# Prior Guided 3D Medical Image Landmark Localization

**Author name(s) withheld** EMAIL(S) WITHHELD
*Address withheld*

**Editors:** Under Review for MIDL 2023

## Abstract

Accurate detection of 3D landmarks is critical for evaluating and characterizing anatomical features and performing preoperative diagnostic screening. However, detecting 3D landmarks can be challenging due to the local structural homogeneity of medical images. To address this issue, physicians often annotate multiple landmarks in a single slice, particularly when estimating 3D distance or volume. In this study, we present a prior guided coarse-to-fine framework for efficient and accurate 3D medical landmark detection; we make use of the prior information that in specific settings physicians annotate multiple landmarks on a same slice. The coarse stage uses coordinate regression on downsampled 3D images to maintain the structural relationships across different landmarks. The fine stage categorizes landmarks as correlated and independent landmarks based on their annotation prior. For independent landmarks, we train multiple models to capture local features and ensure reliable local predictions. For correlated landmarks, we mimic the manual annotation process and propose a correlated landmark detection model that merges information from various patches to query key slices and identify correlated landmarks. Our method is extensively evaluated on two datasets, exhibiting superior performance with an average detection error of respective 3.29 mm and 2.13 mm.

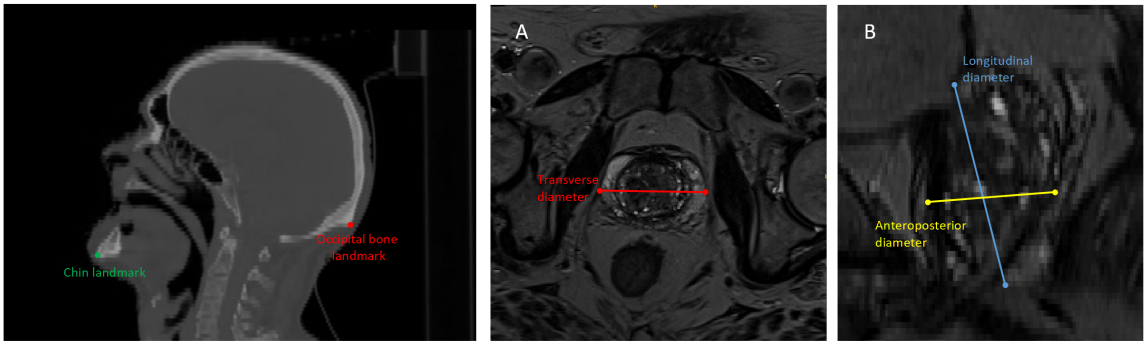
**Keywords:** 3D landmark detection, coarse-to-fine,feature fusion, prior.

## 1. Introduction

Anatomical landmark localization is crucial for various types of medical applications, including orthodontic and maxillofacial surgery planning (Wang et al., 2016), organ volume estimation (Turkbey et al., 2019), and skeletal development assessment (Escobar et al., 2019). As manual annotation of anatomical landmarks is cumbersome and labor-intensive, accurate and automatic landmark detection is highly desired. Unfortunately, local structural similarity often leads to ambiguity in landmark localization, especially in 3D medical images since adjacent slices have very similar structure and intensity profiles. In such context, physicians often annotate landmarks with specific landmarks serving as references (Wang et al., 2016) or annotate multiple landmarks on the key slices (Turkbey et al., 2019; Tan et al., 2022). As shown in Figure 1.(a), the occipital bone landmark is localized with the reference of the chin landmark. In Figure 1.(b), multiple landmarks are annotated on the same slice that has largest organ volume. In these cases, solely focusing on the local features of a single landmark may misidentify the key slice and induce a large localization deviation. Accurately identifying the key slices can improve the accuracy of correlated landmark localization, specifically in 3D cases.

