2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

MULTI-TASK LEARNING NETWORK FOR LANDMARK DETECTION IN ANATOMICAL TREE STRUCTURES

Zimeng Tan[⋆†]*, Jianjiang Feng*[⋆†]*, and Jie Zhou*[⋆†]

[⋆] Department of Automation, Tsinghua University, Beijing, China

[†] Beijing National Research Center for Information Science and Technology, Beijing, China

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9434017

ABSTRACT

Bifurcation landmark detection is an important step in automatic analysis of anatomical tree structures, such as airway and vessel. In this paper, we propose a multi-task learning based deep neural network specifically designed for automatically and accurately localizing bifurcation landmarks . Given an input volume, the network is trained to predict landmark confidence maps, branch segmentation maps, and branch orientation fields simultaneously. In this way, we exploit the spatial relationships among landmarks explicitly. The structural prior is also introduced to the architecture, which guides the network to learn more discriminative features. Experiments on airway tree and aorta tree demonstrate that the proposed method is effective for bifurcation landmark detection, and exploiting segmentation and orientation field regression as auxiliary objectives helps to increase accuracy of landmark detection substantially.

***Index Terms*—** Landmark detection, multi-task learning, anatomical tree structure, segmentation, orientation field

1. INTRODUCTION

Automatic localization of anatomical landmarks is an important enabling technology, providing essential spatial information for centerline extraction, segmentation, registration and many other subsequent medical image analysis tasks [1]. Recently, deep learning based methods have made significant progress [2, 3, 4, 5] benefiting from their superior ability of learning features.

An increasingly popular approach is based on heatmap regression [2, 3]. Given an input volume, the network is trained to predict synthetic heatmap, which denotes the probability of each voxel belonging to the target landmark. In contrast with regressing absolute landmark coordinates directly [4], these voxel-to-voxel heatmap regression methods are intrinsically more suitable for landmark detection, as they focus on each position carefully.

However, in terms of bifurcation landmark detection in anatomical tree structures, which have stronger connected topology and more complex distribution of landmarks, there still remain various challenges. For example, most existing methods only exploit the relationships among landmarks implicitly, which limits their potential for performance improvements. Structural prior and spatial context information are also needed to be taken into account.

Bifurcation landmark detection in anatomical tree structures is similar to human pose estimation [6], which refers to localization of human keypoints in images, since both of them can be regarded as a tree structure. Cao et al. [7] proposed Part Affinity Fields (PAFs) to encode the localization and orientation of limbs, which can be generalized to introduce spatial constrains by learning a set of vector fields among nodes in a tree. This provides a new insight into the problem of bifurcation landmark detection.

Inspired by the recent success of multi-task learning [8, 9, 10], we present a multi-task deep neural network for bifurcation landmark detection in anatomical tree structures. The whole architecture is shown in Fig. 1. Based on the generic pipeline of heatmap regression approach, we propose to design branch segmentation and branch orientation field regression as auxiliary objectives to guide the network to extract more discriminative features. The auxiliary objectives are highly correlated with the main task, and the synergy among them can boost the individual performance for each task.

Given an input volume, the anatomical tree structure is divided into different branch regions according to the distribution of the bifurcation landmarks. The 3D orientation field is defined on each branch depending on the connection direction of the two adjacent landmarks. The network is trained to predict landmark heatmap, branch segmentation and orientation fields simultaneously. In this way, we introduce the global structural prior by segmenting the whole anatomical structure and enhance the context information by considering the predefined branches as different classes. Furthermore, the spatial relationships among landmarks are also incorporated explicitly by predicting orientation fields.

Experiments on two types of anatomical tree structure (airway and aorta) in challenging cases (CT volumes of patients with COVID-19 [11] and CTA volumes of patients with aortic dissection [12], respectively) show that exploiting segmentation and orientation field regression as auxiliary tasks

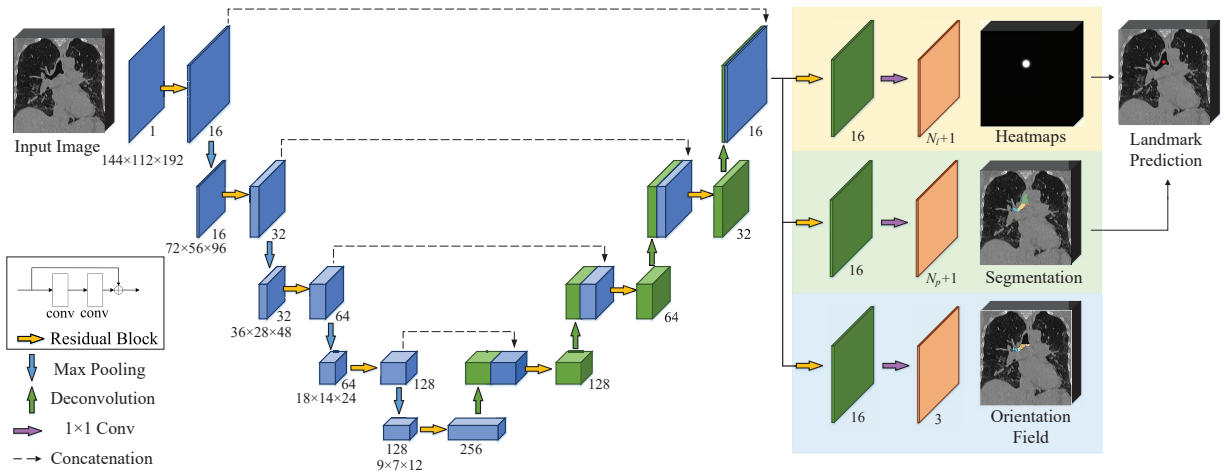


Fig. 1. Architecture of the proposed multi-task network, where N_l and N_p refer to the number of anatomical landmarks and branches, respectively.

can substantially boost accuracy of the landmark localization, especially for difficult situation (i.e., aorta dissection). To our knowledge, this is the first multi-task learning based method specially designed for detecting bifurcation landmarks of anatomical tree structures.

2. METHODS

The backbone of the proposed multi-task model (shown in Fig. 1), similar to U-Net [13], follows encoder-decoder architecture. We replace plain convolution unit with residual block [14] in order to avoid gradient vanishing, which contains two convolution operators and a short-cut connection between input and output. Skip connections between parallel layers are used to incorporate both fine-grain and global context information. Then, the network is split into three branches: the top branch, shown in yellow, regresses the landmark confidence maps; the middle branch, shown in green, predicts the branch segmentation maps; and the bottom branch, shown in blue, predicts the orientation fields.

2.1. Landmark Detection Head Network

Inspired by [2], we convert bifurcation landmark detection problem to a heatmap regression task. The discrete coordinates of a landmark are modeled as a channel heatmap with a Gaussian distribution centered at the landmark position. The heatmap value $\boldsymbol{H}_k^*(\boldsymbol{x})$ of voxel $\boldsymbol{x}$ ranges in [0,1], which can be regarded as the probability to be the kth landmark. The distribution is determined according to the distance from $\boldsymbol{x}$ to the landmark position $\boldsymbol{x}_k$, and the standard deviation δ controls the size. The probability $\boldsymbol{H}_k^*(\boldsymbol{x})$ is set to 1 when $\boldsymbol{x}$ is located at $\boldsymbol{x}_k$, and the father away from $\boldsymbol{x}_k$, the lower the value.

Following the principle of classification, we estimate a shared background to ensure the sum of all classes' probabilities to be 1 for each voxel. So that, for N_l target landmarks, the entire heatmap output of the landmark detection head network contains $N_l + 1$ channels, which can be described as:

$$\boldsymbol{H}_k^*(\boldsymbol{x}) = \begin{cases} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{x}_k)^2], & k = 1, 2, ..., N_l, \\ 1 - \sum_{i=1}^{N_l} \boldsymbol{H}_i^*(\boldsymbol{x}), & k = N_l + 1. \end{cases} \tag{1}$$

Considering the imbalance between background and proximal landmarks, we apply a weighted L_2 loss function L_H between the predicted heatmaps $\boldsymbol{H}_k$ and the groundtruth heatmaps $\boldsymbol{H}_k^*$. The weights are chosen to be the exponential powers of the groundtruth volume.

2.2. Segmentation Head Network

In the segmentation head network, the segmentation task of the whole anatomical tree structure is considered as a multi-class semantic segmentation problem, where different branches growing from the trunk are viewed as different classes. In preparing groundtruth of segmentation, the trunk and branches are separated according to corresponding bifurcation landmarks, namely, the landmark is located at the interface of two adjacent substructures. This makes segmentation task highly correlated with the landmark detection objective. In this way, we emphasize the location distribution of different branches and introduce the shape prior to the architecture, which guides the network to capture more discriminative features.

Specifically, the output contains $N_p + 1$ channels, where the first N_p channels correspond to the N_p foreground classes (containing the trunk and branches) and the last one belongs to the background. We trains our segmentation head network by employing weighted Dice loss function [15] to tackle with the problem of class imbalance. During inference, for each voxel, the class with the highest probability is chosen as the final segmentation prediction.

2.3. