Landmarks Are Alike Yet Distinct: Harnessing Similarity and Individuality for One-Shot Medical Landmark Detection

Xu He[1,2*], Zhen Huang[3,4], Qingsong Yao[5], Xiaoqian Zhou[1,2], and S. Kevin Zhou[1,2✉]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230026, P.R.China
skevinzhou@ustc.edu.cn

[2] Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, 215123, P.R.China

[3] School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, 230026, P.R.China

[4] School of Information Science and Technology, Eastern Institute of Technology (EIT), Ningbo, Zhejiang, 315200, P.R.China

[5] Stanford University, Palo Alto, California, 94305, United States

Abstract. Landmark detection plays a crucial role in medical imaging applications such as disease diagnosis, bone age estimation, and therapy planning. However, training models for detecting multiple landmarks simultaneously often encounters the "seesaw phenomenon", where improvements in detecting certain landmarks lead to declines in detecting others. Yet, training a separate model for each landmark increases memory usage and computational overhead. To address these challenges, we propose a novel approach based on the belief that "landmarks are distinct" by training models with pseudo-labels and template data updated continuously during the training process, where each model is dedicated to detecting a single landmark to achieve high accuracy. Furthermore, grounded on the belief that "landmarks are also alike", we introduce an adapter-based fusion model, combining shared weights with landmark-specific weights, to efficiently share model parameters while allowing flexible adaptation to individual landmarks. This approach not only significantly reduces memory and computational resource requirements but also effectively mitigates the seesaw phenomenon in multi-landmark training. Experimental results on publicly available medical image datasets demonstrate that the single-landmark models significantly outperform traditional multi-point joint training models in detecting individual landmarks. Although our adapter-based fusion model shows slightly lower performance compared to the combined results of all single-landmark models, it still surpasses the current state-of-the-art methods while achieving a notable improvement in resource efficiency.

Keywords: Medical landmark detection · One-shot learning.

[*] X. He and Z. Huang contribute equally to this work.

1 Introduction

Accurate medical landmark detection (MLD) has widespread applications in clinical settings, such as disease diagnosis [20, 26, 6], bone age estimation [4], and therapy planning [1, 23]. It also supports various downstream tasks, such as segmentation [14, 17], image reconstruction [11], and image registration [5]. With the rapid development of deep learning [9, 16], many neural network-based models have been proposed for MLD. For instance, [21] employs a prototypical network for MLD by comparing image features with landmark prototypes, while [28] incorporates dynamic sparse attention into a hybrid Transformer-CNN architecture. However, despite the superior performance of these methods, they generally rely on a large amount of labeled data, which poses a significant challenge due to the time-consuming and labor-intensive annotation process. To address this issue, some studies have proposed *one-shot learning* methods, which use a single annotated medical image for landmark detection [24, 13, 25].

MLD is typically formulated as a multi-label task, where traditional methods often train a single model to detect all landmarks, sharing the same network weights and thus ignoring the individuality of different landmarks. Since different landmarks may have distinct features and local variations, training them in the same network may lead to the so-called "seesaw phenomenon" [19], that is, improving the detection of certain landmarks could degrade the performance of others. In multi-task learning [27], some studies have used hard parameter sharing [2] to facilitate joint learning. [8] introduced the Mixture of Experts (MoE) model, which shares some experts at the bottom layers and combines them through a gating network. [19] proposed Progressive Layered Extraction (PLE) to alleviate the seesaw phenomenon. Inspired by these approaches, we introduce the concept of multi-task learning into landmark detection and propose a novel single-landmark approach (SLA) to harness the *landmark individuality* and improve the accuracy of one-shot landmark detection.

We use the Cascade Comparing to Detect (CC2D) framework [24], a pioneering one-shot MLD method, as the foundation and propose the CC2D-SLA method, which eliminates the interdependencies between different landmarks by training a separate model for each landmark, thereby addressing the seesaw phenomenon at its core. Our experiments demonstrate that CC2D-SLA brings improved MLD accuracy over CC2D. To further enhance the detection accuracy of each landmark, we leverage the idea of augmented template data (ATD) [7], which leads to an improved SLA method called CC2D-SLA-ATD.