With the emergence of 3D convolutional neural networks (3D CNNs) (Çiçek et al., 2016), 3D landmark detection methods have advanced from key slice classification (Yang et al., 2015) to coordinate and heatmap regression with 3D CNNs. However, processing

© 2023 Anonymous.




(a) Cephalograms Landmarks (b) Prostate Landmarks

Figure 1: Representative examples of correlated landmarks.

3D volume data is computationally expensive due to the increased network parameters. Typically, 3D medical images are downsampled to relatively low resolution for regression based landmark detection. Zhang et al. establish a multitask model for image volume classification and landmark localization (Zhang et al., 2020a). Liu et al. simultaneously predict landmarks' positions and the distances between different landmarks for THA pre-operative planning (Liu et al., 2020). Payer et al. introduce a method that combines local appearance with spatial configuration for landmark regression and achieves excellent results on multiple datasets (Payer et al., 2019). These approaches all need downsample the entire image to a low resolution, which weakens local texture features and causes inevitable errors. Numerous studies demonstrate that incorporating high-resolution information is beneficial for landmark localization (Zhang et al., 2020b).

Coarse-to-fine strategies may somehow relieve the accuracy degradation issue caused by downsampling. Tao and Zheng employ a transformer-based structure for coarse landmark detection, and then use a Unet model for heatmap regression based fine landmark detection (Tao and Zheng, 2021). Zeng et al. propose a three-stage coordinate regression framework that progressively utilizes global and local attention (Zeng et al., 2021). Lee et al. integrate global and local feature extractions as well as specific shape constraints into a single-passing CNN for landmark localization (Lee et al., 2022). Chen et al. propose a Structure-Aware Long Short Term Memory (SA-LSTM) framework that fuses the multi-resolution information of input patches to continuously adjust the landmark offsets (Chen et al., 2022). Most of these methods employ coordinate regression in the fine stage, which has an innate potential to incorporate implicit structural knowledge (Jin et al., 2021). Still, the global average pooling operation employed in these methods may undermine spatial structure and has a detrimental effect on precise landmark localization (Mao et al., 2022).

Besides the two categories of methods discussed above, Deep reinforcement learning is also adapted to locate different structures based on the patch information. Alansary et al. use an artifificial RL agent learns to identify the optimal path to the landmark with multi-scale search strategies (Alansary et al., 2019). Bekkoucha et al. propose a multi-agent RL that merges the concepts of the graphical lasso and Morris sensitivity analysis to quantitatively estimate the contribution of individual landmark and landmark subgroup to

the other landmark locations (Bekkouch et al., 2022). Again, search-based method AAM (Gao et al., 2010) can also model the prior knowledge of landmark's structure and the patch texture around landmark. Compared with heatmap-based methods, search-based methods tend to be time-consuming at the inference phase.

Although most of the existing methods implicitly integrate structural knowledge(Gao et al., 2010; Jin et al., 2021; Zeng et al., 2021), they are kind of weak that are usually inferred from the landmarks' distributions. Moreover, these methods tend to ignore the physician's prior knowledge when manually annotating specific landmarks (e.g. correlated landmarks). Physicians' prior knowledge indicates the relevance of landmarks' annotating process and shall be more useful. Incorporating physicians' prior knowledge into the automated landmark localization process is essential but has yet to be explored.

We here propose a prior guided coarse-to-fine landmark localization framework to effectively combine the advantages of heatmap regression and coordinate regression and integrate the prior knowledge from physicians' annotation. We adopt coordinate regression in the coarse stage to preserve structural constraints across landmarks (Jin et al., 2021). In the fine stage, we categorize the landmarks into correlated and independent landmarks based on how physicians annotate them. For independent landmarks, we train multiple Unet models to fully focus on the local features of each landmark. For correlated landmarks located at the same slices, we propose an axial attention fusion module and a key slice detection module for correlated landmark regression. These two modules filter correlated patches' features to attain a high key slice detection accuracy and assist in detecting ambiguous landmarks.

