Anatomical Landmark Detection Using a Multiresolution Learning Approach with a Hybrid Transformer-CNN Model

Thanaporn Viriyasaranon, Serie Ma, and Jang-Hwan Choi(✉)

Division of Mechanical and Biomedical Engineering, Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, South Korea
choij@ewha.ac.kr

Abstract. Accurate localization of anatomical landmarks has a critical role in clinical diagnosis, treatment planning, and research. Most existing deep learning methods for anatomical landmark localization rely on heatmap regression-based learning, which generates label representations as 2D Gaussian distributions centered at the labeled coordinates of each of the landmarks and integrates them into a single spatial resolution heatmap. However, the accuracy of this method is limited by the resolution of the heatmap, which restricts its ability to capture finer details. In this study, we introduce a multiresolution heatmap learning strategy that enables the network to capture semantic feature representations precisely using multiresolution heatmaps generated from the feature representations at each resolution independently, resulting in improved localization accuracy. Moreover, we propose a novel network architecture called hybrid transformer-CNN (HTC), which combines the strengths of both CNN and vision transformer models to improve the network's ability to effectively extract both local and global representations. Extensive experiments demonstrated that our approach outperforms state-of-the-art deep learning-based anatomical landmark localization networks on the numerical XCAT 2D projection images and two public X-ray landmark detection benchmark datasets. Our code is available at https://github.com/seriee/Multiresolution-HTC.git.

Keywords: Anatomical landmark detection · Multiresolution learning · Hybrid transformer-CNN

1 Introduction

Anatomical landmark detection has been used successfully in parametric modeling [19], registration [22], and quantification of various anatomical abnormalities [6,21]. To detect landmarks automatically and accurately, advanced artificial intelligence technologies, including deep learning with convolutional neural net-

T. Viriyasaranon and S. Ma—Equally contributed.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 433–443, 2023.
https://doi.org/10.1007/978-3-031-43987-2_42

work (CNN)-based [13], transformer-based [7], and graph-convolution methods [8], have been developed and have attracted great interest from both academia and industry.

Generally, deep learning-based anatomical landmark detection is based on heatmap regression approaches [2,3], which decode the predicted landmark coordinates from the heatmap corresponding to the landmarks. In previous studies, the networks mostly generate only one resolution of the heatmap to decode the landmark coordinates. However, deriving the landmark coordinate from a high-resolution heatmap exhibits high bias and low variance, whereas the landmark coordinates obtained from a low-resolution heatmap demonstrate low bias and high variance. Typically, the heatmap regression-based detectors generate high-resolution heatmaps by utilizing the high-resolution coarse feature. However, this process results in the loss of specific landmark-related features, including crucial information regarding the geometric relationship between landmarks. Consequently, this impacts the network's ability to accurately localize landmarks.

In this study, we propose a multiresolution heatmap learning strategy that derives the predicted landmark coordinate from multiresolution heatmaps to balance the bias and variance of the predicted landmarks. Moreover, leveraging multiresolution feature representations to generate the heatmap can effectively increase the localization accuracy of the deep learning network.

Typically, the existing methods of anatomical landmark detection are formulated using CNN-based or transformer-based encoder-decoder architecture [16]. The convolution operation collects information by layer, which focus on the local feature information. Meanwhile, the vision transformer has the ability to encode global representations. To combine the advantages of CNNs and transformers, we introduce a novel hierarchical hybrid transformer and CNN architecture called the hybrid transformer-CNN (HTC). HTC introduces a stack of convolutional and transformer modules, which are applied to all stages of the encoder for extracting global information, and local information. Furthermore, we propose a lightweight positional-encoding-free transformer module. Instead of using multi-head attention, we introduce the bilinear pooling operation to capture second-order statistics of features and generate global representations. Moreover, general transformer encoders suffer from the fixed resolution of positional encoding, which results in decreased accuracy when interpolating the positional encoding during testing with resolutions different from the training data. To alleviate this problem, we remove the positional encoding from the transformer modules and employ a 3×3 convolutional operation as the patch embedding to capture location information and generate low-resolution fine features for the hierarchical encoder architecture design.

The main contributions of this paper are as follows:

– Introduction of a multiresolution heatmaps learning strategy, which increases the detection ability of the network by leveraging multiresolution information to derive the predicted landmark.

