3D Image Landmarks Localization Guided by Medical Prior Knowledge

Yijie Pang[1] 12232142@mail.sustech.edu.cn
[1] *Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China*
[2] *Department of Radiology, Shenzhen Second People's Hospital, Shenzhen, China*

Pujin Cheng[1] 12032946@mail.sustech.edu.cn
Junyan Lyu[1] 12063003@mail.sustech.edu.cn
Jialin Li[1] 12222222@mail.sustech.edu.cn
Fan Lin[*2] fanlin@sustech. edu. cn
Xiaoying Tang[*1] tangxy@sustech. edu. cn

Editors: Under Review for MIDL 2023

Abstract

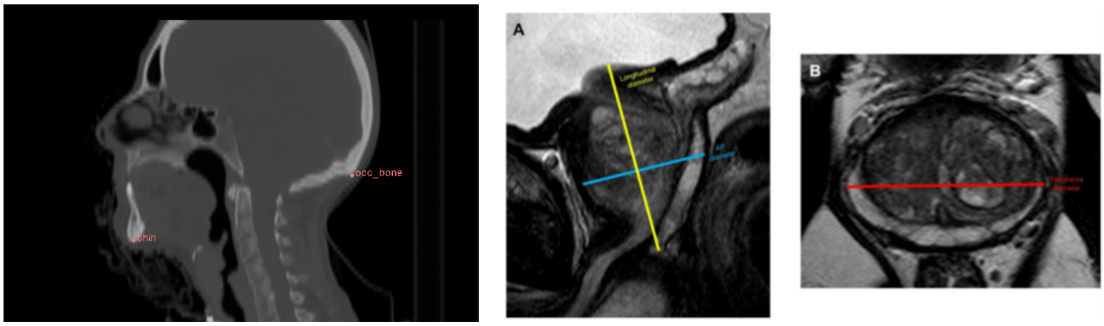
Accurate detection of 3D landmarks is critical for evaluating and measuring organ features and performing preoperative diagnostic procedures. However, detecting these landmarks can be challenging due to local structural similarities, which can cause ambiguities in the annotation. To address this issue, doctors often annotate multiple landmarks in a single slice, particularly when estimating 3D distance and volume. In this study, we present a prior guided coarse-to-fine framework for efficient and accurate 3D landmark detection.We use coordinate regression on downsampled 3D images in the coarse stage to maintain the structural relationships between landmarks. In the fine stage, we categorize landmarks as either correlated or independent based on their manual annotation status. For independent landmarks, we train multiple models to capture local features and ensure reliable local predictions. For correlated landmarks, we replicate the manual annotation process by querying key frames to guide the landmarks annotation. We propose a correlated landmark detection model that merges information from various patches to query key slices and identify landmarks. Our method is extensively evaluated on two datasets and demonstrates superior performance, with an average detection error of 3.28 mm and 2.16 mm, respectively.

Keywords: 3D landmark detection, coarse-to-fine, feature fusion module

1. Introduction

Anatomical landmark localization is crucial in various medical fields, such as orthodontic and maxillofacial surgery planning (Wang et al., 2016), organ volume estimation (Turkbey et al., 2019), and skeletal development assessment (Escobar et al., 2019). Manual annotation of anatomical landmarks is cumbersome and non-producible. Repeatable and accurate automated landmark detection can free physicians from tedious labeling tasks and is useful in clinical applications. Unfortunately, Local structural similarity often leads to clarity in landmark localization, especially in 3D image landmark annotation, where only one

© 2023 Y. Pang, P. Cheng, J. Lyu, J. Li, F. Lin & X. Tang.



(a) Cephalograms Landmarks

(b) Prostate Landmarks

Figure 1: Sample image with correlated landmarks.

slice can be viewed at a time. This will make it difficult to locate landmarks accurately. To minimize labeling errors, experts often annotate landmarks on a single slice using a distinctive landmark as a reference (Wang et al., 2016) or annotate multiple landmarks in key slices (Turkbey et al., 2019; Tan et al., 2022). In these cases, accurately identifying key slices can improve the accuracy of correlated landmark localization and facilitate the observer's analysis of disease.