Our contributions can be summarized as follows: (1) Combining the advantages of the two types of regression methods as well as physician's prior knowledge, we propose a novel coarse-to-fine landmark localization framework. (2) To the best of our knowledge, we are the first here to design a correlated landmark localization module as guided by prior knowledge. (3) We verify our proposed method on two datasets, and the experimental results reveal the superiority of the proposed approach. The source code will be made publicly accessible upon acceptance.

## 2. Methods

As shown in Figure 2, we employ a coarse-to-fine framework for landmark detection. We train a ResNet34 (He et al., 2016) for coordinate regression with a downsampled entire image being the input in the coarse stage. To mitigate the accuracy loss induced by downsampling in the fine stage, we train patch-based Unet models for heatmap regression towards each independent landmark with a cropped region of interest (ROI) from the original high-resolution image. For correlated ambiguous landmarks, we adopt an axis attention module to perform dynamically weighted refining and a key slice detection module to make use of the physicians' annotation prior.

### 2.1. Coarse stage

The coarse stage takes the entire downsampled image as input and incorporates structural knowledge. we employ a 3D ResNet as backbone for coordinate regression. Given an

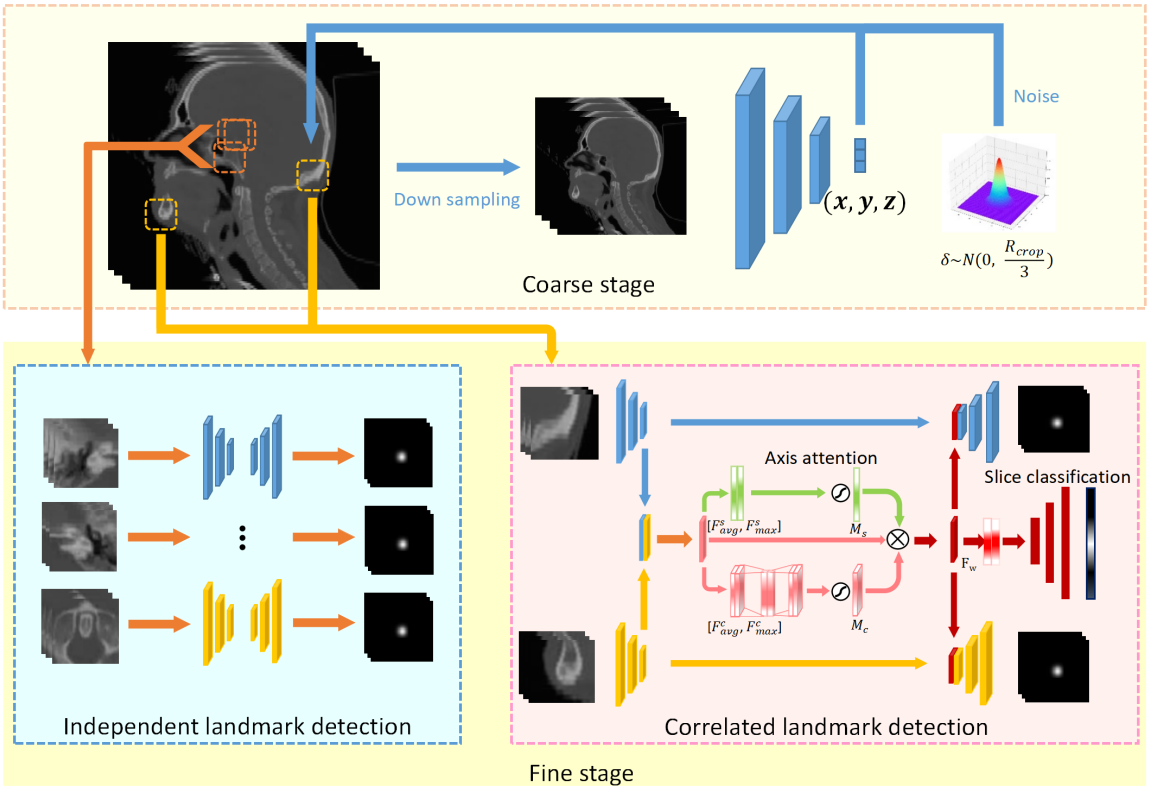



Figure 2: The proposed prior guided coarse-to-fine landmark localization pipeline.

