UOD: Universal One-Shot Detection of Anatomical Landmarks

Heqin Zhu[1,2,3], Quan Quan[3], Qingsong Yao[3], Zaiyi Liu[4,5], 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 230026, Anhui, People's Republic of China
skevinzhou@ustc.edu.cn
[2] Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, Jiangsu, People's Republic of China
[3] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[4] Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China
[5] Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China

Abstract. One-shot medical landmark detection gains much attention and achieves great success for its label-efficient training process. However, existing one-shot learning methods are highly specialized in a single domain and suffer domain preference heavily in the situation of multi-domain unlabeled data. Moreover, one-shot learning is not robust that it faces performance drop when annotating a sub-optimal image. To tackle these issues, we resort to developing a domain-adaptive one-shot landmark detection framework for handling multi-domain medical images, named **Universal One-shot Detection (UOD)**. UOD consists of two stages and two corresponding universal models which are designed as combinations of domain-specific modules and domain-shared modules. In the first stage, a domain-adaptive convolution model is self-supervised learned to generate pseudo landmark labels. In the second stage, we design a domain-adaptive transformer to eliminate domain preference and build the global context for multi-domain data. Even though only one annotated sample from each domain is available for training, the domain-shared modules help UOD aggregate all one-shot samples to detect more robust and accurate landmarks. We investigated both qualitatively and quantitatively the proposed UOD on three widely-used public X-ray datasets in different anatomical domains (i.e., head, hand, chest) and obtained state-of-the-art performances in each domain. The code is at https://github.com/heqin-zhu/UOD_universal_oneshot_detection.

Keywords: One-shot learning · Domain-adaptive model · Anatomical landmark detection · Transformer network

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 24–34, 2023.
https://doi.org/10.1007/978-3-031-43907-0_3

1 Introduction

Robust and accurate detecting of anatomical landmarks is an essential task in medical image applications [24,25], which plays vital parts in varieties of clinical treatments, for instance, vertebrae localization [20], orthognathic and orthodontic surgeries [9], and craniofacial anomalies assessment [4]. Moreover, anatomical landmarks exert their effectiveness in other medical image tasks such as segmentation [3], registration [5], and biometry estimation [1].

In the past years, lots of fully supervised methods [4,8,11,11,20,21,26,27] have been proposed to detect landmarks accurately and automatically. To relieve the burden of experts and reduce the amount of annotated labels, various one-shot and few-shot methods have been come up with. Zhao et al. [23] demonstrate a model which learns transformations from the images and uses the labeled example to synthesize additional labeled examples, where each transformation is composed of a spatial deformation field and an intensity change. Yao et al. [22] develop a cascaded self-supervised learning framework for one-shot medical landmark detection. They first train a matching network to calculate the cosine similarity between features from an image and a template patch, then fine-tune the pseudo landmark labels from coarse to fine. Browatzki et al. [2] propose a semi-supervised method that consists of two stages. They first employ an adversarial auto-encoder to learn implicit face knowledge from unlabeled images and then fine-tune the decoder to detect landmarks with few-shot labels.

However, one-shot methods are not robust enough because they are dependent on the choice of labeled template and the accuracy of detected landmarks may decrease a lot when choosing a sub-optimal image to annotate. To address this issue, Quan et al. [12] propose a novel Sample Choosing Policy (SCP) to select the most worthy image to annotate. Despite the improved performance, SCP brings an extra computation burden. Another challenge is the scalability of model building when facing multiple domains (such as different anatomical regions). While conventional wisdom is to independently train a model for each domain, Zhu et al. [26] propose a universal model YOLO for detecting landmarks across different anatomies and achieving better performances than a collection of single models. YOLO is regularly supervised using the CNN as backbone and it is unknown if the YOLO model works for a one-shot scenario and with a modern transformer architecture.

Motivated by above challenges, to detect robust multi-domain label-efficient landmarks, we design domain-adaptive models and propose a universal one-shot landmark detection framework called **Universal One-shot Detection (UOD)**, illustrated in Fig. 1. A universal model is comprised of domain-specific modules and domain-shared modules, learning the specified features of each domain and common features of all domains to eliminate domain preference and extract representative features for multi-domain data. Moreover, one-shot learning is not robust enough because of the sample selection while multi-domain one-shot learning reaps benefit from different one-shot samples from various domains, in which cross-domain features are excavated by domain-shared modules. Our proposed UOD framework consists of two stages: 1) Contrastive learning for

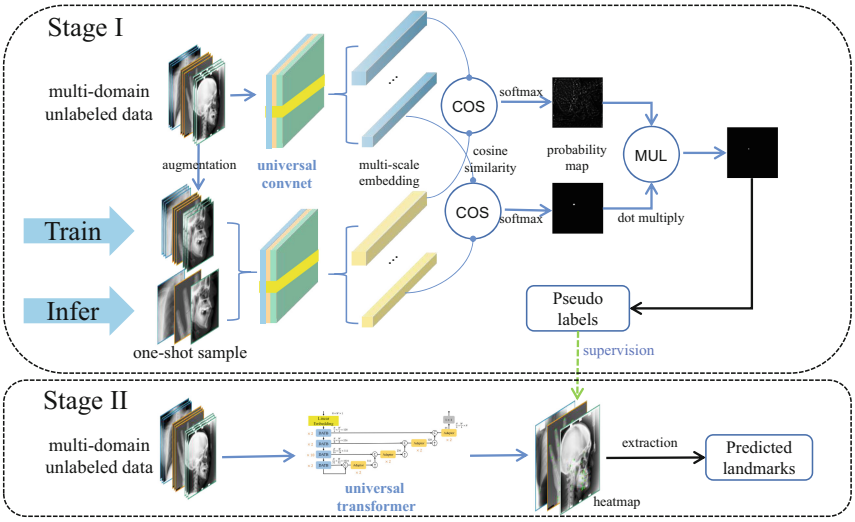


Fig. 1. Overview of UOD framework. In stage I, two universal models are learned via contrastive learning for matching similar patches from original image and augmented one-shot sample image and generating pseudo labels. In stage II, DATR is designed to better capture global context information among all domains for detecting more accurate landmarks.

training a universal model with multi-domain data to generate pseudo landmark labels. 2) Supervised learning for training domain-adaptive transformer (DATR) to avoid domain preference and detect robust and accurate landmarks.

In summary, our contributions can be categorized into three parts: **1)** We design the first universal framework for multi-domain one-shot landmark detection, which improves detecting accuracy and relieves domain preference on multi-domain data from various anatomical regions. **2)** We design a domain-adaptive transformer block (DATB), which is effective for multi-domain learning and can be used in any other transformer network. **3)** We carry out comprehensive experiments to demonstrate the effectiveness of UOD for obtaining SOTA performance on three publicly used X-ray datasets of head, hand, and chest.

2 Method

As Fig. 1 shows, UOD consists of two stages: 1) Contrastive learning and 2) Supervised learning. In stage I, to learn the local appearance of each domain, a universal model is trained via self-supervised learning, which contains domain-specific VGG [15] and UNet [13] decoder with each standard convolution replaced by a domain adaptor [7]. In stage II, to grasp the global constraint and eliminate domain preference, we designed a domain-adaptive transformer (DATR).

2.1 Stage I: Contrastive Learning

As Fig. 1 shows, following Yao et al. [22], we employ contrastive learning to train siamese network for matching similar patches of original image and augmented image. Given a multi-domain input image $X^d \in R^{H^d \times W^d \times C^d}$ belongs to domain d from multi-domain data, we randomly select a target point P and crop a half-size patch X_p^d which contains P. After applying data augmentation on X_p^d, the target point is mapped to P_p. Then we feed X^d and X_p^d into the siamese network respectively and obtain the multi-scale feature embeddings. We compute cosine similarity of two feature embeddings from each scale and apply softmax to the cosine similarity map to generate a probability matrix. Finally, we calculate the cross entropy loss of the probability matrix and ground truth map which is produced with the one-hot encoding of P_p^d to optimize the siamese network for learning the latent similarities of patches. At inferring stage, we replace augmented patch X_p^d with the augmented one-shot sample patch X_s^d. We use the annotated one-shot landmarks as target points to formulate the ground truth maps. After obtaining probability matrices, we apply arg max to extract the strongest response points as the pseudo landmarks, which will be used in UOD Stage II.

2.2 Stage II: Supervised Learning

In stage II, we design a universal transformer to capture global relationship of multi-domain data and train it with the pseudo landmarks generated in stage I. The universal transformer has a domain-adaptive transformer encoder and domain-adaptive convolution decoder. The decoder is based on a U-Net [13] decoder with each standard convolution replaced by a domain adaptor [7]. The encoder is based on Swin Transformer [10] with shifted window and limited self-attention within non-overlapping local windows for computation efficiency. Different from Swin Transformer [10], we design a domain-adaptive transformer block (DATB) and use it to replace the original transformer block.

Domain-Adaptive Transformer Encoder. As Fig. 2(a) shows, the transformer encoder is built up with DATB, making full use of the capability of transformer for modeling global relationship and extracting multi-domain representative features. As in Fig. 2(b), a basic transformer block [17] consists of a multi-head self-attention module (MSA), followed by a two-layer MLP with GELU activation. Furthermore, layer normalization (LN) is adopted before each MSA and MLP and a residual connection is adopted after each MSA and MLP. Given a feature map $x^d \in R^{h \times w \times c}$ from domain d with height h, width w, and c channels, the output feature maps of MSA and MLP, denoted by $\hat{y}^d$ and y^d, respectively, are formulated as:

$$\begin{aligned} \hat{y}^d &= \mathrm{MSA}(\mathrm{LN}(x^d)) + x^d \\ y^d &= \mathrm{MLP}(\mathrm{LN}(\hat{y}^d)) + \hat{y}^d \end{aligned} \tag{1}$$

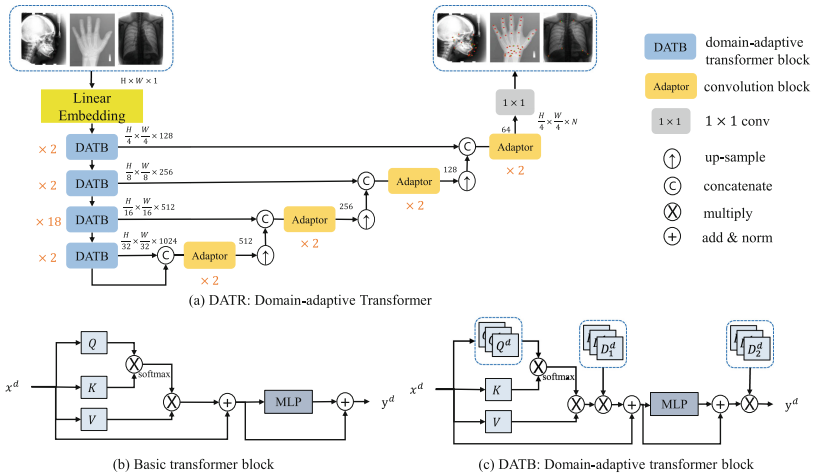


Fig. 2. (a) The architecture of DATR in stage II, which is composed of domain-adaptive transformer encoder and convolution adaptors [7]. (b) Basic transformer block. (c) Domain-adaptive transformer block. Each domain-adaptive transformer is a basic transformer block with query matrix duplicated and domain-adaptive diagonal for each domain. The batch-normalization, activation, and patch merging are omitted.

where $\text{MSA} = \text{softmax}(QK^T)V$.

As illustrated in Fig. 2(b)(c), DATB is based on Eq. (1). Similar to U2Net [7] and GU2Net [26], we adopt domain-specific and domain-shared parameters in DATB. Since the attention probability is dependent on query and key matrix which are symmetrical, we duplicate the query matrix for each domain to learn domain-specific query features and keep key and value matrix domain-shared to learn common knowledge and reduce parameters. Inspired by LayerScale [16], we further adopt learnable diagonal matrix [16] after each MSA and MLP module to facilitate the learning of domain-specific features, which costs few parameters ($O(N)$ for $N \times N$ diagonal). Different from LayerScale [16], proposed domain-adaptive diagonal D_1^d and D_2^d are applied for each domain with D_2^d applied after residual connection for generating more representative and direct domain-specific features. The above process can be formulated as:

$$\begin{aligned} \hat{y}^d &= D_1^d \times \text{MSA}_{Q^d}(\text{LN}(x^d)) + x^d \\ y^d &= D_2^d \times (\text{MLP}(\text{LN}(\hat{y}^d)) + \hat{y}^d) \end{aligned} \tag{2}$$

where $\text{MSA}_{Q^d} = \text{softmax}(Q^d K^T)V$.

Overall Pipeline. Given that a random input $X^d \in R^{H^d \times W^d \times C^d}$ belongs to domain d from mixed datasets on various anatomical regions, which contains N^d landmarks with corresponding coordinates being $\{(i_1^d, j_1^d), (i_2^d, j_2^d), \ldots, (i_{N_d}^d, j_{N_d}^d)\}$, we set the n-th $\in \{1, 2, \ldots, N^d\}$ initial heatmap

$\tilde{Y}_n^d \in R^{H^d \times W^d \times C^d}$ with Gaussian function to be $\tilde{Y}_n^d = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(i-i_n^d)^2+(j-j_n^d)^2}{2\sigma^2}}$ if $\sqrt{(i-i_n^d)^2+(j-j_n^d)^2} \leq \sigma$ and 0 otherwise. We further add an exponential weight to the Gaussian distribution to distinguish close heatmap pixels and obtain the ground truth heatmap $Y_n^d(i,j) = \alpha^{\tilde{Y}_n^d(i,j)}$.

As illustrated in Fig. 2, firstly, the input image from a random batch is partitioned into non-overlapping patches and linearly embedded. Next, these patches are fed into cascaded transformer blocks at each stage, which are merged except in the last stage. Finally, a domain-adaptive convolution decoder makes dense prediction to generate heatmaps, which is further used to extract landmarks via threshold processing and connected components filtering.

3 Experiment

Datasets. For performance evaluation, we adopt three public X-ray datasets from different domains on various anatomical regions of head, hand, and chest. (i) Head dataset is a widely-used dataset for IEEE ISBI 2015 challenge [18,19] which contains 400 X-ray cephalometric images with 150 images for training and 250 images for testing. Each image is of size 2400×1935 with a resolution of $0.1\,\text{mm} \times 0.1\,\text{mm}$, which contains 19 landmarks manually labeled by two medical experts and we use the average labels same as Payer et al. [11]. (ii) Hand dataset is collected by [6] which contains 909 X-ray images and 37 landmarks annotated by [11]. We follow [26] to split this dataset into a training set of 609 images and a test set of 300 images. Following [11] we assume the distance between two endpoints of wrist is 50 mm and calculate the physical distance as $\text{distance}_{\text{physical}} = \text{distance}_{\text{pixel}} \times \frac{50}{\|p-q\|_2}$ where p, q are the two endpoints of the wrist respectively. (iii) Chest dataset [26] is a popular chest radiography database collected by Japanese Society of Radiological Technology (JSRT) [14] which contains 247 images. Each image is of size 2048×2048 with a resolution of $0.175\,\text{mm} \times 0.175\,\text{mm}$. We split it into a training set of 197 images and a test set of 50 images and select 6 landmarks from landmark labels at the boundary of the lung as target landmarks.

Implementation Details. UOD is implemented in Pytorch and trained on a TITAN RTX GPU with CUDA version being 11. All encoders are initialized with corresponding pre-trained weights. We set batch size to 8, σ to 3, and α to 10. We adopt binary cross-entropy (BCE) as loss function for both stages. In stage I, we resize each image to the same shape of 384×384 and train universal convolution model by Adam optimizer for 1000 epochs with a learning rate of 0.00001. In stage II, we resize each image to the same shape of 576×576 and optimize the universal transformer by Adam optimizer for 300 epochs with a learning rate of 0.0001. When calculating metrics, all predicted landmarks are resized back to the original size. For evaluation, we choose model with minimum validation loss as the inference model and adopt two metrics: mean radial error (MRE) $\text{MRE} = \frac{1}{N}\sum_i^N \sqrt{(x_i-\tilde{x}_i)^2+(y_i-\tilde{y}_i)^2}$ and successful detection rates (SDR) within different thresholds t: $\text{SDR}(t) = \frac{1}{N}\sum_i^N \delta(\sqrt{(x_i-\tilde{x}_i)^2+(y_i-\tilde{y}_i)^2} \leq t)$.

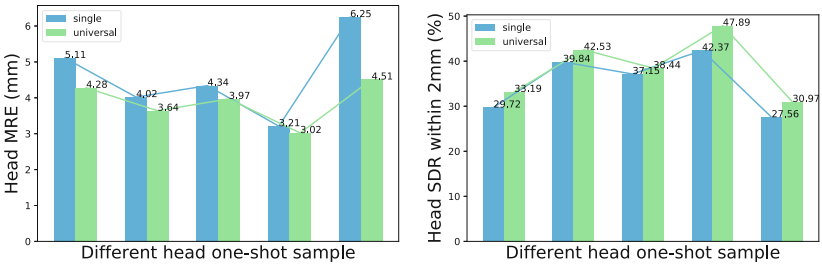


Fig. 3. Comparison of single model and universal model on head dataset.

Table 1. Quantitative comparison of UOD with SOTA methods on head, hand, and chest datasets. * denotes the method is trained on every single dataset respectively while †denotes the method is trained on mixed data.

Method	Label	Head [19]					Hand [6]				Chest [14]			
		MRE↓	SDR↑ (%)				MRE↓	SDR↑ (%)			MRE↓	SDR↑ (%)		
		(mm)	2 mm	2.5 mm	3 mm	4 mm	(mm)	2 mm	4 mm	10 mm	(mm)	2 mm	4 mm	10 mm
YOLO [26]†	all	1.32	81.14	87.85	92.12	96.80	0.85	94.93	99.14	99.67	4.65	31.00	69.00	93.67
YOLO [26]†	25	1.96	62.05	77.68	88.21	97.11	2.88	72.71	92.32	97.65	7.03	19.33	51.67	89.33
YOLO [26]†	10	2.69	47.58	66.47	78.42	90.89	9.70	48.66	76.69	90.52	16.07	11.67	33.67	76.33
YOLO [26]†	5	5.40	26.16	41.32	54.42	73.74	24.35	20.59	48.91	72.94	34.81	4.33	19.00	56.67
CC2D [22]*	1	2.76	42.36	51.82	64.02	78.96	2.65	51.19	82.56	95.62	10.25	11.37	35.73	68.14
Ours†	1	**2.43**	**51.14**	**62.37**	**74.40**	**86.49**	**2.52**	**53.37**	**84.27**	**97.59**	**8.49**	**14.00**	**39.33**	**76.33**

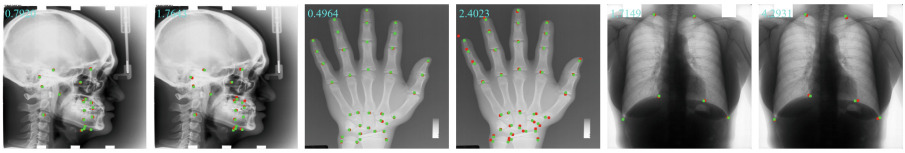
3.1 Experimental Results

The Effectiveness of Universal Model: To demonstrate the effectiveness of universal model for multi-domain one-shot learning, we adopt head and hand datasets for evaluation. In stage I, the convolution models are trained in two ways: 1) single: trained on every single dataset respectively, and 2) universal: trained on mixed datasets together. With a fixed one-shot sample for the hand dataset, we change the one-shot sample for the head dataset and report the MRE and SDR of the head dataset. As Fig. 3 shows, universal model performs much better than single model on various one-shot samples and metrics. It is proved that universal model learns domain-shared knowledge and promotes domain-specific learning. Furthermore, the MRE and SDR metrics of universal model have a smaller gap among various one-shot samples, which demonstrates the robustness of universal model learned on multi-domain data.

Comparisons with State-of-the-Art Methods: As Table 1 shows, we compare UOD with two open-source landmark detection methods, i.e., YOLO [26] and CC2D [22]. YOLO is a multi-domain supervised method while CC2D is a single-domain one-shot method. UOD achieves SOTA results on all datasets under all metrics, outperforming the other one-shot method by a big margin. On the head dataset, benefiting from multi-domain learning, UOD achieves an MRE of 2.43 mm and an SDR of 86.49% within 4 mm, which is comparative with supervised method YOLO trained with at least 10 annotated labels, and much

Table 2. Ablation study of different components of our DATR. Base is the basic transformer block; MSA_{Q^d} denotes the domain-adaptive self-attention and D^d denotes the domain-adaptive diagonal matrix. In each column, the best results are in **bold**.

Transformer	Head [19]					Hand [6]				Chest [14]			
	MRE↓	SDR↑ (%)				MRE↓	SDR↑ (%)			MRE↓	SDR↑ (%)		
	(mm)	2 mm	2.5 mm	3 mm	4 mm	(mm)	2 mm	4 mm	10 mm	(mm)	2 mm	4 mm	10 mm
(a) Base	24.95	2.02	3.17	4.51	5.85	9.83	5.33	16.79	58.64	58.11	0.37	1.96	3.85
(b) $+D^d$	22.75	2.13	3.24	4.61	6.96	7.52	6.13	20.66	68.43	52.98	0.59	2.17	4.68
(c) $+\text{MSA}_{Q^d}$	2.51	49.29	60.89	72.17	84.36	2.72	48.56	80.44	94.38	9.09	12.00	19.33	74.00
(d) $+\text{MSA}_{Q^d}+D^d$	**2.43**	**51.14**	**62.37**	**74.40**	**86.49**	**2.52**	**53.37**	**84.27**	**97.59**	**8.49**	**14.00**	**39.33**	**76.33**



UOD CC2D UOD CC2D UOD CC2D

Fig. 4. Qualitative comparison of UOD and CC2D [22] on head, hand, and chest datasets. The red points • indicate predicted landmarks while the green points • indicate ground truth landmarks. The MRE value is displayed in the top left corner of the image. (Color figure online)

better than CC2D. On the hand dataset, there are some performance improvements in all metrics compared to CC2D, outperforming the supervised method YOLO trained with 25 annotated images. On the chest dataset, UOD shows the superiority of DATR which eliminates domain preference and balances the performance of all domains. In contrast, the performance of YOLO on chest dataset suffers a tremendous drop when the available labels are reduced to 25, 10, and 5. Figure 4 visualizes the predicted landmarks by UOD and CC2D.

Ablation Study: We compare various components of the proposed domain-adaptive transformer. The experiments are carried out in UOD Stage II. As presented in Table 2, the domain-adaptive transformer has two key components: domain-adaptive self-attention MSA_{Q^d} and domain-adaptive diagonal matrix D^d. The performances of (b) and (c) are much superior to those of (a) which demonstrates the effectiveness of D^d and MSA_{Q^d}. Further, (d) combines the two components and achieves much better performances, which illustrates that domain-adaptive transformer improves the accuracy of detecting via cross-domain knowledge and global context information. We take (d) as the final transformer block.

4 Conclusion

To improve the robustness and reduce domain preference of multi-domain one-shot learning, we design a universal framework in that we first train a universal

model via contrastive learning to generate pseudo landmarks and further use these labels to learn a universal transformer for accurate and robust detection of landmarks. UOD is the first universal framework of one-shot landmark detection on multi-domain data, which outperforms other one-shot methods on three public datasets from different anatomical regions. We believe UOD will significantly reduce the labeling burden and pave the path of developing more universal framework for multi-domain one-shot learning.

Acknowledgment. Supported by Natural Science Foundation of China under Grant 62271465 and Open Fund Project of Guangdong Academy of Medical Sciences, China (No. YKY-KF202206).

References

1. Avisdris, N., et al.: BiometryNet: landmark-based fetal biometry estimation from standard ultrasound planes. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 279–289. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_27
2. Browatzki, B., Wallraven, C.: 3fabrec: fast few-shot face alignment by reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6110–6120 (2020)
3. Chen, Z., Qiu, T., Tian, Y., Feng, H., Zhang, Y., Wang, H.: Automated brain structures segmentation from PET/CT images based on landmark-constrained dual-modality atlas registration. Phys. Med. Biol. **66**(9), 095003 (2021)
4. Elkhill, C., LeBeau, S., French, B., Porras, A.R.: Graph convolutional network with probabilistic spatial regression: Application to craniofacial landmark detection from 3D photogrammetry. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 574–583. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_55
5. Espinel, Y., Calvet, L., Botros, K., Buc, E., Tilmant, C., Bartoli, A.: Using multiple images and contours for deformable 3D-2D registration of a preoperative ct in laparoscopic liver surgery. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 657–666. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_63
6. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.: Bone age assessment of children using a digital hand atlas. Comput. Med. Imaging Graph. **31**(4–5), 322–331 (2007)
7. Huang, C., Han, H., Yao, Q., Zhu, S., Zhou, S.K.: 3D U^2-Net: a 3D universal U-net for multi-domain medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 291–299. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_33
8. 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
9. Lang, Y., et al.: DentalPointNet: landmark localization on high-resolution 3d digital dental models. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.)

MICCAI 2022. LNCS, vol. 13432, pp. 444–452. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_43
10. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
11. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal. **54**, 207–219 (2019)
12. Quan, Q., Yao, Q., Li, J., Zhou, S.K.: Which images to label for few-shot medical landmark detection? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20606–20616 (2022)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. **174**(1), 71–74 (2000)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR), pp. 1–14 (2015)
16. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 32–42 (2021)
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Wang, C.W., et al.: Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. IEEE Trans. Med. Imaging **34**(9), 1890–1900 (2015)
19. Wang, C.W., et al.: A benchmark for comparison of dental radiography analysis algorithms. Med. Image Anal. **31**, 63–76 (2016)
20. Wang, Z., et al.: Accurate scoliosis vertebral landmark localization on x-ray images via shape-constrained multi-stage cascaded CNNs. Fundam. Res. (2022)
21. Yao, Q., He, Z., Han, H., Zhou, S.K.: Miss the point: targeted adversarial attack on multiple landmark detection. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 692–702. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_67
22. Yao, Q., Quan, Q., Xiao, L., Kevin Zhou, S.: One-shot medical landmark detection. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 177–188. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_17
23. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8543–8553 (2019)
24. Zhou, S.K., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE (2021)
25. Zhou, S.K., Rueckert, D., Fichtinger, G.: Handbook of Medical Image Computing and Computer Assisted Intervention. Academic Press, Cambridge (2019)

26. Zhu, H., Yao, Q., Xiao, L., Zhou, S.K.: You only learn once: universal anatomical landmark detection. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 85–95. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_9
27. Zhu, H., Yao, Q., Xiao, L., Zhou, S.K.: Learning to localize cross-anatomy landmarks in x-ray images with a universal model. BME Front. **2022** (2022)