With the emergence of 3D convolutional neural networks (3D CNNs) (Çiçek et al., 2016), 3D landmark detection methods have progressed from key slice classification (Yang et al., 2015) to coordinate regression or heatmap regression with 3DCNN. However, processing 3D volume data is computationally expensive as it requires more network parameters and computational resources. 3D images are often downsampled to a lower resolution for heatmap regression. Zhang established a multitask model for image volume classification and landmark localization (Zhang et al., 2020a).Liu predicted the landmarks and edges simultaneously for THA preoperative planning (Liu et al., 2020). Payer introduced a method that combines local appearance with spatial configuration for landmark regression and achieved excellent results on multiple datasets. (Payer et al., 2019). These approaches involve downsampling the image to a low resolution, which can lead to the loss of local texture information and cause errors when decoding from the regressed heatmap. In the human landmarks detection task, numerous studies have demonstrated that using high-resolution information can improve the localization accuracy (Zhang et al., 2020b).

Coarse-to-fine architectures can address the accuracy degradation caused by downsampling. Tao employed a transformer-based structure for coarse landmark detection with a UNet model for heatmap regression in the fine stage (Tao and Zheng, 2021). Zeng proposed three-stage coordinate regression network progressively utilizes global and local attention (Zeng et al., 2021). Lee integrated global and feature extractions and the fusion of all the features in a single-passing CNN (Lee et al., 2022). Chen proposed a Structure-Aware Long Short Term Memory(SA-LSTM) framework , which can fuse the multi-resolution information to continuously adjust the landmark offsets(Chen et al., 2022). Most of these methods use coordinate regression in the second stage, which can produce predictions with reasonable global shapes (Jin et al., 2021). However, this method is inaccurate in local area localization

due to the loss of spatial information caused by global average pooling (GAP) (Mao et al., 2022). How to utilize the high localization accuracy characteristics of heatmap regression in the second stage while maintaining the structural constraints between landmarks has yet to be explored.

Inspired by these aforementioned works and the observer's annotating habit, we propose a coarse-to-fine framework to address the abovementioned issues. In the coarse stage, we adopt coordinate regression to preserve structural constraints between landmarks (Jin et al., 2021). Coordinate regression maintain the shape structure's stability, even if local information around the landmark is lost due to downsampling. In the fine stage, we divide the landmarks into correlated and independent landmarks based on how experts annotate them. For independent landmarks, experts focus on local texture features, so we train multiple models to concentrate on the local features of each landmark. For correlated landmarks, these landmarks are distributed in key slices. To boost accuracy, we propose an axial attention guide information fusion module and key slice detection module for correlated landmark regression. These two modules filter features between different patches to achieve higher accuracy in key slice detection and guide landmark detection in different patches.

Our contributions can be summarized as follows: (1) Based on the characteristics of coordinate regression and heatmap regression, we propose a new coarse-to-fine structure. (2) We design a correlated landmark localization module to detect correlated landmarks. (3) We verify our method on two datasets, and the experimental outputs show our approach has excellent performance.

2. Method

As shown in Figure 2, we use a coarse-to-fine architecture for landmark detection. In the coarse stage, we train a ResNet34 (He et al., 2016) for coordinate regression with the global field. To mitigate the accuracy drop caused by down-sampling, the fine stage searches in the proposal regions with patch-based U-Net models, regressing the heatmap patches at a high resolution. To further improve performance, we design different regression models for different types of landmarks.

2.1. Coarse stage

The coarse stage takes the entire downsampled images as input and informs the underlying global landmark configuration. To maintain the shape stability of the landmarks distribution, we train a 3D ResNet for coordinate regression. Given an entire image containing n landmarks with ground truth, we estimate a regression loss:

$$L_{reg} = \frac{1}{n}\sum_{i=0}^{n}(|x_i - \hat{x}_i| + |y_i - \hat{y}_i| + |z_i - \hat{z}_i|) \tag{1}$$

In this stage, ambiguous landmarks can be predicted based on the locations of other easily detectable landmarks using the fixed shape context. Retention of structural information avoids missing predictions, large deviation predictions, and maintains the stability of the results. However, the compression of spatial information by GAP (Mao et al., 2022) and

the loss of local features due to down-sampling limits the accuracy of prediction. To boost accuracy, we crop the original data according to the predicted position and process the data at a higher resolution in the fine stage.