However, both CC2D-SLA and CC2D-SLA-ATD require training a separate model for each landmark, which leads to a substantial computational overhead. To address the issue of resource inefficiency and redundancy bought by multiple single-landmark models and further mitigate the seesaw phenomenon, we introduce an adapter to harness the *landmark similarity and individuality.* By combining shared weights with landmark-specific weights, we enable the model to learn the features of all landmarks through a single end-to-end model. Ultimately, this leads to the proposed CC2D-SLA-ATD-Adapter method, which

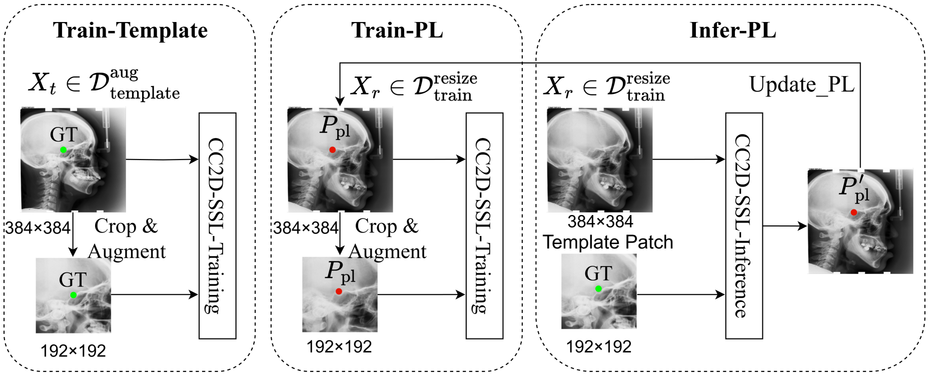

Fig. 1: Training framework of CC2D-SLA-ATD, which consists of three stages at each epoch: Train-Template, Train-PL, and Infer-PL. CC2D-SLA training, on the other hand, is composed of only the Train-PL and Infer-PL stages.

not only reduces computational and memory overhead but also maintains high precision in landmark detection performance.

We evaluate our models on the ISBI 2015 Challenge dataset [20]. Experimental results show that the single-landmark models significantly outperform traditional multi-point joint training models in landmark detection. Although our adapter-based fusion model (CC2D-SLA-ATD-Adapter) slightly underperforms compared to the best results from combining all single-landmark models, it still outperforms the current state-of-the-art (SOTA), demonstrating the potential of our method for medical imaging applications.

2 Method

Below, we first introduce CC2D-SLA and its improved variant, CC2D-SLA-ATD. We then describe how to integrate the adapter. Finally, we present the architecture of our CC2D-SLA-ATD-Adapter. Fig. 1 shows the training framework.

2.1 CC2D-SLA

The training of CC2D-SLA in each epoch consists of two stages: 1) Train-PL, in which the model is trained using pseudo-labels (PLs), and 2) Infer-PL, in which the pseudo-labels are inferred and updated.

Train-PL Stage. Let $\mathcal{D}_{\text{train}}$ be the training dataset. We first resize all images to 384×384, forming the resized set $\mathcal{D}^{\text{resize}}_{\text{train}}$. Initially, each image $X_r \in \mathcal{D}^{\text{resize}}_{\text{train}}$ is assigned a random pseudo-label $P_{pl} = (x_{pl}, y_{pl})$. As training proceeds, these pseudo-labels are updated at the end of each epoch. During the Train-PL stage, for each image X_r, we first crop a patch centered at P_{pl} and then apply data augmentation to obtain X_p. Both X_r and X_p are subsequently used as inputs to the CC2D-SSL framework in its training stage, referred to as CC2D-SSL-Training. Note that CC2D-SSL-Training encompasses the entire training pipeline of CC2D-SSL after receiving the image inputs, and is distinct from our CC2D-SLA approach. Here, we simply replace the typical input of CC2D-SSL-Training with the pair (X_r, X_p) to train on the PL-based patches.

Infer-PL Stage. After one pass through the training set, the model enters the Infer-PL stage to update the pseudo-labels. We denote CC2D-SSL-Inference as the complete inference pipeline of the CC2D-SSL framework. Specifically, for a given landmark ID k, a template patch X_{tp} is cropped from the template image around the ground-truth location of the k-th landmark. Next, every image $X_r \in \mathcal{D}_{\text{train}}^{\text{resize}}$ serves as a query image. We feed (X_{tp}, X_r) into the CC2D-SLA-Inference module to infer the new pseudo-label P'_{pl} for the k-th landmark on X_r. Thus, all pseudo-labels in $\mathcal{D}_{\text{train}}^{\text{resize}}$ are updated at the end of the epoch.