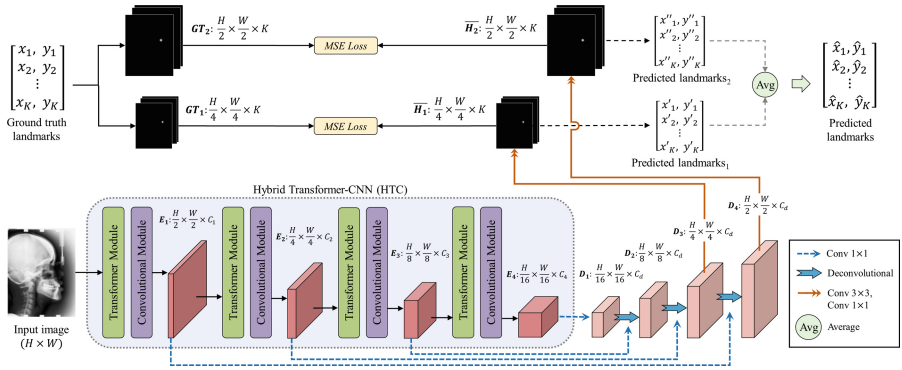

Fig. 1. Illustration of the overall architecture of the anatomical detector, including multiresolution heatmap learning and the hybrid transformer-CNN (HTC).

– Development of a hierarchical hybrid transformer and CNN architecture named the HTC, which sequentially combines transformer and convolutional modules.
– Our proposed HTC model trained with a multiresolution heatmap learning approach clearly outperforms previous state-of-the-art models on three datasets: XCAT 2D projections of head CBCT volumes, X-ray dataset from ISBI2023 Challenge [1], and hand X-ray dataset [13].

2 Methods

In this section, we introduce our anatomical landmark detector, which consists of the proposed multiresolution heatmap learning method and the HTC backbone network, as shown in Fig. 1.

2.1 Multiresolution Heatmap Learning

As depicted in Fig. 1, we generate the multiresolution prediction heatmap using multiresolution feature representations (D_i) from the decoder layers and extra convolutional layers. The output feature representations of the decoder layers are a combination of the feature representations from each of the stages of the encoder and upsampled feature representations from the previous stage. The number of channels of all stages of the decoder feature representations are set to be equal to C_d. In this work, we defined C_d as 256.

The lowest-resolution feature representation of the encoder E_4 of size $\frac{H}{16} \times \frac{W}{16} \times C_4$ is passed through 1×1 convolutional operation, and the output feature representation D_1 has size $\frac{H}{16} \times \frac{W}{16} \times C_d$. Then, D_1 is upsampled using 3×3 deconvolution operations and aggregated with the encoder feature representations E_3 to generate the next stage feature representations D_2 having size $\frac{H}{8} \times \frac{W}{8} \times C_d$.

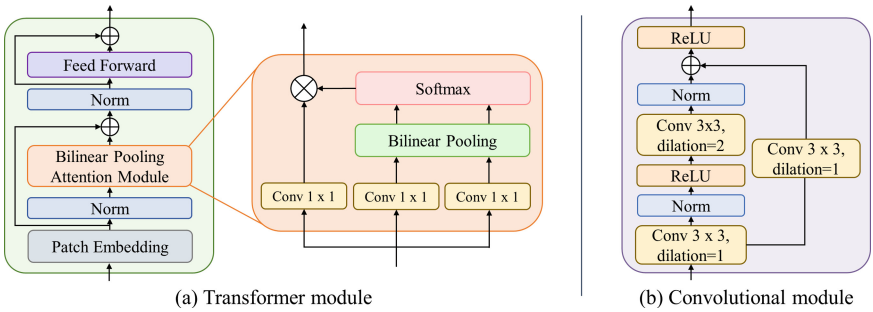

Fig. 2. Architecture of the components within each stage of the HTC (a) transformer modules; (b) convolutional modules.

Similarly, the high-resolution decoder feature representations D_3 and D_4 are the combinations of the previous decoder features D_2 and D_3, and encoder feature representation E_2 and E_1, respectively.

In this study, we generated predicted heatmaps H_1 and H_2 with two different resolution sizes $\frac{H}{4} \times \frac{W}{4} \times K$ and $\frac{H}{2} \times \frac{W}{2} \times K$ by utilizing 3×3 and 1×1 convolutional operations with decoder feature representations D_3 and D_4, respectively. Then, H_1 and H_2 are generated as follows:

$$\bar{H}_1 = Conv_1(Conv_3(D_3)),$$
$$\bar{H}_2 = Conv_1(Conv_3(D_4)),$$

where $Conv_1$ and $Conv_3$ are 1×1 and 3×3 convolutional operations, respectively. In addition, K is the number of landmarks.