entire image containing $n$ landmarks with ground truth $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$, we employ the following regression loss:

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

where $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ donates the coarse estimation of landmarks' positions. To boost the accuracy, we crop the original image according to the predicted landmarks in the coarse stage and obtain ROI images of the original high resolution for fine localization.

### 2.2. Fine stage

The fine stage focuses on extracting local features around multiple individual landmarks using patch-based Unets (Çiçek et al., 2016). The locally functioning Unets have a better distinguishing ability for high-resolution local features. Considering the coarse stage's predictions $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ will converge to the ground truth, cropping patches centered on $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ will result in the lack of diversity of training samples. Inspired by (Chen et al., 2022), we add Gaussian noise $\delta \sim \mathrm{N}(0\ ,\ \sigma)$ to the cropping centers. We set $\sigma$ to $l/6$ to ensure the patches contain the to-be-predicted landmarks, and $l$ is the patch's edge length.

To effectively exploit physicians' prior knowledge, we categorize the landmarks of interest into two types according to the physicians' annotation practice. For independent landmarks,

physicians separately annotate them based on local texture features. As shown at the bottom left of Figure 2, Each U-Net will extract features and predict heatmaps for the corresponding patch. The loss function $L_{patch}$ simply the Mean Squared Error (MSE) which is consistent with previous works (Chen et al., 2022) on heatmap regression. For the correlated landmarks, physicians typically identify particular slices containing the complete organ's characteristics and annotate correlated landmarks on those slices. In light of this, we reset the predicted centers of correlated ROI images in the same slice. This ensures that the position of key slice is same for the correlated patches. We design an Axis attention module and Key slice detection for key slice querying and landmark detection.

**Axis attention module** Since we utilize different encoders to separately deal with the correlated patches, we aim at deeply fusing axis features for key slice detection. Inspired by the CBAM module (Woo et al., 2018), we propose axis channel and spatial attention to perform dynamic weighted refining. Axis is defined as the direction perpendicular to the key slice, and we assume the axis follows the z-axis.

We sequentially concatenate correlated patch features and feed it into a $3 \times 3$ convolution layer, obtaining a fused feature map $F = [f_1, f_2, ..., f_C] \in R^{C \times H \times W \times D}$. To encourage channel attention blocks to capture spatial interactions with axis information, we factorize the axis avgpooling as formulated in Eq. 2, which can preserve axis positional information and generate axis avg channel context descriptor $F^c_{avg}$. The formulation of axis maxpooling is similar, which only preserve the maximum value to generate descriptor $F^c_{max}$.

$$F^c_{avg} = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} F(i,j), F^c_{avg} \in R^{C \times D}. \tag{2}$$

(H,W) are the squeezed dimensions to build axis channel descriptor. Both descriptors are then forwarded to a shared 1-dimentional convolution layer and merged by element-wise summation. After sigmoid activation $\sigma$ , we obtain axis channel attention maps $M_c \in R^{C \times D}$

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

To exploit the axis spatial attention, we squeeze the channel infomation of [ $F^c_{avg}, F^c_{max}$], yielding a pair of direction-aware feature maps: $[F^s_{avg}, F^s_{max}] \in R^{2 \times D}$. feeding it into a convolution layer and sigmoid activation $\sigma$, we obtain an axis spatial attention map $M_s \in R^D$. After element-wise multiplication of F, $M_s$ and $M_c$, we get the dynamically weighted feature maps $F_w = F \times M_s \times M_c$.