CC2D-SLA-ATD. In CC2D-SLA, training relies on pseudo-labels generated from a single template image. To better utilize the template data, we perform data augmentation on the template, following the procedure in FM-OSD [13]. Specifically, we apply random shifting, rotation, and scaling to produce 500 augmented versions of the template, forming the dataset $\mathcal{D}_{\text{template}}^{\text{aug}}$.

As illustrated in Fig. 1, for each augmented template image $X_t \in \mathcal{D}_{\text{template}}^{\text{aug}}$, we crop a patch centered at its ground-truth location and apply additional data augmentation to obtain X_{tp}. We then feed X_t and X_{tp} into the CC2D-SSL-Training pipeline. We refer to this procedure as the Train-Template stage, which is prepended to CC2D-SLA to get CC2D-SLA-ATD.

2.2 Adapter Integration for Multi-Landmark Training

To address the seesaw phenomenon and the resource inefficiency of multiple single-landmark models, we introduce adapter layers into our network. By incorporating adapters, all landmarks can be jointly trained using shared and landmark-specific weights, while ensuring only minor performance degradation.

Pre-Adapter Feature Extraction. As shown in the upper-left part of Fig. 2, without adapters, a feature map F_i of shape $H \times W \times C$ is transformed by a convolutional layer $Conv$, yielding $F_{i+1} \in \mathbb{R}^{H \times W \times C'}$: $F_{i+1} = Conv(F_i)$.

Adapter-Incorporated Layers. After integrating adapters, each landmark k has its own dedicated convolution $Conv^{A_k}$ alongside the shared convolution $Conv$. The primary difference lies in the output channel dimension: $Conv^{A_k}$ produces a feature map $F_{i+1}^{A_k} \in \mathbb{R}^{H \times W \times C_A}$, where C_A (e.g., 16) is typically

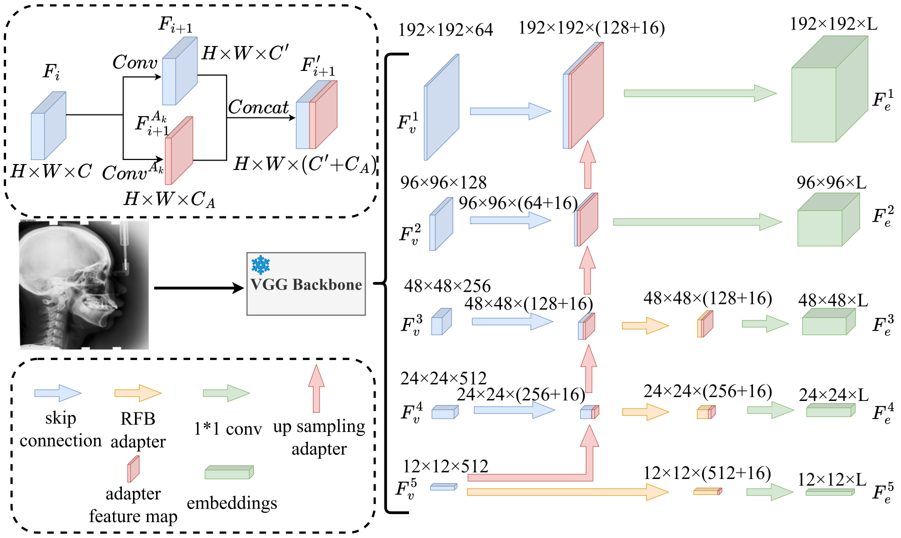

Fig. 2: Network architecture of CC2D-SLA-ATD-Adapter. The upper-left part of the figure shows an illustration of the adapter-based feature map transformation.

much smaller than C'. Formally, $F_{i+1}^{A_k} = Conv^{A_k}(F_i)$. We then concatenate F_{i+1} and $F_{i+1}^{A_k}$ along the channel dimension to obtain

$$F'_{i+1} = \text{Concat}\big(F_{i+1}, F_{i+1}^{A_k}\big) = \text{Concat}\Big(Conv(F_i),\, Conv^{A_k}(F_i)\Big), \tag{1}$$

where $F'_{i+1} \in \mathbb{R}^{H\times W\times(C'+C_A)}$. Equation (1) is simplified as $F'_{i+1} = C^{A_k}(F_i)$. Consequently, the subsequent layer's input channel size is adjusted to $C' + C_A$.