During training, we use mean squared error (MSE) as the loss function between the predicted and ground-truth heatmaps for each resolution. Moreover, we enforce the detector to learn global and local information from the heatmap generated by the high-resolution coarse feature and the low-resolution fine-grained features, through the weighted summation of the heatmap loss from each resolution heatmap as follows:

$$\mathcal{L} = \mathcal{L}_{H_1} + \lambda \mathcal{L}_{H_2}, \tag{1}$$

where $\mathcal{L}_{H_1}$ and $\mathcal{L}_{H_2}$ are the respective losses calculated from the predicted heatmap H_1 and ground truth heatmap of size $\frac{H}{4} \times \frac{W}{4} \times K$ as well as the predicted heatmap H_2 and ground truth heatmap of size $\frac{H}{2} \times \frac{W}{2} \times K$. Additionally, λ is the loss weight, which is set as three in all the experiments. During inference, we calculate the output landmark prediction coordinates of the model by averaging the corresponding coordinates decoded from the heatmap at each resolution.

2.2 Hybrid Transformer-CNN (HTC)

We introduce a novel hierarchical encoder architecture named HTC, which consists of four stages of a stack of the convolutional $ConvU_i$ and transformer mod-

ules $TransU_i$ to generate multi-level feature representation. The overall architecture design of the HTC is shown in Fig. 1. At each stage, the transformer module captures global information such as the geometric relation between landmarks, while the convolutional module extracts local information. The architecture of the transformer and convolutional module are shown in Fig. 2 (a) and (b), respectively. In this study, we proposed a lightweight positional-encoding-free transformer module. Instead of using multi-head attention, we introduce the bilinear pooling operation to capture second-order statistics of features and generate global representations. Moreover, we eliminate the usage of positional encoding to address the challenge of reduced model performance when testing the model with resolutions that differ from the training data. This modification aims to mitigate the issue and improve the overall performance under varying resolution scenarios.

In the first stage, an input image of size $H \times W \times 3$ is fed to the patch embedding, which is a convolutional 3×3 operation to obtain a patch token of size $\frac{H}{2} \times \frac{W}{2} \times C_1$. Then, the patch token is passed through the bilinear pooling attention module, which requires three inputs: a query Q, a key K, and a value V, which are the patch token that passes through the 1×1 convolutional operations, separately. Then, Q, K, and V are flattened to size $\frac{HW}{2^2} \times C_1$. The key and query are then fed to bilinear pooling [9], which is an effective way of gathering the key features and capturing the global representations of the images as follows:

$$F = \sum_{n \in N} K_n Q_n{}^{\intercal}, \tag{2}$$

where N is the set of spatial locations (combinations of rows and columns). We further applied the softmax function to the output of bilinear pooling F to generate the attention weighting vector. Thereafter, matrix multiplication was performed between F and V, and the output of the bilinear pooling attention module was passed through to the feed-forward layer [23]. Then, the output of the feed-forward layer was reshaped as feature representations G_i of size $\frac{H}{2} \times \frac{W}{2} \times C_1$. In addition, the output of the transformer module $TransU_i$ was fed to the convolutional module $ConvU_i$, comprising 3×3 convolutional layers with dilated rates equal to one and two ($Conv_{3,d1}$ and $Conv_{3,d2}$), a batch normalization operation ($Norm$), and a rectified linear unit ($ReLU$) activation function as follows:

$$E_i = ReLU(Conv_{3,d1}(G_i) + Norm(Conv_{3,d2}(ReLU(Norm(Conv_{3,d1}(G_i)))))). \tag{3}$$

Similarly, using the feature representations from the previous stages as inputs, we obtained E_2, E_3, and E_4 with spatial reduction ratios of 4, 8, and 16 pixels, respectively, with respect to the input image.