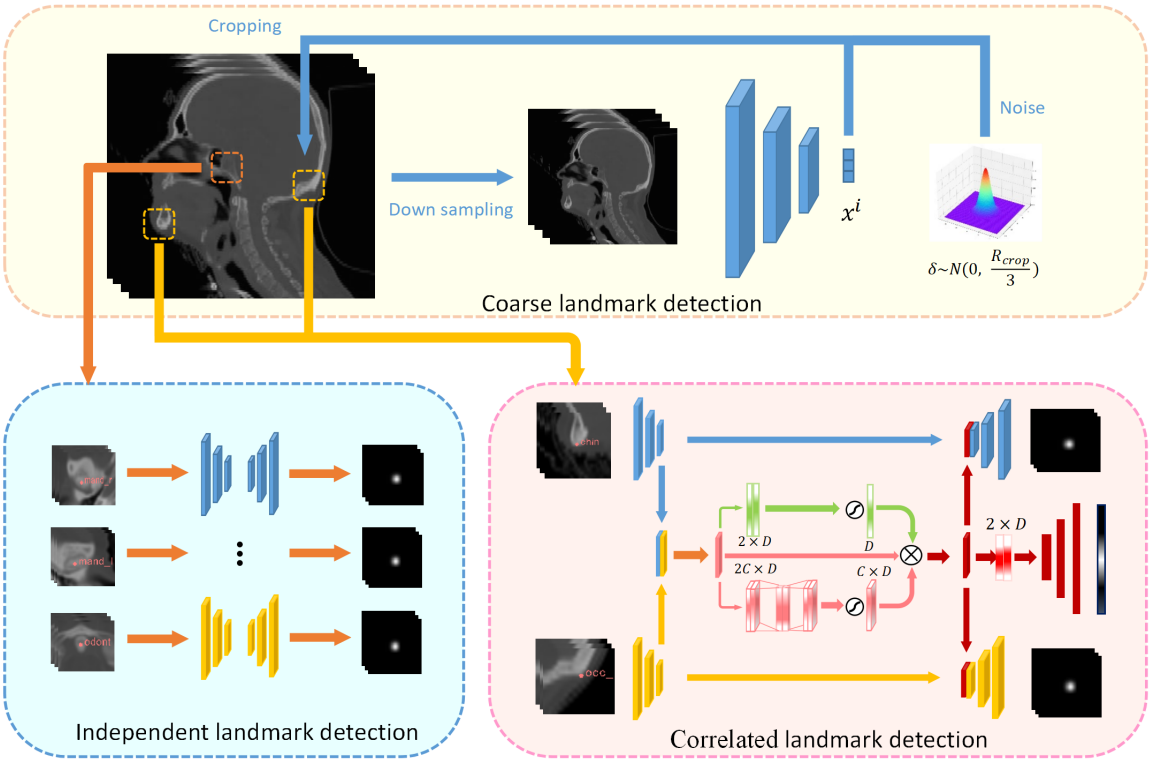


Figure 2: Overall framework of our prior guided corse-to-fine model. See text for detailed explanation

2.2. Fine stage

The fine stage with patch-based U-Net (Çiçek et al., 2016) focuses on learning the local appearance around landmarks. The local stage U-Net has a better distinguishing ability for high-resolution local features. It is noteworthy that cropping patches centered on x_i will result in a lack of diversity of training samples. x_i is the result of the prediction in the coarse stage for the i-th landmark. Inspired by (Chen et al., 2022), we add Gaussian noise $\delta^i \sim$ N(0 , σ) to the cropping centrals for i-th landmark. We set σ to r/6 to ensure the patches contain landmarks, and r is the radius of the cropping patches. Note that for correlated landmarks distributed in the same slice, the cropping centrals of correlated landmarks are in the same slice.

We divide landmarks into two types to improve performance based on the physician labeling process. For independent landmarks, physicians annotate them based on local edge features. As shown in the independent landmark detection module of Figure 2, multiple patch-based Unet models are used for heatmap regression on these specific regions. For

the correlated landmarks, physicians need to find particular slices containing the complete organ's characteristics and annotate landmarks on these slices. We designed a correlated landmark detection model to fuse the features of different patches for key slice querying. $L_{heatmap}$ is formulated by Mean Square Error (MSE) that is consistent with previous literature (Chen et al., 2022). Finally, we use the DARK (Zhang et al., 2020b) to decode the heatmaps and add the cropping centrals as the final prediction result.

