

Rethinking Medical Landmark Localization with Prototype Learning-based Progressive Offset Correction

Jingxian Xu^{1*}, Yuhao Huang^{1,2,3*}, Rusi Chen¹, Yanfeng Zhou³, and
Dong Ni^{1,4,5,6} (✉)

¹ Medical Ultrasound Image Computing (MUSIC) Lab, Shenzhen University,
Shenzhen, China
nidong@szu.edu.cn

² Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science &
Innovation, Chinese Academy of Sciences, Hong Kong, China

³ Boston Children’s Hospital, Harvard Medical School, Boston, USA

⁴ School of Artificial Intelligence, Shenzhen University, Shenzhen, China

⁵ School of Biomedical Engineering and Informatics, Nanjing Medical University,
Nanjing, China

⁶ National Engineering Laboratory for Big Data System Computing Technology,
Shenzhen University, Shenzhen, China

Abstract. Accurate landmark localization in medical images is a fundamental step for quantitative clinical measurement and downstream analysis. Existing localization methods have advanced, among which multi-stage refinement is a superior solution. Although this strategy mitigates the anatomical ambiguity inherent in single-stage global predictions, its high computational cost limits practical applicability. In this work, we propose a parameter-economic model, PPOC-LL, which leverages **Prototype learning-based Progressive Offset Correction for Landmark Localization**. Our contribution is three-fold. **First**, to drive coarse-to-fine landmark optimization, we introduce a multi-scale dynamic perception strategy for patch-level feature pyramid modeling. **Second**, to effectively handle anatomically similar patterns, we design a similarity-driven prototype learning mechanism that captures informative local semantics for robust offset prediction. **Last**, to stabilize the model learning and improve the overall performance, we incorporate a novel error-aware reliability regularization via tolerance-based balancing. We collected a large validation cohort, including two public and one private datasets spanning X-ray and ultrasound modalities, covering cephalometric, symphysis-fetal head, and fetal heart landmarks. Extensive experiments demonstrate that PPOC-LL achieves satisfactory performance with a favorable trade-off between accuracy and model complexity.

Keywords: Landmark localization · Multi-scale · Prototype Learning.

* Jingxian Xu and Yuhao Huang contributed equally to this work.

1 Introduction

Landmark localization constitutes a core element of quantitative measurements in diverse clinical applications [25,2], including craniofacial evaluation, intra-partum fetal monitoring, and fetal cardiac assessment, etc. However, manual annotation is time-consuming and labor-intensive. Moreover, anatomical ambiguity and inter-observer variability may introduce systematic bias, potentially compromising downstream measurements and clinical decision-making. Therefore, robust automated localization methods are urgently needed.

Recent deep learning-based studies have made significant progress in intelligent medical image analysis [8,3,16,10,9]. Specifically, for landmark localization, regression-based methods are one of the most common solutions, and can be broadly categorized into direct coordinate prediction and heatmap-based formulations. The former predicts absolute landmark coordinates in an end-to-end manner [27], while the latter estimates per-landmark heatmaps and extracts coordinates from their peaks [4]. Zhou et al. [30] explored metric learning to strengthen the representation capability of heatmap-based landmark detectors. Then, subsequent efforts integrated topology-aware constraint [11] and uncertainty quantification [12] were further leveraged to refine the quality and reliability of heatmap predictions. However, most of them rely primarily on global information and perform prediction in a single forward pass, lacking explicit modeling of local contextual information around each landmark.

To address the above issues, several studies have investigated iterative approaches that incorporate local contextual information, with representative approaches including reinforcement learning (RL) and cascaded models. Early RL typically learn a policy to sequentially navigate toward target landmarks within medical volumes [1]. Alternatively, cascaded models first generate coarse predictions and subsequently refine them using local cues or higher-resolution representations. Alignment-proposal-refinement joint learning [29] and discretized ordinary differential equation formulation [7] have been specifically designed for cephalometric landmark localization. Subsequently, Khan et al. [13] extended the two-stage cascaded model to multi-resolution and multi-modal datasets. In addition, a cascaded iterative Transformer was proposed for facial landmark detection under large pose and occlusion [15]. However, RL-based methods may encounter training difficulties, and cascaded network designs typically increase model complexity, potentially limiting their applicability in clinical environments.

