Unsupervised Domain Adaptation for Anatomical Landmark Detection

Haibo Jin[1], Haoxuan Che[1], and Hao Chen[1,2(✉)]

[1] Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Kowloon, Hong Kong
{hjinag,hche,jhc}@cse.ust.hk
[2] Department of Chemical and Biological Engineering,
The Hong Kong University of Science and Technology, Kowloon, Hong Kong

Abstract. Recently, anatomical landmark detection has achieved great progresses on single-domain data, which usually assumes training and test sets are from the same domain. However, such an assumption is not always true in practice, which can cause significant performance drop due to domain shift. To tackle this problem, we propose a novel framework for anatomical landmark detection under the setting of unsupervised domain adaptation (UDA), which aims to transfer the knowledge from labeled source domain to unlabeled target domain. The framework leverages *self-training* and *domain adversarial learning* to address the domain gap during adaptation. Specifically, a self-training strategy is proposed to select reliable landmark-level pseudo-labels of target domain data with dynamic thresholds, which makes the adaptation more effective. Furthermore, a domain adversarial learning module is designed to handle the unaligned data distributions of two domains by learning domain-invariant features via adversarial training. Our experiments on cephalometric and lung landmark detection show the effectiveness of the method, which reduces the domain gap by a large margin and outperforms other UDA methods consistently.

1 Introduction

Anatomical landmark detection is a fundamental step in many clinical applications such as orthodontic diagnosis [11] and orthognathic treatment planning [6]. However, manually locating the landmarks can be tedious and time-consuming. And the results from manual labeling can cause errors due to the inconsistency in landmark identification [5]. Therefore, it is of great need to automate the task of landmark detection for efficiency and consistency.

In recent years, deep learning based methods have achieved great progresses in anatomical landmark detection. For supervised learning, earlier works [6,20,27] adopted heatmap regression with extra shape constraints. Later,

(a) Histogram   (b) Domain A   (c) Domain B
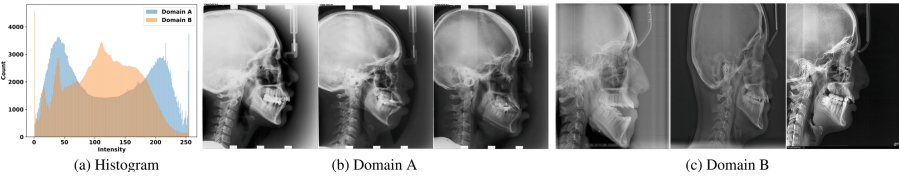
Fig. 1. Domain A vs. Domain B. (a) Image histogram. (b)–(c) Visual samples.

graph network [16] and self-attention [11] were introduced to model landmark dependencies in an end-to-end manner for better performance.

Despite the success of recent methods, they mostly focus on single-domain data, which assume the training and test sets follow the same distribution. However, such an assumption is not always true in practice, due to the differences in patient populations and imaging devices. Figure 1 shows that cephalogram images from two domains can be very different in both histogram and visual appearance. Therefore, a well trained model may encounter severe performance degradation in practice due to the domain shift of test data. A straightforward solution to this issue is to largely increase the size and diversity of training set, but the labeling is prohibitively expensive, especially for medical images. On the other hand, unsupervised domain adaptation (UDA) [10] aims to transfer the knowledge learned from the labeled source domain to the unlabeled target domain, which is a potential solution to the domain shift problem as unlabeled data is much easier to collect. The effectiveness of UDA has been proven in many vision tasks, such as image classification [10], object detection [7], and pose estimation [3,19,24]. However, its feasibility in anatomical landmark detection still remains unknown.

In this paper, we aim to investigate anatomical landmark detection under the setting of UDA. Our preliminary experiments show that a well-performed model will yield significant performance drop on cross-domain data, where the mean radial error (MRE) increases from 1.22 mm to 3.32 mm and the success detection rate (SDR) within 2 mm drops from 83.76% to 50.05%. To address the domain gap, we propose a unified framework, which contains a base landmark detection model, a self-training strategy, and a domain adversarial learning module. Specifically, self-training is adopted to effectively leverage the unlabeled data from the target domain via pseudo-labels. To handle confirmation bias [2], we propose landmark-aware self-training (LAST) to select pseudo-labels at the landmark-level with dynamic thresholds. Furthermore, to address the covariate shift [26] issue (i.e., unaligned data distribution) that may degrade the performance of self-training, a domain adversarial learning (DAL) module is designed to learn domain-invariant features via adversarial training. Our experiments on two anatomical datasets show the effectiveness of the proposed framework. For example, on cephalometric landmark detection, it reduces the domain gap in MRE by 47% (3.32 mm → 1.75 mm) and improves the SDR (2 mm) from 50.05% to 69.15%. We summarize our contributions as follows.
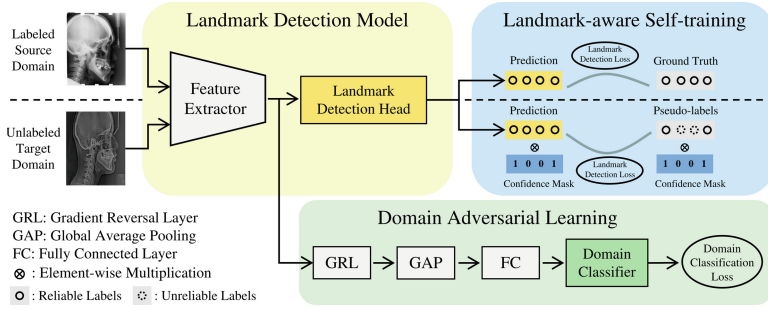
Fig. 2. The overall framework. Based on 1) the landmark detection model, it 2) utilizes LAST to leverage the unlabeled target domain data via pseudo-labels, and 3) simultaneously conducts DAL for domain-invariant features.

1. We investigated anatomical landmark detection under the UDA setting for the *first time*, and showed that domain shift indeed causes severe performance drop of a well-performed landmark detection model.
2. We proposed a novel framework for the UDA of anatomical landmark detection, which significantly improves the cross-domain performance and consistently outperforms other state-of-the-art UDA methods.

2   Method

Figure 2 shows the overall framework, which aims to yield satisfactory performance in target domain under the UDA setting. During training, it leverages both labeled source domain data $\mathcal{S} = \{x_i^{\mathcal{S}}, y_i^{\mathcal{S}}\}_{i=1}^N$ and unlabeled target domain data $\mathcal{T} = \{x_j^{\mathcal{T}}\}_{j=1}^M$. For evaluation, it will be tested on a hold-out test set from target domain. The landmark detection model is able to predict landmarks with confidence, which is detailed in Sect. 2.1. To reduce domain gap, we further propose LAST and DAL, which are introduced in Sects. 2.2 and 2.3, respectively.

2.1   Landmark Detection Model

Recently, coordinate regression [11,16] has obtained better performance than heatmap regression [6,27]. However, coordinate based methods do not output confidence scores, which are necessary for pseudo-label selection in self-training [4,14]. To address this issue, we designed a model that is able to predict accurate landmarks while providing confidence scores. As shown in Fig. 3 (a), the model utilizes both coordinate and heatmap regression, where the former provides coarse but robust predictions via global localization, then projected to the local maps of the latter for prediction refinement and confidence measurement.

Global Localization. We adopt Transformer decoder [12,15] for coarse localization due to its superiority in global attentions. A convolutional neural network
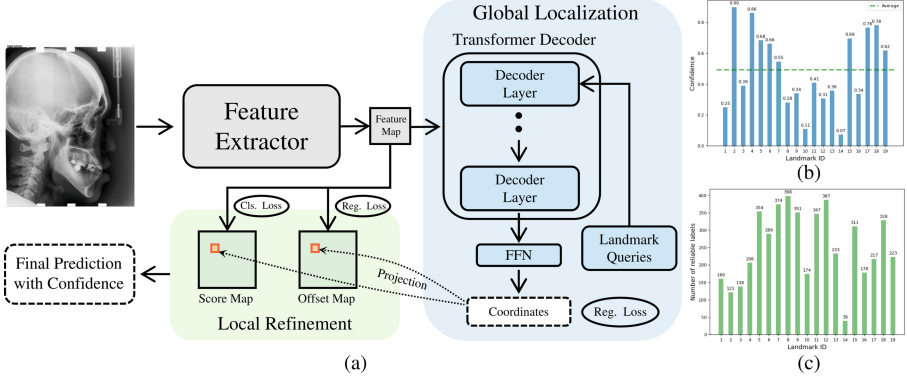
Fig. 3. (a) Our landmark detection model. (b) Confidence scores of different landmarks for a random target domain image. (c) Statistics of reliable landmark-level pseudo-labels with a fixed threshold $\tau = 0.4$ over 500 images.

(CNN) is used to extract feature $f \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represents number of channels, map height and width, respectively. By using f as memory, the decoder takes landmark queries $q \in \mathbb{R}^{L \times C}$ as input, then iteratively updates them through multiple decoder layers, where L is the number of landmarks. Finally, a feed-forward network (FFN) converts the updated landmark queries to coordinates $\hat{y}_c \in \mathbb{R}^{L \times 2}$. The loss function L_{coord} is defined to be the L1 loss between the predicted coordinate $\hat{y}_c$ and the label y_c.

Local Refinement. This module outputs a score map $\hat{y}_s \in \mathbb{R}^{L \times H \times W}$ and an offset map $\hat{y}_o \in \mathbb{R}^{2L \times H \times W}$ via 1×1 convolutional layers by taking f as input. The score map indicates the likelihood of each grid to be the target landmark, while the offset map represents the relative offsets of the neighbouring grids to the target. The ground-truth (GT) landmark of the score map is smoothed by a Gaussian kernel [23], and L2 loss is used for loss function L_{score}. Since the offset is a regression problem, L1 is used for loss L_{offset}, and only applied to the area where its GT score is larger than zero. During inference, different from [6,18,23], the optimal local grid is not selected by the maximum score of $\hat{y}_s$, but instead the projection of the coordinates from global localization. Then the corresponding offset value is added to the optimal grid for refinement. Also, the confidence of each prediction can be easily obtained from the score map via projection.

The loss function of the landmark detection model can be summarized as

$$L_{\text{base}} = \sum_{\mathcal{S}} \lambda_s L_{\text{score}} + \lambda_o L_{\text{offset}} + L_{\text{coord}}, \tag{1}$$

where $\mathcal{S}$ is source domain data, λ_s and λ_o are balancing coefficients. Empirically, we set $\lambda_s = 100$ and $\lambda_o = 0.02$ in this paper.

2.2 Landmark-Aware Self-training

Self-training [14] is an effective semi-supervised learning (SSL) method, which iteratively estimates and selects reliable pseudo-labeled samples to expand the labeled set. Its effectiveness has also been verified on several vision tasks under the UDA setting, such as object detection [7]. However, very few works explored self-training for the UDA of landmark detection, but mostly restricted to the paradigm of SSL [9,21].

Since UDA is more challenging than SSL due to domain shift, reliable pseudo-labels should be carefully selected to avoid confirmation bias [2]. Existing works [9,19,21] follow the pipeline of image classification by evaluating reliability at the image-level, which we believe is not representative because the landmarks within an image may have different reliabilities (see Fig. 3 (b)). To avoid potential noisy labels caused by the image-level selection, we propose LAST, which selects reliable pseudo-labels at the landmark-level. To achieve this, we use a binary mask $m \in \{0,1\}^L$ to indicate the reliability of each landmark for a given image, where value 1 indicates the label is reliable and 0 the opposite. To decide the reliability of each landmark, a common practice is to use a threshold τ, where the l-th landmark is reliable if its confidence score $s^l > \tau$. During loss calculation, each loss term is multiplied by m to mask out the unreliable landmark-level labels. Thus, the loss for LAST is

$$L_{\text{LAST}} = \sum_{\mathcal{S} \cup \mathcal{T}'} M(L_{\text{base}}),$$ (2)

where M represents the mask operation, $\mathcal{T}' = \{x_j^{\mathcal{T}}, y_j'^{\mathcal{T}}\}_{j=1}^M$, and $y'^{\mathcal{T}}$ is the estimated pseudo-labels from the last self-training round. Note that the masks of the source domain data $\mathcal{S}$ always equal to one as they are ground truths.

However, the landmark-level selection leads to unbalanced pseudo-labels between landmarks, as shown in Fig. 3 (c). This is caused by the fixed threshold τ in self-training, which cannot handle different landmarks adaptively. To address this issue, we introduce percentile scores [4] to yield dynamic thresholds (DT) for different landmarks. Specifically, for the l-th landmark, when the pseudo-labels are sorted based on confidence (high to low), τ_r^l is used as the threshold, which is the confidence score of r-th percentile. In this way, the selection ratio of pseudo-labels can be controlled by r, and the unbalanced issue can be addressed by using the same r for all the landmarks. We set the curriculum to be $r = \Delta \cdot t$, where t is the t-th self-training round and Δ is a hyperparameter that controls the pace. We use $\Delta = 20\%$, which yields five training rounds in total.

2.3 Domain Adversarial Learning

Although self-training has been shown effective, it inevitably contains bias towards source domain because its initial model is trained with source domain data only. In other words, the data distribution of target domain is different from the source domain, which is known as covariate shift [26]. To mitigate it,

we introduce DAL to align the distribution of the two by conducting an adversarial training between a domain classifier and the feature extractor. Specifically, the feature f further goes through a global average pooling (GAP) and a fully connected (FC) layer, then connects to a domain classifier D to discriminate the source of input x. The classifier can be trained with binary cross-entropy loss:

$$L_{\mathrm{D}} = -d \log D(F(x)) - (1 - d) \log(1 - D(F(x))), \qquad (3)$$

where d is domain label, with $d = 0$ and $d = 1$ indicating the images are from source and target domain, respectively. The domain classifier is trained to minimize L_{D}, while the feature extractor F is encouraged to maximize it such that the learned feature is indistinguishable to the domain classifier. Thus, the adversarial objective function can be written as $L_{\mathrm{DAL}} = \max_{F} \min_{D} L_{\mathrm{D}}$. To simplify the optimization, we adopt gradient reversal layer (GRL) [10] to mimic the adversarial training, which is placed right after the feature extractor. During backpropagation, GRL negates the gradients that pass back to the feature extractor F so that F is actually maximized. In this way, the adversarial training can be done via the minimization of L_{D}, i.e., $L_{\mathrm{DAL}} = L_{\mathrm{D}}$.

Finally, we have the overall loss function as follows:

$$L = \sum_{\mathcal{S} \cup \mathcal{T}'} L_{\mathrm{LAST}} + \lambda_D L_{\mathrm{DAL}}, \qquad (4)$$

where λ_D is a balancing coefficient.

3   Experiments

3.1   Experimental Settings

In this section, we present experiments on cephalometric landmark detection. See lung landmark detection in Appendix A.

Source Domain. The ISBI 2015 Challenge provides a public dataset [22], which is widely used as a benchmark of cephalometric landmark detection. It contains 400 images in total, where 150 images are for training, 150 images are Test 1 data, and the remaining are Test 2. Each image is annotated with 19 landmarks by two experienced doctors, and the mean values of the two are used as GT. In this paper, we only use the training set as the labeled source domain data.

Target Domain. The ISBI 2023 Challenge provides a new dataset [13], which was collected from 7 different imaging devices. By now, only the training set is released, which contains 700 images. For UDA setting, we randomly selected 500 images as unlabeled target domain data, and the remaining 200 images are for evaluation. The dataset provides 29 landmarks, but we only use 19 of them, i.e., the same landmarks as the source domain [22]. Following previous works [11,16], all the images are resized to 640×800. For evaluation metric, we adopt MRE and SDR within four radius (2 mm, 2.5 mm, 3 mm, and 4 mm).

Table 1. Results on the target domain test set, under UDA setting.

Method	MRE↓	2 mm	2.5 mm	3 mm	4 mm
Base, Labeled Source	3.32	50.05	56.87	62.63	70.87
FDA [25]	2.16	61.28	69.73	76.34	84.57
UMT [8]	1.98	63.94	72.52	78.89	87.05
SAC [1]	1.94	65.68	73.76	79.63	87.81
AT [17]	1.87	66.82	74.81	80.73	88.47
Ours	**1.75**	**69.15**	**76.94**	**82.92**	**90.05**
Base, Labeled Target	1.22	83.76	89.71	92.79	96.08

Implementation Details. We use ImageNet pretrained ResNet-50 as the backbone, followed by three deconvolutional layers for upsampling to stride 4 [23]. For Transformer decoder, three decoder layers are used, and the embedding length $C = 256$. Our model has 41M parameters and 139 GFLOPs when input size is 640×800. The source domain images are oversampled to the same number of target domain so that the domain classifier is unbiased. Adam is used as the optimizer, and the model is trained for 720 epochs in each self-training round. The initial learning rate is 2×10^{-4}, and decayed by 10 at the 480th and 640th epoch. The batch size is set to 10 and λ_D is set to 0.01. For data augmentation, we use random scaling, translation, rotation, occlusion, and blurring. The code was implemented with PyTorch 1.13 and trained with one RTX 3090 GPU. The training took about 54 h.

3.2 Results

For the comparison under UDA setting, several state-of-the-art UDA methods were implemented, including FDA [25], UMT [8], SAC [1], and AT [17]. Additionally, the base model trained with source domain data only is included as the lower bound, and the model trained with equal amount of labeled target domain data is used as the upper bound. Table 1 shows the results. Firstly, we can see that the model trained on the target domain obtains much better performance than the one on source domain in both MRE (1.22 mm vs. 3.32 mm) and SDR (83.76% vs. 50.05%, within 2 mm), which indicates that the domain shift can cause severe performance degradation. By leveraging both labeled source domain and unlabeled target domain data, our model achieves 1.75 mm in MRE and 69.15% in SDR within 2 mm. It not only reduces the domain gap by a large margin (3.32 mm → 1.75 mm in MRE and 50.05% → 69.15% in 2 mm SDR), but also outperforms the other UDA methods consistently. However, there is still a gap between the UDA methods and the supervised model in target domain.

3.3 Model Analysis

We first do ablation study to show the effectiveness of each module, which can be seen in Table 2. The baseline model simply uses vanilla self-training [14] for
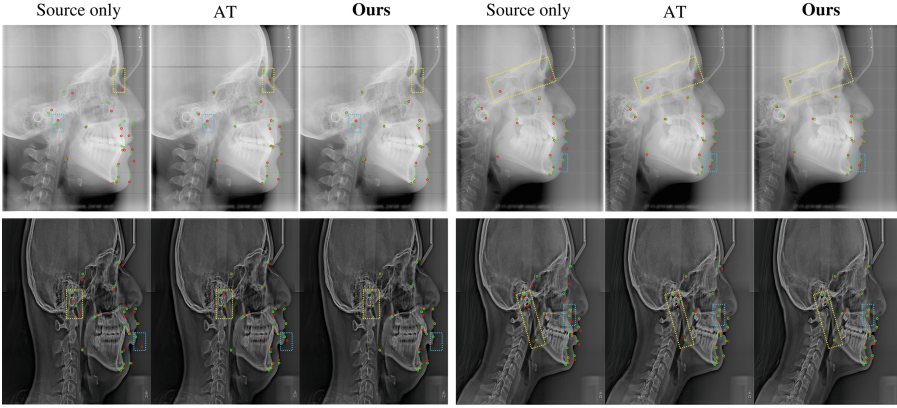
Fig. 4. Qualitative results of three models on target domain test data. Green dots are GTs, and red dots are predictions. Yellow rectangles indicate that our model performs better than the other two, while cyan rectangles indicate that all the three fail. (Color figure online)

Table 2. Ablation study of different modules.

Method	MRE↓	2 mm	2.5 mm	3 mm	4 mm
Self-training [14]	2.18	62.18	69.44	75.47	84.36
LAST w/o DT	1.98	65.34	72.53	78.03	86.11
LAST	1.91	66.21	74.39	80.23	88.42
DAL	1.96	65.92	74.18	79.73	87.60
LAST+DAL	**1.75**	**69.15**	**76.94**	**82.92**	**90.05**
LAST+DAL w/ HM	1.84	66.45	75.09	81.82	89.55

domain adaptation, which achieves 2.18 mm in MRE. By adding LAST but without dynamic thresholds (DT), the MRE improves to 1.98 mm. When the proposed LAST and DAL are applied separately, the MREs are 1.91 mm and 1.96 mm, respectively, which verifies the effectiveness of the two modules. By combining the two, the model obtains the best results in both MRE and SDR. To show the superiority of our base model, we replace it by standard heatmap regression [23] (HM), which obtains degraded results in both MRE and SDR. Furthermore, we conduct analysis on subdomain discrepancy, which shows the effectiveness of our method on each subdomain (see Appendix B).

3.4   Qualitative Results

Figure 4 shows the qualitative results of the source-only base model, AT [17], and our method on target domain test data. The green dots are GTs, and red dots are predictions. It can be seen from the figure that our model makes better predictions than the other two (see yellow rectangles). We also notice that

some landmarks are quite challenging, where all the three fail to give accurate predictions (see cyan rectangles).

4 Conclusion

In this paper, we investigated anatomical landmark detection under the UDA setting. To mitigate the performance drop caused by domain shift, we proposed a unified UDA framework, which consists of a landmark detection model, a self-training strategy, and a DAL module. Based on the predictions and confidence scores from the landmark model, a self-training strategy is proposed for domain adaptation via landmark-level pseudo-labels with dynamic thresholds. Meanwhile, the model is encouraged to learn domain-invariant features via adversarial training so that the unaligned data distribution can be addressed. We constructed a UDA setting based on two anatomical datasets, where the experiments showed that our method not only reduces the domain gap by a large margin, but also outperforms other UDA methods consistently. However, a performance gap still exists between the current UDA methods and the supervised model in target domain, indicating more effective UDA methods are needed to close the gap.

References

1. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15384–15394 (2021)
2. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
3. Bigalke, A., Hansen, L., Diesel, J., Heinrich, M.P.: Domain adaptation through anatomical constraints for 3D human pose estimation under the cover. In: International Conference on Medical Imaging with Deep Learning, pp. 173–187. PMLR (2022)
4. Cascante-Bonilla, P., Tan, F., Qi, Y., Ordonez, V.: Curriculum labeling: revisiting pseudo-labeling for semi-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 6912–6920 (2021)
5. Chen, M.H., et al.: Intraobserver reliability of landmark identification in cone-beam computed tomography-synthesized two-dimensional cephalograms versus conventional cephalometric radiography: a preliminary study. J. Dental Sci. **9**(1), 56–62 (2014)
6. Chen, R., Ma, Y., Chen, N., Lee, D., Wang, W.: Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 873–881. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_97

7. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3339–3348 (2018)

8. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4091–4101 (2021)

9. Dong, X., Yang, Y.: Teacher supervises students how to learn from partially labeled images for facial landmark detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 783–792 (2019)

10. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning, pp. 1180–1189. PMLR (2015)

11. Jiang, Y., Li, Y., Wang, X., Tao, Y., Lin, J., Lin, H.: CephalFormer: incorporating global structure constraint into visual features for general cephalometric landmark detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 227–237. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_22

12. Jin, H., Li, J., Liao, S., Shao, L.: When liebig's barrel meets facial landmark detection: a practical model. arXiv preprint arXiv:2105.13150 (2021)

13. Khalid, M.A., et al.: Aariz: a benchmark dataset for automatic cephalometric landmark detection and CVM stage classification. arXiv:2302.07797 (2023)

14. Lee, D.H.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning (2013)

15. Li, J., Jin, H., Liao, S., Shao, L., Heng, P.A.: RepFormer: refinement pyramid transformer for robust facial landmark detection. In: IJCAI (2022)

16. Li, W., et al.: Structured landmark detection via topology-adapting deep graph learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 266–283. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_16

17. Li, Y.J., et al.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7581–7590 (2022)

18. Liu, W., Wang, Yu., Jiang, T., Chi, Y., Zhang, L., Hua, X.-S.: Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 670–679. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_65

19. Mu, J., Qiu, W., Hager, G.D., Yuille, A.L.: Learning from synthetic animals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12386–12395 (2020)

20. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal. (2019)

21. Wang, C., et al.: Pseudo-labeled auto-curriculum learning for semi-supervised keypoint localization. In: ICLR (2022)

22. Wang, C.W., et al.: A benchmark for comparison of dental radiography analysis algorithms. Med. Image Anal. (2016)

23. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 472–487. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_29

24. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3D human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5255–5264 (2018)
25. Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095 (2020)
26. Zhao, S., et al.: A review of single-source deep unsupervised visual domain adaptation. IEEE Tran. Neural Netw. Learn. Syst. **33**(2), 473–493 (2020)
27. Zhong, Z., Li, J., Zhang, Z., Jiao, Z., Gao, X.: An attention-guided deep regression model for landmark detection in cephalograms. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 540–548. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_60