Axis attention fusion:

Since we utilize different encoders to deal separately with heatmap regression, we aim at fusing features extracted by correlated patch encoders for key slices querying. We hope that the feature fusion module can guide the positioning of ambiguous landmarks through correlated landmarks. Inspired by CBAM (Woo et al., 2018), we designed an axis attention fusion module to perform dynamic weighted refining and multi-patch feature fusion. Axis is defined as the direction perpendicular to the key slice.

In Figure 2, we use a combination of average pooling and max-pooling to extract spatial information from the feature map fused by a 3x3 convolution layer. Generating two spatial context descriptors: F^s_{max} and F^s_{avg}. Unlike CBAM, we only retain axial features rather than all spatial information. This allows us to focus on the relevant features for key slice detection and helps improve our model's accuracy. After sigmoid activation, we obtain spatial attention map $M_s \in R^{1\times1\times D}$. This axial information characterizes the probability of the optimal slice distribution. Inspired by Coordinate Attention (Hou et al., 2021) we design an axis channel attention module to exploit the inter-channel relationship of axial features. We also use avg-pooling and max-pooling operations to aggregate features along the axis direction, yielding a pair of direction-aware channel feature maps:F^c_{max} and F^c_{avg}. This method can preserve precise positional information along the axis direction. After the shared MLP and sigmoid activation, we merge the output feature vectors with element-wise summation and produce our axis channel attention map $M_c \in R^{C\times1\times1\times D}$. In short, the axis spatial, axis channel attention and the output of our attention block can be written as :

$$M_s = \sigma(cov^3[F^s_{max}, F^s_{avg}]) \tag{2}$$

$$M_c = \sigma(MLP(F^c_{max}) + MLP(F^c_{avg})) \tag{3}$$

$$F = F \times M_s \times M_c \tag{4}$$

Heatmap regression and Key slice detection:

After filtering and fusion, the features can assist landmark regression and search for critical slices. We concatenate the fused features and patch features to guide patch-based heatmap regression, improving localization accuracy in the axial direction. We introduce a slice detection branch to determine the key slice, which helps to constrain feature filtering and enables doctors to more accurately assess the volume, distance, and shape of bones or organs. Like axis channel attention, we preserve precise positional information along the axis direction and use 1D convolution to decode high-level features for key slice detection. Unlike the hard labels of traditional classification tasks, we use a heatmap to represent the probability of where the key slice is located. The loss function for slice detection, L_{slice} is formulated using MSE. The overall loss function for the second stage is:

$$L_{second} = \lambda_0 L_{heatmap} + \lambda_1 L_{slice} \tag{5}$$

3. Experiments and Results

3.1. Dataset

In-house prostate Dataset:
The dataset consists of 857 T2 MRI scans, each with a slice thickness of 1.5 mm and a slice spacing of 0.67mm. The scans are saved in NII format and contain six prostate landmarks annotated by an expert for prostate volume estimation (Turkbey et al., 2019): the endpoints of the maximum longitudinal diameter, maximum anteroposterior diameter, and maximum transverse diameter. The dataset is randomly divided into 514 training scans, 171 validation scans, and 172 test scans.

PDDCA Dataset:
The PDDCA dataset consists of 48 CT images from the Radiation Therapy Oncology Group (RTOG) 0522 study. These images are in NRRD format with various resolutions and spacings. The dataset was used in the MICCAI Head-Neck Challenge 2015 (Raudaschl et al., 2017), and 33 of the images were labeled with five bony landmarks for gross alignment evaluation. The landmarks are: Chin (chin), Right condyloid process (mand r), Left condyloid process (mand l), Odontoid process (odont proc), and Occipital bone (occ bone). Similar to previous work (Chen et al., 2022), we resize the images to 518×518×384 with an isotropic voxel spacing of 1 mm. The dataset is equally divided into three folds for three-fold cross-validation. We report the average accuracy of best results across the three folds.