$$M_s = \sigma(cov[F^s_{max}, F^s_{avg}]). \tag{4}$$

**Heatmap regression and Key slice detection** Then, we concatenate the dynamically weighted features and the patch features to guide patch-based heatmap regression, with a goal of improving the localization accuracy in the axial direction. We introduce a slice detection branch to determine the key slice and constrain the attention map. Same as axis channel attention, we use pooling operations to preserve precise positional features along the axis direction and use 1D convolutions to decode the high-level features. Unlike the traditional classification, we use a Gaussian heatmap to represent the probability of

where the key slice is located. The loss function for slice detection $L_{slice}$ is formulated with MSE and the overall loss function for the fine stage is:

$$L_{fine} = \lambda_0 L_{patch} + \lambda_1 L_{slice}. \tag{5}$$

Where $\lambda_0$ and $\lambda_1$ are the weights of 3D heatmap regression and slice classification. In the inference phase, the DARK (Zhang et al., 2020b) method is adopted to decode the two different heatmaps, obtaining the locations of landmarks and the key slices.

## 3. Experiments and Results

### 3.1. Datasets

**In-house prostate dataset** This dataset consists of 857 T2-weighted MRI scans with a slice space of 0.67 mm, acquired from the Health Science Center, Shenzhen Second People's Hospital. The scans are saved in NII format and each contains six prostate landmarks annotated by an experienced clinician for prostate volume estimation (Turkbey et al., 2019): the endpoints of the maximum longitudinal diameter, the maximum transverse diameter as well as the maximum diameter. The dataset were randomly split into 514 scans for training, 171 scans for validation and 172 scans for testing.

**PDDCA dataset** The PDDCA dataset consists of 48 CT images selected from the Radiation Therapy Oncology Group 0522 study. The images are in NRRD format with various resolutions and spacings. The dataset were used in the MICCAI Head-Neck Challenge 2015 (Raudaschl et al., 2017), and 33 images were labeled with five bony landmarks for gross alignment evaluation. The landmarks include: the right condyloid process landmark (mand_r), the left condyloid process landmark (mand_l), the chin landmark(chin), odontoid process landmark(odont_p), and the occipital bone landmark(occ). Following a previous work (Chen et al., 2022), we resize all images to 518×518×384 with an isotropic voxel spacing of 1 mm. The dataset are equally divided for three-fold cross-validation.

### 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 coarse and fine stages are respectively set to be to $1e^{-3}$ and $1e^{-6}$ . The training epochs are 50 on the prostate dataset and 300 on the PDDCA dataset. We use the SGD optimizer with a weight decay of $1e^{-8}$. Random translation (10% of the input size), random scaling (up to 10% difference), and random intensity shift (up to 0.1) are applied for data augmentation. The input size of the coarse stage is set to be 64×64×64 for the prostate dataset and 96×96×72 for PDDCA. To balance the trade-off between computational memory consumption and contextual feature extraction, we set the cropping size to be 48×48×48 in the fine stage and the cropping resolutions to be 1mm and 2mm for the two datasets.

### 3.3. Comparison to state-of-the-art

To evaluate the performance of our proposed method, we conduct comprehensive experiments on the two datasets. We compare our method to three baseline deep learning-based

medical landmark localization methods: 3D Unet (Çiçek et al., 2016), SCN (Payer et al., 2019), and DRM (Zhong et al., 2019). All methods are assessed with the evaluation metrics proposed by (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). For one-stage methods, all 3D images are downsampled to the maximum size allowed by memory as the input. Two-stage methods use the same hyperparameters as our proposed method, such as downsampling volume, patch size and spacing.