3 Experiments

3.1 Dataset

To evaluate our method, we conducted experiments on a total of three datasets, including one 4D XCAT phantom CT dataset and two public X-ray datasets. Here, we generated head models from the 4D XCAT dataset [17] for 27 patients with varying anatomical sizes and genders. We manually labeled 13 cephalometric landmarks on CT phantom volumes. Moreover, we perform forward projection on both the 3D phantom CT volumes and landmarks at 360 angles per patient to obtain 2D images and landmark labels. We randomly selected 70% of the patients' CT scans as the training dataset (18 patients) and the remaining 30% as the test dataset (9 patients). The size of each image was 620×480, with a pixel spacing of 0.66 mm.

We also evaluated our method on the public X-ray dataset from the IEEE ISBI 2023 Challenge [1]. A total of 29 ground truth landmarks were labeled by two experts. The image sizes and pixel spacings vary over patients. We randomly selected 75% of the X-ray images of the provided training dataset as the network training dataset(525 images) and the remaining 25% as the test dataset.

Finally, we performed experiments on a public hand dataset containing X-ray images from 895 patients and having 37 landmarks [13]. The sizes of these images are not all the same, so we resized the images to 1024×1216. Owing to their lack of physical pixel resolution, we calculated the pixel spacing based on the assumption that the distance between two landmarks at both endpoints of the wrist is approximately 50 mm.

3.2 Implementation Details

In our experiments, we implemented our framework using MMPose [4], an open-source toolbox for pose estimation based on PyTorch. For XCAT CT landmark detection, our HTC with multiresolution heatmap learning was trained for 50 epochs using the AdamW optimizer with the initial learning rate set to 0.0003. Furthermore, we trained the ISBI2023 landmark detection method for 180 epochs using the AdamW optimizer, with an initial learning rate of 0.00045. In addition, for hand landmark detection, we trained for 300 epochs using the AdamW optimizer at an initial learning rate of 0.0004. For all the experiments, the evaluation metrics are the mean radial error (MRE, mm) and successful detection rate (SDR, %) under 2 mm, 2.5 mm, 3 mm, and 4 mm conditions.

3.3 Performance Evaluation

In this section, we compare the performances of the proposed and state-of-the-art methods as well as analyze ablation studies on the proposed method. In each of the tables below, the metrics showing the best and second-best performances are indicated by boldface and underlined, respectively.

Table 1. Performance comparison of the proposed method with previous state-of-the-art methods on the XCAT CT and ISBI2023 datasets.

Model	#Param(M)	XCAT CT dataset					ISBI2023 dataset				
		MRE(SD)	SDR(%)				MRE(SD)	SDR(%)			
			2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm
Hourglass [11]	94.85	5.35(10.21)	38.71	49.54	58.36	71.19	1.32(1.88)	82.90	88.41	92.12	95.53
HRNet-W48 [20]	65.33	3.91(6.22)	36.38	48.47	58.56	73.01	1.26(1.49)	83.94	88.85	92.26	96.00
HRFormer-S [26]	44.04	3.40(2.82)	36.50	48.58	58.71	72.67	1.34(1.45)	81.83	88.41	91.74	95.70
UNet [16]	35.35	4.08(9.49)	**47.28**	58.44	66.47	76.81	3.97(15.04)	82.13	86.58	89.28	92.22
PVT-Tiny [24]	16.91	5.93(7.72)	18.18	26.17	34.39	49.48	2.00(4.77)	72.99	80.43	85.34	90.54
Conformer-Ti [15]	22.32	3.88(4.17)	42.26	31.65	52.12	66.99	1.70(2.93)	75.69	83.23	87.35	93.10
GU2Net [27]	2.74	5.96(15.36)	38.66	48.41	56.39	66.67	1.78(5.10)	81.97	86.31	88.87	92.28
FARNet [2]	78.97	3.51(3.57)	35.62	46.54	56.21	71.33	1.10(**1.33**)	87.59	91.88	94.48	**96.87**
AFPF [3]	20.68	4.59(11.58)	41.57	52.98	61.74	72.51	4.13(16.35)	79.09	83.67	86.72	90.05
Multi-task UNet [25]	13.58	4.27(8.71)	40.80	51.46	59.74	71.87	5.91(15.63)	74.58	79.29	81.81	84.71
HTC+Multiresolution learning	16.20	**2.88**(**2.51**)	46.82	**59.02**	**67.97**	**79.68**	**1.08**(1.37)	**88.05**	**92.17**	**94.50**	96.69

Table 2. Performance comparison of the proposed method with other state-of-the-art models on the hand dataset. The symbol " * " indicates that we repeated the training with the same environment used to develop our method.