In this scheme, $Conv$ captures shared features across all landmarks, while $Conv^{A_k}$ learns landmark-specific features, activated only during training or inference of landmark k. This design alleviates the seesaw phenomenon, and enables landmark-specific learning without affecting others.

Adapters allow joint training of multiple landmarks, reducing the need for separate models for each, thus improving memory and computational efficiency. This shared-plus-specific design balances accuracy and resource usage, maintaining performance when scaling to multiple landmarks.

2.3 CC2D-SLA-ATD-Adapter

By enhancing CC2D-SLA-ATD with an adapter, we achieve CC2D-SLA-ATD-Adapter, enabling multi-landmark training while mitigating the seesaw phenomenon and improving overall performance.

As illustrated in Fig. 2, for an input image X of size 384×384, we use a pretrained VGG [18] network to extract five layers of feature maps: $F_v = \{F_v^1, F_v^2, \ldots, F_v^5\} = \text{VGG}(X)$. Here, F_v^5 is the deepest layer. We adopt the same

VGG19 model [18] pretrained on ImageNet [3] as in CC2D [24], and keep its weights frozen throughout training. Following [15], each upsampled feature map is concatenated with the corresponding lower-level feature map. Specifically,

$$F^i_{\text{concat}} = \text{Concat}\Big(\text{Up}\big(F_c^{i+1}\big),\, F_v^i\Big), \quad i = 1, 2, 3, 4. \tag{2}$$

Here, F_c is the feature map obtained by applying the adapter-based convolution:

$$F_c^i = C^{A_k}\big(F^i_{\text{concat}}\big), \quad i = 1, 2, 3, 4, \tag{3}$$

We set $F_c^5 = F_v^5$ to represent the deepest layer (where no upsampling occurs).

In CC2D [24], the third, fourth, and fifth layers employ Receptive Field Block (RFB) modules [10] to enlarge the receptive field. We also integrate adapters into the RFB modules by replacing the $Conv$ and $Conv^{A_k}$ modules with RFB and RFB^{A_k}, then concatenate their outputs. Formally,

$$F^i_{RFB} = R^{A_k}\big(F_c^i\big) = \text{Concat}\Big(\text{RFB}(F_c^i),\, \text{RFB}^{\text{A}_\text{k}}(F_c^i)\Big), \quad i = 3, 4, 5, \tag{4}$$

where F^i_{RFB} is the resulting feature map after applying the RFB with adapters.

Finally, each feature map is passed through a 1 × 1 convolution to yield a fixed-dimensional embedding F_e^i. In particular,

$$F_e^i = \begin{cases} \text{Conv}_{1\times1}\big(F_c^i\big), & \text{if } i = 1, 2, \\ \text{Conv}_{1\times1}\big(F^i_{RFB}\big), & \text{if } i = 3, 4, 5. \end{cases} \tag{5}$$

3 Experiments

3.1 Settings

Dataset. For this study, we use the widely recognized IEEE ISBI 2015 Challenge dataset [20], which contains 400 radiographs annotated with 19 landmarks by two expert clinicians. The average of their annotations serves as the ground truth. The images are 1935 × 2400 pixels with a 0.1mm pixel spacing. The dataset is split into 150 training and 250 testing images. One image is selected as the template, and the others are treated as unlabeled data for model training.

Evaluation Metrics. We evaluate the model performance using two common metrics: Mean Radial Error (MRE) and Successful Detection Rate (SDR). MRE calculates the average Euclidean distance between predicted landmarks and ground truth. SDR measures the proportion of landmarks detected within various thresholds (2mm, 2.5mm, 3mm, and 4mm) from the ground truth. These metrics are widely used in previous studies on landmark detection [24, 13, 29].

Table 1: Performance comparison of different methods on the Head [20] dataset.