Table 1: Performance comparisons between our proposed methods. † and ‡ respectively indicates one-stage methods and coarse-to-fine methods.

| Dataset | method | MRE (SD) | SDR (%) | | | | |
|---|---|---|---|---|---|---|---|
| | | | 2mm | 2.5mm | 3mm | 4mm | 8mm |
| PDDCA | 3D-Unet † | 7.69 (5.24) | 2.03 | 3.25 | 5.05 | 16.28 | 67.90 |
| | SCN † | 7.44 (4.26) | 2.65 | 6.74 | 10.98 | 21.36 | 69.30 |
| | 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 |
| | SA-LSTM ‡ | 2.37 (1.60) | **56.36** | **71.60** | 80.00 | 89.99 | 95.91 |
| | **Proposed ‡** | **2.13 (1.18)** | 55.23 | 70.12 | **86.20** | **93.50** | **99.40** |
| Prostate | 3D-Unet † | 3.57 (2.27) | 23.84 | 36.82 | 48.23 | 68.12 | 96.32 |
| | SCN † | 3.48 (2.31) | 25.68 | 39.34 | 51.74 | 69.57 | 95.73 |
| | DRM ‡ | 3.44 (**2.21**) | 26.74 | 38.24 | 52.13 | 70.54 | **97.58** |
| | **Proposed ‡** | **3.29** (2.26) | **31.17** | **41.67** | **54.13** | **73.22** | 95.62 |

**Results on the PDDCA dataset** We show the results directly copied from previous literature(Chen et al., 2022) as well as our method in Table 1. The results demonstrate that two-stages methods generally deliver higher localization accuracies since they can effectively utilize high-resolution features. Among all methods, our proposed method performs the best in terms of MRE (2.13 mm) and SD (1.18 mm). In terms of SDR, our method achieves much higher successful rates than the second best-performing method SA-LSTM for the target radii of 3mm, 4mm, and 8mm. For the other two target radii, namely 2mm and 2.5mm, the proposed method's performance is on par with SA-LSTM. To further analyze the superiority of our method, we conducted experiments on three baseline methods with optimal hyperparameters. Detailed analysis of different landmarks' localization accuracy is presented in Appendix A.1. From the obtained results, we conjecture a potential reason is that the proposed method can more accurately locate ambiguous landmarks with the specifically-designed axis attention fusion module.

**Results on the in-house prostate dataset** LA-GCN adopts Mask-RCNN in the fine stage, which uses the segmentation labels of bones. As such, LA-GCN does not work for our in-house prostate dataset. Due to the complexity of SA-LSTM method and the instability of landmarks' local features, SA-LSTM fails to continuously adjust the landmark offsets on prostate dataset. Therefore, its only present the results of the comparison results of the three baseline, namely 3D Unet, SCN, and DRM in Table 1. Due to the small volume of prostate data, one-stage methods can use high-resolution data as the input and thus

achieve decent accuracy. With that being said, our proposed method still obtains the best localization performance, in terms of almost all evaluation metrics. This may be because our method has a stronger local feature extraction capability. More detailed experimental results are presented in Appendix A.2.

### 3.4. Ablation study

The main contribution of our method is the correlated landmark regression module. We conducted ablation experiments on the three components to verify their importance. The fusion module refers to the convolutional layer and information interaction between correlated patches. Without the fusion module, all correlated landmarks will be treated as independent landmarks. The axis attention module is used for dynamically weighting. The slice detection branch is designed for key slice detection. We run five times per experiment with different seed and report the average of the highest performance. The ablation analysis results on the PDDCA dataset(22 samples for training and 11 samples for validation) are listed in Table 2. The feature fusion module guides the axial localization of ambiguous landmarks, which significantly enhances the performance. The implementation of the axial attention module and the slice detection branch aggregate axis features to further guide the localization of ambiguous landmarks. These modules enable our model to achieve a lower MSE (1.82 mm) on correlated landmarks. Collectively, our proposed method can effectively use prior information from physicians' annotation practice and constrain the location of correlated landmarks. With such constraint, our method boosts the localization accuracy, especially for ambiguous landmarks ( see Appendix A.3).