Model	Hand dataset			
	MRE(SD)	SDR(%)		
		2 mm	4 mm	10 mm
Payer et al. [12]	1.13(0.98)	87.60	98.66	99.96
GIRRF [5]	0.97(2.45)	91.60	97.84	99.31
DATR [28]	0.86(-)	94.04	99.20	99.97
Lindner et al. [10]	0.85(1.01)	93.68	98.95	99.94
Urschler et al. [21]	0.80(0.93)	92.19	98.46	98.46
Štern et al. [18]	0.80(0.91)	92.20	98.45	99.83
FARNet* [2]	0.67(0.64)	95.65	99.58	99.99
Payer et al. [14]	0.66(0.74)	94.99	99.27	99.99
HRNet-W48* [20]	0.65(0.62)	96.10	99.60	**100**
GU2Net* [27]	0.63(1.36)	96.01	99.39	99.98
HTC+Multiresolution learning	**0.56**(**0.58**)	**96.84**	**99.63**	**100**

Comparisons with State-of-the-art Methods: Performance comparisons on the XCAT CT and ISBI2023 dataset are shown in Table 1. We compared our proposed model with both natural and medical-domain landmark detectors. Our method remarkably achieved the best performance on the XCAT CT dataset with an MRE of 2.88 mm and on the ISBI2023 dataset with an MRE of 1.08 mm. Furthermore, our approach showed effects in lowering the standard deviation of the mean error on the XCAT CT dataset and improved the percentage of successful detection by 2.87%. Moreover, our method outperformed the highest detection rate among previous studies, reaching an SDR of 88.05% on the ISBI2023 dataset. Furthermore, the HTC was intentionally designed to have a smaller number of parameters compared to transformer-based architectures such as HRFormer and Conformer, despite achieving superior detection accu-

racy. Table 2 presents the performance comparison between the proposed and existing methods on the hand dataset. Our method significantly outperformed the best performance for all metrics on the hand dataset, with 0.56 mm MRE and 0.58 mm standard deviation of MRE. Additionally, qualitative comparisons of the images of the detection results are shown in Fig. 3.

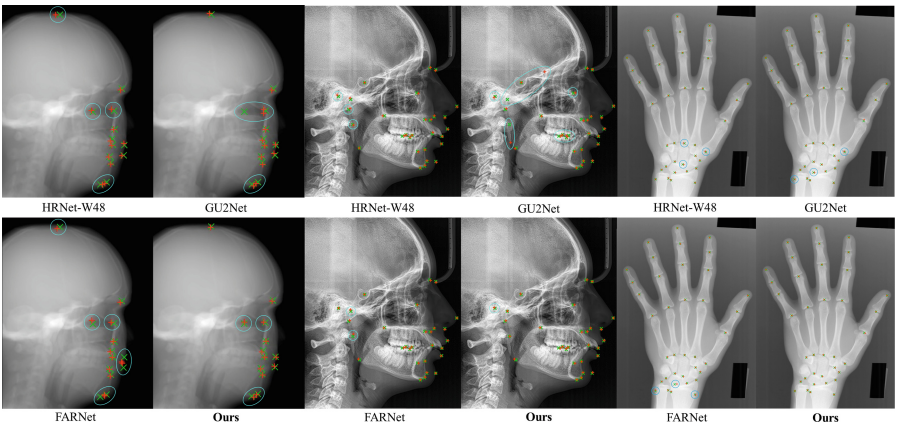

Fig. 3. Comparison of the proposed method and other models on the XCAT CT, ISBI2023, and hand datasets (from left to right). Ground-truth landmarks are marked in green, and the predictions are marked in red. We mark landmarks that do not overlap between predictions and ground truth with circles. (Color figure online)

Table 3. Comparison of the effects of the proposed HTC and multiresolution heatmap learning.

