement. Medical Image Analysis **110**, 103960 (2026)
11. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**, 252–256 (2020)
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. arXiv:2304.07193 (2023)
13. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Darcet, T., Moutakanni, T., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P., et al.: DINOv3. arXiv:2508.10104 (2025)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024)
15. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
16. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, vol. 30, pp. 1195–1204 (2017)
17. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac MR image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017, pp. 253–260. Springer (2017)
18. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems, vol. 33, pp. 596–608 (2020)
19. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7093–7102 (2020)