Table 2: Ablation study results for PDDCA dataset by evaluation metrics of MRE(SD).

| Component | | | Correlated | | Independent | | |
|---|---|---|---|---|---|---|---|
| Fusion | Axis attention | Slice detection | Chin | Occ | Mand_l | Mand_r | Odont_p |
| | | | 2.00 (0.65) | 4.32 (3.10) | 2.05 (0.53) | 2.18 (0.74) | 1.68 (1.14) |
| ✓ | | | 1.87 (**0.57**) | 2.18 (1.18) | | | |
| ✓ | ✓ | | **1.76** (0.66) | 2.28 (1.56) | | | |
| ✓ | ✓ | ✓ | 1.85 (0.62) | **1.79** (**1.15**) | | | |

## 4. Conclusion

This paper proposes a coarse-to-fine framework for localizing independent and correlated anatomical landmarks from 3D medical images. In the fine stage, for independent landmarks, we employ multiple Unet models for landmarks regression, ensuring that each model solely focuses on a patch centering the specific single landmark of interest. For the correlated landmarks, we propose a feature fusion module and a key slice detection module. It successfully identifies the position of the key slice from multiple patches and uses the fused features to assist in landmark localization. Our method outperforms state-of-the-art methods, according to extensive experiments on the publicly-accessible PDDCA dataset and our in-house prostate dataset. We shall further incorporate more types of prior knowledge, such as the shape prior of the anatomical region of interest and other prior knowledge utilized in the manual annotation process into patch-based heatmap regression methods in the future work.

## Acknowledgments

Acknowledgments withheld.

## References

Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Konstantinos Kamnitsas, Athanasios Vlontzos, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. Evaluating reinforcement learning agents for anatomical landmark detection. *Medical Image Analysis*, 53:156–164, 2019.

Imad Eddine Ibrahim Bekkouch, Bulat Maksudov, Semen Kiselev, Tamerlan Mustafaev, Tomaž Vrtovec, and Bulat Ibragimov. Multi-landmark environment analysis with reinforcement learning for pelvic abnormality detection and quantification. *Medical Image Analysis*, 78:102417, 2022.

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.

Xinbo Gao, Ya Su, Xuelong Li, and Dacheng Tao. A review of active appearance models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(2):145–158, 2010.

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.

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.

## Appendix A. More Experimental Results

### A.1 More results on PDDCA dataset

To convincingly demonstrate the advantages of our approach, we re-implement four deep learning-based methods for extensive comparison, ie 3D Unet, SCN, DRM and SA-LSTM. We set the optimal hyperparameters to maximize their performance, such as the cropping size, kernel size of Gaussian heatmap and input resolution. The PDDCA dataset is equally distributed into three folds to perform three-fold cross-validation, and the average individual landmark localization accuracy of three folds is reported in Table 3. We also report the SDR of our method with SA-LSTM at different radii in Figure 3

Table 3: Individual landmark localization performance comparisons by metric MSE.

| Method | Overall | Chin | Mand_l | Mand_r | Odont_proc | Occ_bone |
|---|---|---|---|---|---|---|
| 3D Unet | 4.00 (2.47) | 2.67 | 3.80 | 3.56 | 3.51 | 6.47 |
| SCN | 3.71 (2.02) | 2.43 | 3.67 | 3.43 | 3.75 | 5.26 |
| DRM | 2.72 (2.85) | **1.60** | 2.00 | **1.97** | 1.94 | 6.09 |
| SA-LSTM | 2.44 (1.99) | 1.95 | **1.88** | 2.15 | **1.79** | 4.45 |
| **Proposed** | **2.13 (1.18)** | 1.75 | 1.99 | 2.08 | 2.01 | **2.81** |