Model	Multi-resolution learning	XCAT CT dataset					ISBI2023 dataset				
		MRE(SD)	SDR(%)				MRE(SD)	SDR(%)			
			2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm
Conformer [15]	✗	3.88(4.16)	42.26	31.65	52.11	66.98	1.70(2.93)	75.69	83.23	87.35	93.10
HTC	✗	3.03(2.59)	43.64	55.72	64.73	76.99	1.11(1.37)	87.56	**91.80**	**94.26**	**96.65**
	✓	**2.88**(**2.51**)	**46.82**	**59.02**	**67.97**	**79.68**	**1.08**(**1.35**)	**87.80**	**91.80**	94.18	96.45

Ablation Study: We conducted an additional study to observe the effects of the proposed multiresolution heatmap learning and HTC. For this study, we compared the HTC with Conformer [15], which is a representative hybrid transformer and CNN architecture. As shown in Table 3, the HTC outperforms Conformer with MRE values of 3.03 and 1.11 mm for the XCAT CT and ISBI2023 datasets, respectively. Furthermore, the implementation of multiresolution heatmap learning can enhance the MRE of 0.15 and 0.03 mm for the XCAT CT and ISBI2023 datasets, respectively.

4 Conclusion

This study presents a new feature extraction architecture, referred to as the hybrid transformer-CNN (HTC), along with multiresolution heatmap learning for automatic anatomical landmark detection. The HTC architecture comprises of multiple stages of stacked transformer modules, which incorporate a bilinear pooling attention module to capture the global information of images and convolutional modules to extract local and specific feature representations relevant to landmarks. Additionally, we introduced multiresolution heatmap learning to improve the network's ability to capture global and local representations more accurately than learning from a single heatmap resolution, thereby enhancing network localization. Our experimental evaluations on three benchmark datasets demonstrate that the proposed method surpasses state-of-the-art approaches across various modalities and anatomical regions. These findings highlight the potential of our method for automatic anatomical landmark detection in various medical applications.

Acknowledgements. This research was partly supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF-5199990614253); by the Technology development Program of MSS [S3146559]; by the National Research Foundation of Korea (NRF-2022R1A2C1092072); and Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2022-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)).

References

1. Anwaar Khalid, M., et al.: CEPHA29: automatic cephalometric landmark detection challenge 2023. arXiv e-prints arxiv.org/abs/2212.04808 (2022)
2. Ao, Y., Wu, H.: Feature aggregation and refinement network for 2D anatomical landmark detection. J. Digit. Imaging **36**(2), 547–561 (2022). https://doi.org/10.1007/s10278-022-00718-4
3. 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
4. Contributors, M.: OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)
5. Ebner, T., Stern, D., Donner, R., Bischof, H., Urschler, M.: Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 421–428. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10470-6_53
6. Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T.: Shape representation for efficient landmark-based segmentation in 3-D. IEEE Trans. Med. Imaging **33**(4), 861–874 (2014)

7. 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.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol. 13433, pp. 227–237. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_22
8. Lang, Y., et al.: Automatic localization of landmarks in craniomaxillofacial CBCT images using a local attention-based graph convolution network. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 817–826. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_79
9. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1457 (2015)
10. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1862–1874 (2014)
11. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
12. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using CNNs. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 230–238. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_27
13. 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)
14. 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)
15. Peng, Z., et al.: Conformer: local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 367–376 (2021)
16. 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
17. Segars, W.P., Sturgeon, G., Mendonca, S., Grimes, J., Tsui, B.M.: 4D XCAT phantom for multimodality imaging research. Med. Phys. **37**(9), 4902–4915 (2010)
18. Štern, D., Ebner, T., Urschler, M.: From local to global random regression forests: exploring anatomical landmark localization. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 221–229. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_26
19. Štern, D., Likar, B., Pernuš, F., Vrtovec, T.: Parametric modelling and segmentation of vertebral bodies in 3D CT and MR spine images. Phys. Med. Biol. **56**(23), 7505 (2011)
20. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)

21. Urschler, M., Ebner, T., Štern, D.: Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med. Image Anal. **43**, 23–36 (2018)
22. Urschler, M., Zach, C., Ditt, H., Bischof, H.: Automatic point landmark matching for regularizing nonlinear intensity registration: application to thoracic CT images. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 710–717. Springer, Heidelberg (2006). https://doi.org/10.1007/11866763_87
23. Vaswani, A., et al.: Attention is all you need. ArXiv abs/1706.03762 (2017)
24. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
25. 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
26. Yuan, Y., et al.: HRFormer: high-resolution vision transformer for dense predict. Adv. Neural. Inf. Process. Syst. **34**, 7281–7293 (2021)
27. 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
28. Zhu, H., Yao, Q., Zhou, S.K.: DATR: domain-adaptive transformer for multi-domain landmark detection. arXiv preprint arXiv:2203.06433 (2022)