3.2. Implementation details

Our method is implemented using PyTorch and trained on an NVIDIA GeForce RTX 2080 Ti GPU. The learning rates for the first and second stages are set to 1e-3 and 1e-6, respectively. We train our network for 50 epochs on the prostate dataset and 300 epochs on the PDDCA dataset. We use the SGD optimizer with a weight decay of 1e-8. For training data augmentation, we apply random translation (10% of the input size), random scaling (up to 10% difference), and random intensity shift (up to 0.1). Note that we do not perform random rotation. We set the input size of the first stage image to 96×96×72 for PDDCA and 64×64×64 for the prostate dataset. To balance the trade-off between memory usage and coverage area, we set the crop size to 48 in the second stage. For the PDDCA dataset, we use a cropping resolution of 2mm, and for the in-house dataset, we use a cropping resolution of 1mm.

3.3. Comparison to SOTA

To evaluate the performance of our proposed method, we conduct comprehensive experiments on two datasets: the PDDCA dataset (Raudaschl et al., 2017) and an in-house prostate dataset. We use the evaluation metrics proposed in (Chen et al., 2022), including the mean radial error (MRE), standard deviation (SD), and successful detection rate (SDR) at five target radii (2 mm, 2.5 mm, 3 mm, 4 mm, 8 mm). We compare our method to three deep learning-based methods: 3D Unet (Çiçek et al., 2016), SCN (Payer et al., 2019), and DRM (Zhong et al., 2019). For the one-stage methods, the 3D image is downsampled to

the maximum size allowed by memory as input. The two-stage methods use the same hyperparameters as our method, such as patch size and down-sampled volume size.

Table 1: Performance comparisons between our proposed method and other SOTA methods on the PCDDA and In-house dataset. For PDDCA dataset, we present the results of SA-LSTM (Chen et al., 2022)

Dataset	method	MRE(SD)	2mm	2.5mm	3mm	4mm	8mm
PDDCA	3D-Unet	7.69(5.24)	2.03	3.25	5.05	16.28	67.9
	SCN	7.44(4.26)	2.65	6.74	10.98	21.36	69.3
	DRM	6.39(3.37)	7.27	12.72	16.36	29.09	74.54
	LA-GCN	3.23(2.52)	35.68	46.76	58.19	69.48	94.74
	LSTM	2.37(1.6)	**56.36**	**71.6**	80	89.99	95.91
	proposed	**2.13(1.18)**	55.23	70.12	**86.20**	**93.50**	**99.40**
In-house	3D-Unet	3.57(2.27)	23.84	35.67	48.23	68.12	98.58
	SCN	3.48 (2.31)	25.68	37.67	51.74	69.57	97.58
	DRM	3.44 (**2.21**)	26.68	38.67	52.74	69.57	**98.58**
	proposed	**3.29**(2.26)	**31.2**	**41.67**	**54**	**72.2**	98.2

Results on the PDDCA Dataset:
We compare the performance of our method with the State-Of-The-Art (SOTA) methods described in (Chen et al., 2022). Similar to this article's experiment settings, three-fold cross-validation is performed on this dataset. Due to the unavailability of the data division in the article, there may be discrepancy in the data set division for three-fold cross-validation. As shown in Table 1, The results demonstrate that two-stage methods, which effectively utilize high-resolution features, have higher annotating accuracy compared to one-stage methods. Compared with other two-stage methods, Our proposed method enables efficient heatmap regression for different types of landmarks. With the prior information, our model can accurately locate the ambiguous landmarks (Occipital bone). Our method significantly outperforms the state-of-the-art methods. Especially in the precision of 3 mm $\sim$ 8 mm, the improvement of SDRs is increased by 4% $\sim$ 5%. High-precision localization of these ambiguous landmarks makes our method achieve lower MRE(2.13) and SD(1.18) .

Results on the In-house prostate Dataset:
LA-GCN adopts Mask-RCNN in the fine stage for landmark detection, which uses the segmentation labels of bones. For the SA-LSTM method, we find that the fine-stage cannot continuously adjust the landmark offsets on our dataset. Therefore, we only present the results of the methods we reproduced (UNet, SCN, and DRM) in Table 1. Due to the small size of prostate data, we can use high-resolution data as input for the one-stage method, which also achieves good accuracy. Our model achieves higher accuracy in MRE and SD evaluation indicators than in these methods. At the same time, our method can precisely query key slices to help doctors better judge the prostate volume.

3.4. Ablation study

The main contribution of our method is the correlated landmark regression module, for which we conducted ablation experiments on various components. The results of these experiments on the PDDCA dataset correlated landmarks are shown in Table 2. The use of the information fusion module helps guide the axial localization of ambiguous landmarks, and the results show that it significantly impacts accuracy. The axial attention module enhances the fusion of axial features, and the keyframe query layer helps guide the attention module to filter these features. These modules enable our model to achieve an MSE of 1.82 mm on correlated landmarks.

Table 2: Ablation study results for related landmarks.

Component			**MRE(SD)**	**SDR**			
fusion	**axis attention**	**classification**		**2mm**	**3mm**	**4mm**	**8mm**
			3.16(2.49)	31.88	65.4	80	93.7
✓			2.06(**0.916**)	74.7	86.48	95.5	100
✓	✓		2.02(1.26)	83.66	87.3	94.58	99.1
✓	✓	✓	**1.82**(0.932)	**81.82**	**89.12**	**95.5**	**100**

4. Conclusion

This paper proposes a coarse-to-fine architecture for digitizing independent and related anatomical landmarks on 3D images. For independent landmarks, we use multiple Unet models for landmarks regression, ensuring that each model only focuses on the features within the patches. We propose a feature fusion module and a key slice classification module for the landmarks on the same slice. It realizes the position of the key slice from multiple patches and uses the fused information to assist in annotating each patch. Comprehensive experiments conducted on public datasets and one in-house dataset demonstrates that our method achieves better results than state-of-the-art methods. In future work, we will further explore how to add shape constraints of organs or bones and the constraints of human annotation to patch-based heatmap regression methods.

Acknowledgments

This work was supported by the Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innoation. ("Climbing Program" Special Funds). pdjh2022c0021

References

Runnan Chen, Yuexin Ma, Nenglun Chen, Lingjie Liu, Zhiming Cui, Yanhong Lin, and Wenping Wang. Structure-aware long short-term memory network for 3d cephalometric landmark detection. *IEEE Transactions on Medical Imaging*, 2022.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

María Escobar, Cristina González, Felipe Torres, Laura Daza, Gustavo Triana, and Pablo Arbeláez. Hand pose estimation for pediatric bone age assessment. In *International conference on medical image computing and computer-assisted intervention*, pages 531–539. Springer, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, 2021.

Minkyung Lee, Minyoung Chung, and Yeong-Gil Shin. Cephalometric landmark detection via global and local encoders and patch-wise attentions. *Neurocomputing*, 470:182–189, 2022.

Wei Liu, Yu Wang, Tao Jiang, Ying Chi, Lei Zhang, and Xian-Sheng Hua. Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 670–679. Springer, 2020.

Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *European Conference on Computer Vision*, pages 72–88. Springer, 2022.

Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical image analysis*, 54:207–219, 2019.

Patrik F Raudaschl, Paolo Zaffino, Gregory C Sharp, Maria Francesca Spadea, Antong Chen, Benoit M Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, et al. Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. *Medical physics*, 44(5):2020–2036, 2017.

Zimeng Tan, Jianjiang Feng, Wangsheng Lu, Yin Yin, Guangming Yang, and Jie Zhou. Cerebrovascular landmark detection under anatomical variations. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

Rong Tao and Guoyan Zheng. Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine ct with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 93–103. Springer, 2021.

Baris Turkbey, Andrew B Rosenkrantz, Masoom A Haider, Anwar R Padhani, Geert Villeirs, Katarzyna J Macura, Clare M Tempany, Peter L Choyke, Francois Cornud, Daniel J Margolis, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76 (3):340–351, 2019.

Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaž Vrtovec, Olaf Ronneberger, et al. A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis*, 31:63–76, 2016.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Dong Yang, Shaoting Zhang, Zhennan Yan, Chaowei Tan, Kang Li, and Dimitris Metaxas. Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 17–21. IEEE, 2015.

Minmin Zeng, Zhenlei Yan, Shuai Liu, Yanheng Zhou, and Lixin Qiu. Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis*, 68:101904, 2021.

Dongqing Zhang, Jianing Wang, Jack H Noble, and Benoit M Dawant. Headlocnet: Deep convolutional neural networks for accurate classification and multi-landmark localization of head cts. *Medical image analysis*, 61:101659, 2020a.

Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020b.

Zhusi Zhong, Jie Li, Zhenxi Zhang, Zhicheng Jiao, and Xinbo Gao. An attention-guided deep regression model for landmark detection in cephalograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer, 2019.