In this study, we propose the Prototype learning-based Progressive Offset Correction framework for Landmark Localization (named PPOC-LL). Our highlights are as follows. First, we develop a multi-scale feature perception mechanism to extract feature pyramids and obtain initial landmark predictions. These cues facilitate coarse-to-fine attention modeling and enable patch-level progressive refinement. Second, we introduce a prototype learning scheme combined with local matching to capture vital semantic representations, enabling robust offset estimation. Last, we design a tolerance-guided regularization that ensures stable model learning and improved detection accuracy. Experiments on three medical datasets spanning two imaging modalities show that PPOC-LL outperforms

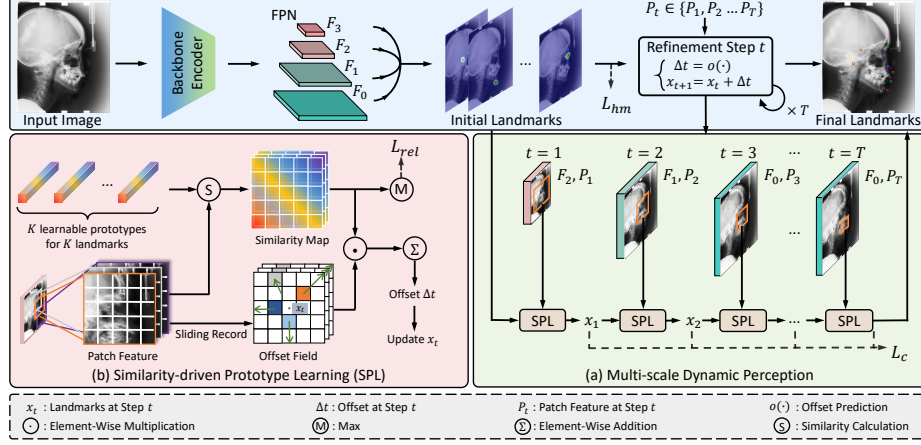


Fig. 1: Illustration of our proposed PPOC-LL.

state-of-the-art methods. Moreover, PPOC-LL effectively balances performance and model complexity, indicating strong potential for clinical applicability.

2 Method

Fig. 1 shows our PPOC-LL framework, which treats landmark localization as progressive correction via offset prediction. First, a regular backbone equipped with a feature pyramid network (FPN) outputs multi-scale knowledge and initial landmarks via *soft-argmax*. Then, the proposed multi-scale patch-based dynamic perception and similarity-driven prototype learning strategies capture the offset relationships and update the landmark position progressively. Besides, an error-aware reliability loss is equipped to stabilize the training process.

2.1 Multi-scale Dynamic Perception for Progressive Optimization

Existing single-pass landmark localization relied on global context and struggled to resolve local anatomical similarity. Although FPN provides multi-scale representations, a fixed-scale sampling strategy is suboptimal for progressive refinement, since early corrections require a larger search range while later updates benefit from finer spatial details. Therefore, we build PPOC-LL on top of an FPN with feature maps F_0 - F_3 , and perform iterative refinement by sampling multi-scale patches in a coarse-to-fine schedule. Specifically, each refinement step t samples a compact $p_t \times p_t$ feature patch centered at the current estimate x_t on the selected pyramid scale to decode a displacement for location update.

In our setting, early steps rely on lower-resolution features with larger receptive fields for coarse correction, whereas later steps refine landmarks using higher-resolution features for improved precision. The patch is always centered

at the x_t , yielding step-specific updates that form a continuous correction trajectory. This step-wise re-centering of the sampled patch around x_t is termed as a *sliding record*. We define the perception scale at step t using an input image space search radius R_t , and map it to a selected pyramid level l_t from $\{F_0, F_1, F_2\}$. The conversion factor between the input and the l_t -th feature is:

$$s_{l_t} = (W - 1)/(W_{l_t} - 1), \quad (1)$$

where W and W_{l_t} denote spatial resolutions of input image and the l_t -th pyramid, respectively. With the *ceiling* function, the patch size can be formulated as:

$$p_t = 2 \lceil R_t/s_{l_t} \rceil + 1. \quad (2)$$

2.2 Similarity-driven Prototype Learning for Offset Correction

In iterative refinement, each update is driven by evidence collected around the current landmark location. However, relying solely on local responses can be unstable under appearance variations across subjects and acquisition conditions, where the same landmark may exhibit diverse textures and different landmarks may appear similar. To enhance discriminability, we associate each landmark with a learnable prototype vector as a semantic reference.

Specifically, at step t , the local patch feature for landmark k is defined as $P_{t,k}$, centered on $x_{t,k}$. We then project the $P_{t,k}$ to the prototype dimension using 1×1 convolution (*conv*) and compute a cosine similarity $M_{t,k}(i, j)$ map between the learnable prototype q_k and each spatial location (i, j) in the patch, as follows:

$$M_{t,k}(i, j) = \cos(q_k, \text{conv}(P_{t,k}(i, j))). \quad (3)$$

We then compute a normalized local matching distribution $m_{t,k}(i, j)$ by applying a softmax over the similarity map $M_{t,k}(i, j)$ across spatial locations in the patch. Finally, we decode the displacement as the expectation over a predefined offset field $\Delta(i, j)$ on the sampled patch, and update the landmark locations:

$$\begin{aligned} \Delta_{t,k} &= \sum_{i,j} m_{t,k}(i, j) \Delta(i, j), \\ x_{t+1,k} &= x_{t,k} + \Delta_{t,k}, \end{aligned} \quad (4)$$

where $\Delta(i, j) = (\Delta x, \Delta y)$ is fixed and computed from the patch's relative indices. We set $\Delta x = (j - \lfloor p_t/2 \rfloor) \cdot 2/(W_{l_t} - 1)$ and $\Delta y = (i - \lfloor p_t/2 \rfloor) \cdot 2/(H_{l_t} - 1)$, where (H_{l_t}, W_{l_t}) is the spatial resolution of the feature map selected at step t , so that $\Delta(i, j)$ is expressed in the same normalized coordinate system as $x_{t,k}$.

2.3 Error-aware Reliability Regularization for Stable Training

We observe that the sharpness of the local matching distribution is not always calibrated to the underlying localization error during iterative refinement. Therefore, coordinate-only supervision can not explicitly enforce consistency between

the distribution confidence and the actual localization error, which can destabilize training. To address this issue, we propose an error-aware reliability regularizer to provide a step-wise scalar reliability signal. To be specific, the peak probability of local matching distribution $m_{t,k}$ serves as the predicted reliability.

$$r_{t,k} = \max_{i,j} m_{t,k}(i,j) \in [0, 1]. \quad (5)$$

Then, we derive the soft supervision target $\tilde{r}_{t,k}$ from the Euclidean localization error between the current estimate and the ground truth x_k^* :

$$\tilde{r}_{t,k} = 1 - \text{clip}\left(\frac{\|x_{t,k} - x_k^*\|_2}{\gamma}, 0, 1\right). \quad (6)$$

When constructing the target, we detach $x_{t,k}$ so that this regularizer supervises the reliability prediction only and does not backpropagate to the coordinate updates. The $\text{clip}(\cdot, 0, 1)$ operator clamps the normalized error to $[0, 1]$, ensuring $\tilde{r}_{t,k} \in [0, 1]$, where γ is the tolerance parameter specifically defined by different datasets to ensure geometric consistency under multi-scale resizing. This yields a monotonic and continuous relationship between localization error and reliability, casting the supervision as error regression rather than a hard threshold classification. Finally, the reliability regularization term can be defined by:

$$\mathcal{L}_{\text{rel}} = \sum_{t=1}^T \sum_{k=1}^K |r_{t,k} - \tilde{r}_{t,k}|. \quad (7)$$