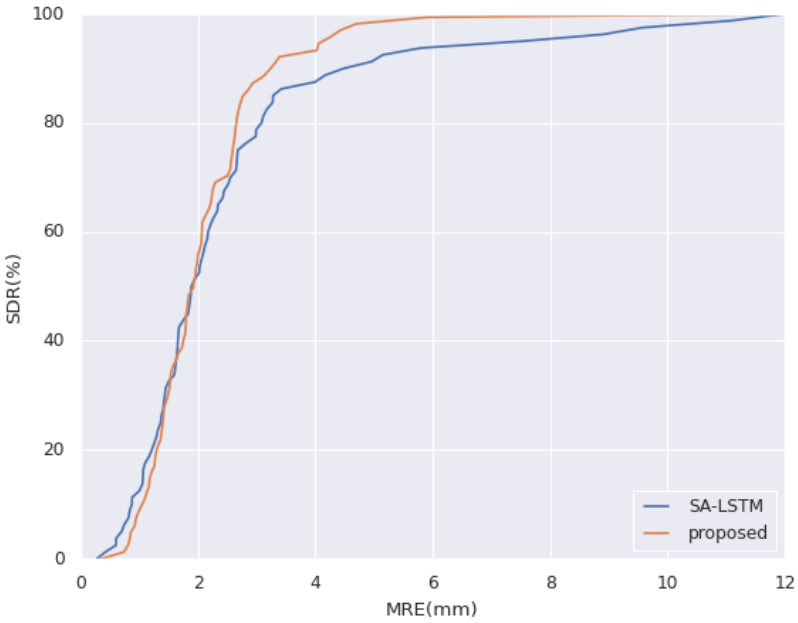



Figure 3: SDR under continuous acceptance radius .

### A.2 More results on in-house prostate dataset

Table 4 shows the MSE for individual landmarks predicted by three baseline methods. L_1 ∼ L_6 represent the endpoints of the maximum longitudinal diameter, the maximum transverse diameter as well as the maximum diameter. Results demonstrate that the embedding of physician's prior knowledge raises the localization accuracy of all landmarks.

Table 4: Individual landmark localization performance comparisons by metric MSE.

| Method | Overall | L_1 | L_2 | L_3 | L_4 | L_5 | L_6 |
|---|---|---|---|---|---|---|---|
| 3D Unet | 3.57 (2.27) | 3.58 | 3.61 | 3.23 | 3.95 | **3.32** | 3.74 |
| SCN | 3.48 (2.31) | 3.36 | 3.18 | 3.49 | 3.63 | 3.43 | 3.78 |
| DRM | 3.44 **(2.21)** | 3.33 | 3.17 | 3.32 | 3.59 | 3.45 | 3.80 |
| **Proposed** | **3.29** (2.26) | **3.16** | **3.04** | **3.02** | **3.52** | 3.43 | **3.61** |

### A.3 Further demonstration of ablation study

Physician annotates the occipital bone landmark (Occ_bone) with the reference of the Chin landmark. Ablation study is conducted for these two landmarks. We run five times per experiment and report the average of the highest performance of five times for each experiment. Here we show that three modules effectively improve the localization accuracy of ambiguous landmark and successfully incorporate the physician's a prior knowledge into the model.

Table 5: Evaluation of localization accuracy for correlated landmarks

| Landmark | Fusion | Axis attention | Slice detection | SDR (%) | | | |
|---|---|---|---|---|---|---|---|
| | | | | 2mm | 3mm | 4mm | 8mm |
| Chin | | | | 45.45 | **98.18** | **100.00** | **100.00** |
| | ✓ | | | 56.36 | 98.18 | 100.00 | 100.00 |
| | ✓ | ✓ | | 63.56 | 94.18 | 99.18 | 100.00 |
| | ✓ | ✓ | ✓ | **65.42** | 91.06 | 98.40 | 100.00 |
| Occ_bone | | | | 18.35 | 32.62 | 60.05 | 87.40 |
| | ✓ | | | 49.04 | 74.82 | 91.00 | 100.00 |
| | ✓ | ✓ | | 61.82 | 80.42 | 89.89 | 98.18 |
| | ✓ | ✓ | ✓ | **61.84** | **87.18** | **92.61** | **100.00** |