Method	Model Count	MRE(↓) (mm)	SDR(↑)(%)			
			2mm	2.5mm	3mm	4mm
SAM [22]	1	2.56	54.11	63.66	70.25	80.84
UOD [29]	1	2.43	51.14	62.37	74.40	86.49
CC2D [24]	1	2.04	62.46	71.62	80.00	89.45
FM-OSD(coarse) [13]	1	1.93	63.60	75.43	83.03	91.94
FM-OSD(fine) [13]	2	1.82	67.35	77.92	84.59	91.92
CC2D-SLA(ours)	19	1.82	69.73	76.69	84.04	**92.17**
C-ATD(ours)	19	**1.79**	**72.02**	**78.02**	**84.72**	92.00
C-Adapter(ours)	1	1.96	67.83	75.33	81.89	90.82
C-F2(ours)	3	1.83	70.48	77.35	83.96	91.43

Implementation Details. All experiments are conducted using PyTorch on an NVIDIA RTX 3090 GPU with a learning rate of 0.0001, the Adam optimizer, a batch size of 8, and 300 epochs. In CC2D-SLA-ATD (or C-ATD in short), 19 models are trained, each dedicated to a single landmark. For CC2D-SLA-ATD-Adapter (or C-Adapter in short), we introduce 19 adapters, each with an output channel size of 16. A frozen VGG19 network serves as the feature extractor.

3.2 Performance Comparison

As shown in Table 1, we compare our proposed models with several SOTA methods, including SAM [22], UOD [29], CC2D [24], and the two stages of FM-OSD [13]. We re-trained SAM, UOD, and CC2D on the ISBI 2015 Challenge dataset under the one-shot setting and reported their best results following our experimental protocol. For FM-OSD [13], the fine-stage results are taken from the original paper, while the coarse-stage results are reproduced by us using the official implementation to ensure consistency with our setup. We also implement CC2D-SLA-ATD-Adapter-F2 (or C-F2 in short) , which uses our C-Adapter model as the coarse stage of FM-OSD, followed by the fine stage of FM-OSD to perform landmark detection on high-resolution medical images. Note that all models except C-F2 are evaluated on low-resolution images (384×384). We report the MRE and SDR results for these methods on the ISBI 2015 Challenge dataset and compare the number of models used.

It is evident that our C-ATD model achieves the best performance, with a 2mm SDR of 72.02% and an MRE of 1.79mm, significantly surpassing previous SOTA methods. This confirms that training individual landmark models effectively enhances detection accuracy, though it requires 19 separate models. To improve efficiency, we introduce adapters, allowing for the fusion of all landmarks into a single model. While this results in a slight performance drop compared to C-ATD, it still performs similarly to FM-OSD's coarse stage. Furthermore, applying FM-OSD's fine stage for high-resolution inference boosts the performance of C-Adapter, achieving a 2mm SDR of 70.48%. This not only exceeds the previous SOTA methods but also sets a new benchmark for MLD performance.

Table 2: Performance of different methods on the 19 landmarks. The landmark indices correspond to the positions described in [12]. The MRE is measured in millimeters, and the SDR refers to the 2mm SDR, with units in percentage.

Landmark	CC2D		FM-OSD		C-ATD		C-F2	
	MRE	SDR(2mm)	MRE	SDR(2mm)	MRE	SDR(2mm)	MRE	SDR(2mm)
1	1.35	85.2	1.55	84.0	**0.98**	**97.6**	1.26	92.0
2	1.60	74.0	**1.49**	73.2	1.51	**78.0**	1.54	73.6
3	1.58	72.4	1.66	68.4	**1.37**	**84.4**	1.46	78.4
4	1.93	66.4	2.28	55.6	**1.80**	**70.8**	2.00	67.2
5	1.86	62.8	1.72	65.6	**1.54**	**76.0**	1.61	75.2
6	2.50	**50.0**	**2.41**	48.4	2.47	49.2	2.43	48.4
7	1.42	81.6	1.05	88.0	0.96	**94.0**	**0.94**	93.6
8	1.46	78.4	**0.99**	91.6	1.10	**93.2**	1.93	89.6
9	1.08	88.8	**0.83**	94.8	0.84	**96.8**	0.84	95.2
10	4.19	20.4	3.23	33.6	3.59	24.8	**3.17**	**37.6**
11	2.58	42.8	**2.44**	**52.0**	2.85	48.8	2.66	48.8
12	2.60	48.8	1.91	68.0	**1.43**	**86.4**	1.50	80.8
13	1.63	70.0	1.60	68.4	1.56	78.4	**1.52**	**79.2**
14	1.69	71.2	**1.43**	80.0	1.54	79.6	1.47	**80.0**
15	1.73	65.6	**1.69**	**68.8**	2.50	50.0	1.89	59.6
16	3.54	26.4	2.61	47.2	**2.52**	**51.2**	2.56	48.8
17	1.73	67.6	1.55	74.8	**1.24**	**84.0**	1.51	77.6
18	1.77	65.6	**1.67**	**67.6**	2.00	63.2	1.82	66.4
19	2.44	48.8	2.49	49.6	**2.25**	**62.0**	2.75	47.2
Mean	2.04	62.5	1.82	67.4	**1.79**	**72.0**	1.83	70.5