We train the model with three losses: (i) a typical heatmap regression loss $\mathcal{L}_{hm}$ based on mean squared error, where the target heatmaps are generated from the ground-truth coordinates; (ii) a step-wise coordinate loss $\mathcal{L}_c$ that applying L1 penalties to each x_t with larger weights on later steps; and (iii) the above-mentioned tolerance-balanced regularization $\mathcal{L}_{rel}$. Finally, the total loss $\mathcal{L}$ is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{hm} + \lambda_2 \mathcal{L}_c + \lambda_3 \mathcal{L}_{rel}. \quad (8)$$

3 Experiments

Datasets and Evaluation Metrics. We collected two public datasets (Cephalograms [25], CE and Intrapartum Ultrasound Grand Challenge, IUGC [2]) and one private dataset (fetal heart ultrasound, FHU) for method validation. *CE* comprises 400 lateral cephalogram images, each annotated with 19 landmarks. The dataset was officially split into 150, 150 and 100 images for training, test1 and test2, respectively. *IUGC* includes 801 mid-sagittal view images, with 3 landmarks on the pubic symphysis and fetal head annotated for Angle of Progression (AoP) calculation. Following the fully-supervised setting, we use 300 and 501 images for training and testing. *FHU* contains 905 mid-gestation four-chamber planes annotated with 24 anatomical landmarks, of which 724 are used for training and 181 for testing. For *CE* and *FHU*, Mean Radial Error (MRE)

Table 1: Quantitative comparison on CE, FHU, and IUGC datasets. The best results are shown in **bolded**, and the second-best ones are underlined.

CE										
Methods	Test1					Test2				
	MRE↓	SDR↑				MRE↓	SDR↑			
		2mm	2.5mm	3mm	4mm		2mm	2.5mm	3mm	4mm
CenterNet [31]	2.13 _{1.41}	55.43	70.32	80.57	90.49	2.49 _{1.95}	47.68	62.53	73.72	84.42
Multi-Reg [14]	1.19 _{0.80}	86.42	92.00	95.54	<u>98.53</u>	-	74.58	81.79	87.53	94.26
APR [29]	1.34 _{0.92}	81.37	89.09	93.79	97.86	1.64 _{0.91}	70.58	79.53	86.05	93.32
ContextNet [22]	1.18 _{1.01}	86.20	91.20	94.40	97.70	<u>1.46</u> _{0.82}	75.90	83.40	<u>89.30</u>	94.70
AFPF-RV [5]	1.17	86.67	<u>92.67</u>	95.54	<u>98.53</u>	1.48	75.05	82.84	88.53	95.05
CH-Net [19]	1.20	83.47	89.16	92.60	96.49	<u>1.46</u>	74.63	<u>83.58</u>	87.21	93.79
NFDP [11]	<u>1.14</u> _{0.93}	<u>87.02</u>	92.38	<u>95.76</u>	98.35	<u>1.46</u> _{1.31}	75.11	82.48	89.16	<u>95.16</u>
PPOC-LL	1.09 _{<u>0.87</u>}	88.49	93.51	95.89	98.53	1.43 _{1.25}	<u>75.79</u>	83.74	89.42	95.53

FHU					IUGC							
Methods	MRE↓	SDR↑				Strategy	Methods	MRE↓	MRE (<i>PS_R</i>)	MRE (<i>PS_L</i>)	MRE (<i>FH_T</i>)	ΔAoP (°)
		3px	4px	6px	8px							
IntegralNet [23]	9.73 _{8.86}	11.35	18.14	34.28	49.70	Semi-supervised	T1 [17]	13.16 _{6.25}	8.56 _{5.01}	9.19 _{6.77}	21.73 _{16.05}	4.42 _{4.79}
DSNT [21]	8.92 _{8.25}	15.12	23.45	40.80	55.90		T2 [18]	11.67 _{6.34}	6.53 _{4.38}	8.60 _{5.06}	19.90 _{17.54}	3.81 _{3.12}
HRNet [26]	9.95 _{9.15}	8.75	15.65	30.57	47.01		T8 [6]	12.82 _{6.14}	7.81 _{5.15}	9.25 _{7.64}	21.42 _{15.88}	4.57 _{7.90}
SHG [20]	10.30 _{9.62}	8.66	14.71	30.02	45.86	Fully-supervised	T4 [28]	<u>14.77</u> _{5.91}	9.06 _{5.70}	<u>11.57</u> _{7.53}	<u>23.69</u> _{14.83}	<u>4.70</u> _{5.26}
NFDP [11]	<u>7.57</u> _{6.62}	22.19	32.83	51.93	66.85		T10 [24]	21.83 _{19.89}	10.67 _{9.27}	15.62 _{25.74}	39.19 _{51.87}	8.37 _{14.64}
PPOC-LL	7.34 _{6.62}	<u>22.18</u>	34.12	53.91	67.70		PPOC-LL	12.28 _{<u>12.30</u>}	8.12 _{5.46}	9.49 _{<u>9.54</u>}	19.23 _{<u>16.18</u>}	4.33 _{4.54}

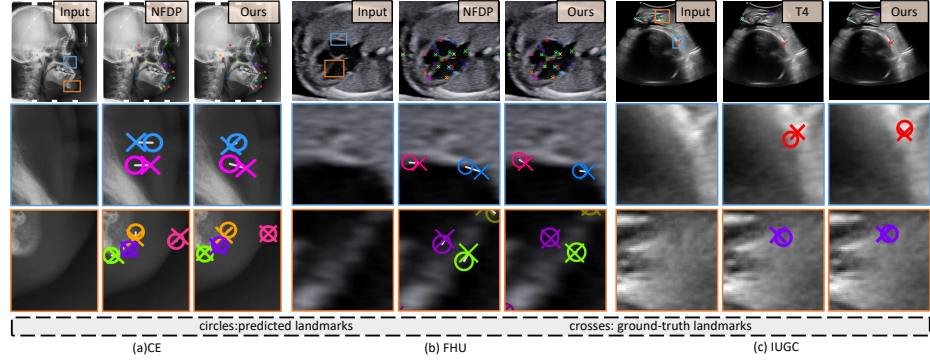


Fig. 2: Qualitative results on CE, FHU, and IUGC datasets. Rows 2-3 show zoomed-in patches (see Row 1) for clearer landmark comparison.

and Successful Detection Rate (SDR) were selected as the evaluation metrics. The SDR thresholds were set to 2, 2.5, 3, and 4 mm for *CE*, and 3, 4, 6, and 8 pixels for *FHU*. For *IUGC*, we follow [2] to report MRE and also landmark-level statistics ($PS_R/PS_L/FH_t$), with AoP error for clinical measurement evaluation.

Implementation Details. We implement our model in PyTorch, using a single NVIDIA RTX 4090 GPU. ResNet18 is used as the backbone for all models. AdamW was used to optimize model training with a weight decay of $1e-4$, a batch size of 16, and 400 epochs. A cosine annealing learning rate schedule

Table 2: Effectiveness of components on CE dataset, including Refinement Paradigm (RP) and Error-aware Regularization (ER). The best and second-best results are **bolded** and underlined, respectively.

<i>RP</i>	<i>ER</i>	MRE↓	SDR ↑			
			2.0mm	2.5mm	3.0mm	4.0mm
		1.37	78.48	86.04	91.05	96.46
✓		1.31	80.38	<u>87.50</u>	<u>91.91</u>	<u>97.06</u>
	✓	<u>1.30</u>	80.07	87.45	91.69	96.67
✓	✓	1.26	82.14	88.60	92.78	97.12

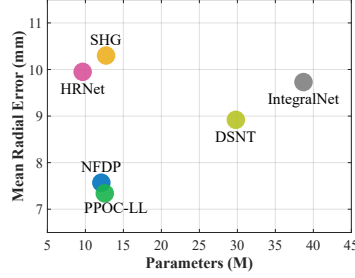


Fig. 3: Comparison of different methods in terms of MRE and model size.

Table 3: Comparison of different iteration step T and tolerance parameter γ on CE-Test1 dataset. Best results are **bolded**, second-best are underlined.

T	MRE↓	SDR ↑			
		2.0mm	2.5mm	3.0mm	4.0mm
1	1.47 _{0.96}	76.95	87.19	93.23	97.93
3	1.15 _{<u>0.93</u>}	86.35	92.42	95.54	98.42
4	<u>1.10</u> 0.87	88.04	<u>93.12</u>	95.82	<u>98.53</u>
5	1.09 0.87	88.49	93.51	<u>95.89</u>	<u>98.53</u>
6	<u>1.10</u> 0.87	<u>88.18</u>	92.98	96.00	98.60

γ	MRE↓	SDR ↑			
		2.0mm	2.5mm	3.0mm	4.0mm
2	1.09 _{<u>0.88</u>}	87.72	<u>93.16</u>	95.75	<u>98.53</u>
3	1.09 0.87	88.49	93.51	95.89	<u>98.53</u>
5	<u>1.11</u> _{<u>0.88</u>}	87.47	92.60	96.18	98.50
7	1.13 _{0.91}	<u>87.86</u>	92.84	95.86	98.74
10	1.16 _{<u>0.88</u>}	86.77	92.53	95.47	97.96

is adopted with an initial learning rate of $2e-4$. Data augmentation includes flipping, rotation, and transposition. Input images are resized to 512×512 for *CE* and *IUGC*, and to 256×256 for *FHU*. We set the input image space search radius R_t to $\{112, 80, 48, 48, 28\}$ for 512×512 inputs, and to $\{119, 56, 20, 20, 12, 8\}$ for 256×256 inputs. We use $\lambda_1 = 1$, $\lambda_2 = 20$, and $\lambda_3 = 0.2$ for the loss weights. Based on validation performance, we set the refinement steps $T=5, 6, 6$ and tolerance parameter $\gamma=3, 3, 8$ for *CE*, *IUGC*, and *FHU*, respectively. For PPOC-LL and ablation variants, experiments were repeated with three random seeds, and the averaged results are reported. For MRE, the small values shown in the right-down corner indicate the standard deviation of radial errors on the test set.

Comparisons with State-of-the-art Models. Table 1 compares PPOC-LL with various strong methods. Specifically, for CE and FHU, we included 7 and 5 competitors, respectively. For IUGC, we report results in both semi-supervised (T1/T2/T8) and fully-supervised (T4/T10) settings [2]. Results on CE and FHU indicate that PPOC-LL consistently outperforms most methods, ranking first or second across the evaluated metrics. Specifically, PPOC-LL reports MREs of 1.09 mm, 1.43 mm, and 7.34 px on CE-Test1, CE-Test2, and FHU, respectively, surpassing the performance of the state-of-the-art NFDP. Across the two CE test sets, PPOC-LL exhibits lower standard deviations (0.87/1.25) than NFDP (0.93/1.31), demonstrating superior cross-sample robustness. On IUGC, PPOC-

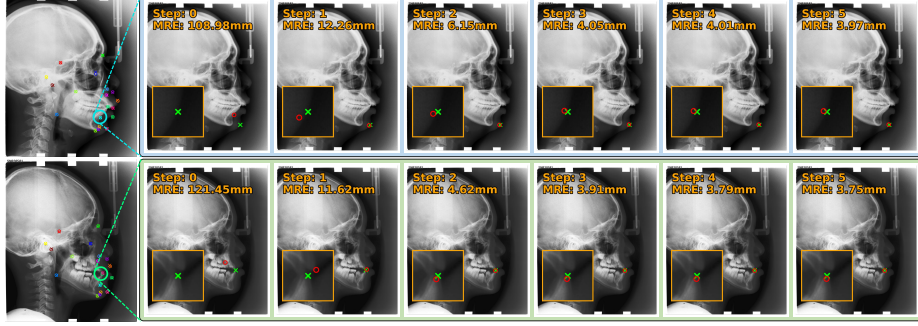


Fig. 4: Visualization of the refinement process on the CE dataset. The lower-left zoomed-in panels show the progressive convergence of predicted landmarks (circles) toward the ground truths (crosses) over iterations.

LL ranks first among fully supervised methods and, despite not using any unlabeled data, achieves performance comparable to semi-supervised approaches. These gains strongly demonstrate the effectiveness of our overall design, including prototype-guided local matching and progressive offset correction. Fig. 2 visualizes that the predictions from our PPOC-LL method are closer to the ground truth points. Fig. 3 illustrates the trade-off between MRE and model size. It shows that our PPOC-LL obtains the lowest MRE among all methods, specifically outperforming NFDG with only a slight increase in parameters.

Ablation Studies. As shown in Table 2, compared with directly predicting landmarks in a single pass on the finest pyramid feature, introducing the progressive refinement paradigm can reduce the MRE from 1.37 mm to 1.31 mm and improve SDR across all thresholds, with the 2.0 mm SDR increasing from 78.48% to 80.38%. Incorporating the error-aware regularization will also decrease the MRE to 1.30 mm with consistent SDR gains. Equipped with both components can bring further improvement, achieving the lowest MRE of 1.26 mm and the highest SDR (82.14% at 2.0 mm, 88.60% at 2.5 mm, 92.78% at 3.0 mm, and 97.12% at 4.0 mm). These results suggest that the proposed strategies provide complementary benefits, leading to precise and stable landmark localization.

Table 3 further evaluates the impact of the iteration steps T and the tolerance parameter γ on model performance using the CE dataset. Results show that increasing T substantially improves both MRE ($\sim 0.38\downarrow$) and SDR ($\sim 12\uparrow$, 2.0 mm), with the performance gradually plateauing around $T = 5/6$. We also present cephalometric examples in Fig. 4 to show the refinement process. These results demonstrate the effectiveness of our designed multi-scale iterative refinement and suggest an optimal number of iteration steps. For different γ , PPOC-LL exhibits high robustness, with MRE remaining stable within a narrow range (1.09-1.16 mm) as γ varies from 2 to 10. Such insensitivity shows that our regularization strategy can stabilize training without extensive hyperparameter tuning.

4 Conclusion

In this work, we propose a novel framework, named PPOC-LL, for automated landmark localization that integrates progressive offset correction with prototype learning across three medical datasets. Specifically, we first design a patch-level dynamic perception strategy to exploit multi-scale feature pyramids and iteratively refine landmarks in a coarse-to-fine manner. At each iteration, prototype learning with similarity-based matching is employed to capture local cues and predict robust offsets. Finally, an error-aware regularization is introduced to stabilize training and further improve localization performance. In future work, we plan to extend our method to more organs, modalities, and 3D scenarios.

Acknowledgments. This work was supported by the Frontier Technology Development Program of Jiangsu Province (No. BF2024078) and National Natural Science Foundation of China (No. 12326619).

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

References

1. Alansary, A., Oktay, O., Li, Y., et al.: Evaluating reinforcement learning agents for anatomical landmark detection. *Medical image analysis* **53**, 156–164 (2019)
2. Bai, J., Tang, Y., et al.: Iugc: A benchmark of landmark detection in end-to-end intrapartum ultrasound biometry. *Medical Image Analysis* p. 103960 (2026)
3. Chen, C., Yang, X., Huang, R., et al.: Region proposal network with graph prior and iou-balance loss for landmark detection in 3d ultrasound. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2020)
4. Chen, J., Che, H., et al.: An automatic cephalometric landmark detection method based on heatmap regression and monte carlo dropout. In: 2023 45th Annual International Conference of the IEEE EMBC. pp. 1–4. IEEE (2023)
5. Chen, R., Ma, Y., et al.: Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 873–881. Springer (2019)
6. Deng, B., Chen, Y., Peng, Z.: A two-stage semi-supervised ensemble framework for automated angle of progression measurement in intrapartum ultrasound. In: Intrapartum Ultrasound Grand Challenge, pp. 88–99. Springer (2025)
7. He, T., et al.: Cascade-refine model for cephalometric landmark detection in high-resolution orthodontic images. *Knowledge-Based Systems* **265**, 110332 (2023)
8. Hu, X., Wang, L., Yang, X., Zhou, X., Xue, W., et al.: Joint landmark and structure learning for automatic evaluation of developmental dysplasia of the hip. *IEEE Journal of Biomedical and Health Informatics* **26**(1), 345–358 (2021)
9. Huang, Y., Chang, A., Dou, H., et al.: Flip learning: Weakly supervised erase to segment nodules in breast ultrasound. *Medical Image Analysis* **102**, 103552 (2025)
10. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., et al.: Segment anything model for medical images? *Medical Image Analysis* **92**, 103061 (2024)
11. Huang, Z., et al.: Landmark localization from medical images with generative distribution prior. *IEEE transactions on medical imaging* **43**(7), 2679–2692 (2024)

12. Jonkers, J., Coopman, F., et al.: Reliable uncertainty quantification for 2d/3d anatomical landmark localization using multi-output conformal prediction. *Medical Image Analysis* p. 103953 (2026)
13. Khan, R., Khalid, M.A., et al.: Enhancing cephalometric landmark detection with a two-stage cascaded cnn on multi-resolution multi-modal data. In: *Annual conference on medical image understanding and analysis*. pp. 3–18. Springer (2024)
14. Lee, M., Chung, M., Shin, Y.G.: Cephalometric landmark detection via global and local encoders and patch-wise attentions. *Neurocomputing* **470**, 182–189 (2022)
15. Li, Y., Tan, G., Gou, C.: Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. *International Journal of Computer Vision* **132**(4), 1242–1257 (2024)
16. Liang, J., Yang, X., Huang, Y., Li, H., He, S., et al.: Sketch guided and progressive growing gan for realistic and editable ultrasound image synthesis. *Medical image analysis* **79**, 102461 (2022)
17. Liu, X., Hu, J., Li, Y., Chen, X., Wang, Y.: Noisy student-based self-training enhances landmark detection in intrapartum ultrasound. In: *Intrapartum Ultrasound Grand Challenge*, pp. 1–13. Springer (2025)
18. Ma, C., Li, Y., Guo, B., Jiao, J., Huang, Y., Wang, Y., Guo, Y.: Unlabeled data-driven fetal landmark detection in intrapartum ultrasound. In: *Intrapartum Ultrasound Grand Challenge*, pp. 14–23. Springer (2025)
19. McCouat, J., Voiculescu, I.: Contour-hugging heatmaps for landmark detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20597–20605 (2022)
20. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: *European conference on computer vision*. pp. 483–499. Springer (2016)
21. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372* (2018)
22. Oh, K., Oh, I.S., Lee, D.W., et al.: Deep anatomical context feature learning for cephalometric landmark detection. *IEEE Journal of Biomedical and Health Informatics* **25**(3), 806–817 (2020)
23. Sun, X., Xiao, B., Wei, F., et al.: Integral human pose regression. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 529–545 (2018)
24. Tang, Y., Zhou, Z., Lu, Y., Bai, J., et al.: Heatmap regression for automated angle of progression measurement: The baseline method for the iugc2025. In: *Intrapartum Ultrasound Grand Challenge*, pp. 105–117. Springer (2025)
25. Wang, C.W., Huang, C.T., Lee, J.H., et al.: A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis* **31**, 63–76 (2016)
26. Wang, J., Sun, K., Cheng, T., Jiang, B., et al.: Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **43**(10), 3349–3364 (2020)
27. Wu, Y., Hassner, T., Kim, K., Medioni, G., Natarajan, P.: Facial landmark detection with tweaked convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence* **40**(12), 3067–3074 (2017)
28. Yang, Z., et al.: Dsnt-deepunet: A coordinate prediction method for intrapartum ultrasound. In: *Intrapartum Ultrasound Grand Challenge*, pp. 33–46. Springer (2025)
29. Zeng, M., et al.: Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis* **68**, 101904 (2021)
30. Zhou, G.Q., Miao, J., Yang, X., Li, R., Huo, E.Z., et al.: Learn fine-grained adaptive loss for multiple anatomical landmark detection in medical images. *IEEE Journal of Biomedical and Health Informatics* **25**(10), 3854–3864 (2021)
31. Zhou, X., et al.: Objects as points. *arXiv preprint arXiv:1904.07850* (2019)