Fig. 3 presents the predicted landmark detection results from different methods on the dataset. As shown in the figure, CC2D has the lowest prediction accuracy among the methods displayed, while FM-OSD achieves relatively better accuracy. The prediction accuracy of C-Adapter is comparable to that of FM-OSD. And the C-ATD model shows the best prediction accuracy overall.

As shown in Table 2, we provide the MRE and SDR results for each of the 19 landmarks across different methods. C-ATD achieves the best performance on most landmarks, with significant improvements for landmarks 1, 3, 5, 7, 12, 17, and 19, highlighting the effectiveness of the single-landmark approach. However, its performance is less optimal for landmarks 10, 15, and 18, suggesting that different landmarks have distinct characteristics, and these particular landmarks require absorbing more knowledge in order to achieve higher prediction accuracy. To address this, we introduce adapters, allowing landmarks to learn through shared weights as well as landmark-specific weights. After testing, this adaptation leads to a more balanced performance across all landmarks.

3.3 Ablation Study

As shown in the upper half of Table 3, we investigate the effect of varying the output channel size, denoted as C_A, for each adapter in the C-Adapter model. We set C_A to 0, 4, 8, 16, and 32. When no adapter is used ($C_A = 0$), the model's performance significantly degrades. However, with adapters, performance remains stable across different C_A values, suggesting that the adapters help the model recognize feature patterns, while shared weights handle main feature extraction.

Table 3: The performances of our methods with different channel sizes.

Para.	Value	MRE(↓)	SDR(↑)(%)			
		(mm)	2mm	2.5mm	3mm	4mm
C_A	0	2.30	63.92	72.15	79.68	89.09
	4	1.95	67.64	**75.92**	82.61	**91.07**
	8	1.95	67.07	74.53	81.18	90.48
	16	1.96	67.83	75.33	81.89	90.82
	32	**1.91**	**68.21**	75.64	**82.67**	90.99
CC2D's channels	+16	2.06	62.00	70.38	79.16	88.76
	+16×19	**2.02**	**63.68**	**72.32**	**80.38**	**89.68**

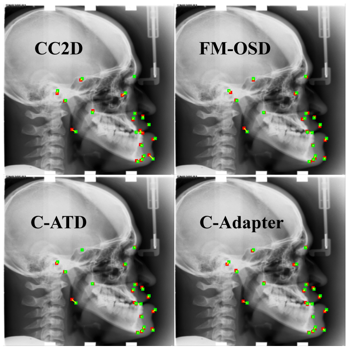

Fig. 3: Visualizations of the prediction results from different methods. The landmarks in red and green represent the predictions and ground truths.

In the lower half of Table 3, we explore the impact of increasing the number of channels in all convolution layers of the decoder in the original CC2D model, in order to examine how weight scaling affects performance. Specifically, we increased the number of channels by 16 and by 16×19. While increasing weights improves performance slightly, the gain is small, indicating that C-Adapter's performance enhancement is not primarily due to increased model weights.

4 Conclusion

In this paper, we present a progression from exploiting each landmark's individuality—through single-landmark training—to utilizing inter-landmark similarity by incorporating adapters into a unified model. This two-fold approach demonstrates a feasible and effective strategy for improving MLD accuracy. The proposed C-Adapter represents an initial endeavor toward jointly learning multiple landmarks via shared and landmark-specific weights. However, further investigation is needed into more advanced methods of inter-landmark collaboration to simultaneously enhance performance for all landmarks. We believe that continued exploration of this balance between individuality and similarity will yield more robust, efficient, and accurate solutions for one-shot MLD.

Acknowledgment. This work is supported by Natural Science Foundation of China under Grant 62271465 and Suzhou Basic Research Program under Grant SYG202338.

References

1. Bier, B., Unberath, M., Zaech, J.N., Fotouhi, J., Armand, M., Osgood, G., Navab, N., Maier, A.: X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 55–63. Springer (2018)
2. Caruana, R.: Multitask learning. Machine learning **28**, 41–75 (1997)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.: Bone age assessment of children using a digital hand atlas. Computerized medical imaging and graphics **31**(4-5), 322–331 (2007)
5. Han, D., Gao, Y., Wu, G., Yap, P.T., Shen, D.: Robust anatomical landmark detection for mr brain image registration. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17. pp. 186–193. Springer (2014)
6. Huang, Z., Li, H., Shao, S., Zhu, H., Hu, H., Cheng, Z., Wang, J., Kevin Zhou, S.: Pele scores: pelvic x-ray landmark detection with pelvis extraction and enhancement. International Journal of Computer Assisted Radiology and Surgery **19**(5), 939–950 (2024)
7. Huang, Z., Wang, S., Hu, H., Xu, Y.: Retigan: A hybrid image enhancement method for medical images. In: 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL). pp. 25–29. IEEE (2024)
8. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
10. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 385–400 (2018)
11. Liu, X., Wang, J., Liu, F., Zhou, S.K.: Universal undersampled mri reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 211–221. Springer (2021)
12. Lu, G., Zhang, Y., Kong, Y., Zhang, C., Coatrieux, J.L., Shu, H.: Landmark localization for cephalometric analysis using multiscale image patch-based graph convolutional networks. IEEE Journal of Biomedical and Health Informatics **26**(7), 3015–3024 (2022)
13. Miao, J., Chen, C., Zhang, K., Chuai, J., Li, Q., Heng, P.A.: FM-OSD: Foundation model-enabled one-shot detection of anatomical landmarks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 297–307. Springer (2024)
14. Oktay, O., Bai, W., Guerrero, R., Rajchl, M., De Marvao, A., O'Regan, D.P., Cook, S.A., Heinrich, M.P., Glocker, B., Rueckert, D.: Stratified decision forests for accurate anatomical landmark localization in cardiac images. IEEE transactions on medical imaging **36**(1), 332–342 (2016)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

16. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)
17. Shao, S., Yuan, X., Huang, Z., Qiu, Z., Wang, S., Zhou, K.: Diffuseexpand: Expanding dataset for 2d medical image segmentation using diffusion models. arXiv preprint arXiv:2304.13416 (2023)
18. Simonyan, K.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Tang, H., Liu, J., Zhao, M., Gong, X.: Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In: Proceedings of the 14th ACM conference on recommender systems. pp. 269–278 (2020)
20. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. Medical image analysis **31**, 63–76 (2016)
21. Wu, H., Wang, C., Mei, L., Yang, T., Zhu, M., Shen, D., Cui, Z.: Cephalometric landmark detection across ages with prototypical network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 155–165. Springer (2024)
22. Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L.: Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Transactions on Medical Imaging **41**(10), 2658–2669 (2022)
23. Yang, D., Zhang, S., Yan, Z., Tan, C., Li, K., Metaxas, D.: Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI). pp. 17–21. IEEE (2015)
24. Yao, Q., Quan, Q., Xiao, L., Kevin Zhou, S.: One-shot medical landmark detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 177–188. Springer (2021)
25. Yin, Z., Gong, P., Wang, C., Yu, Y., Wang, Y.: One-shot medical landmark localization by edge-guided transform and noisy landmark refinement. In: European Conference on Computer Vision. pp. 473–489. Springer (2022)
26. Zhang, J., Liu, M., An, L., Gao, Y., Shen, D.: Alzheimer's disease diagnosis using landmark-based features from longitudinal structural mr images. IEEE journal of biomedical and health informatics **21**(6), 1607–1616 (2017)
27. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE transactions on knowledge and data engineering **34**(12), 5586–5609 (2021)
28. Zhou, X., Huang, Z., Zhu, H., Yao, Q., Zhou, S.K.: Hybrid attention network: An efficient approach for anatomy-free landmark detection. arXiv preprint arXiv:2412.06499 (2024)
29. Zhu, H., Quan, Q., Yao, Q., Liu, Z., Zhou, S.K.: Uod: Universal one-shot detection of anatomical landmarks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2023)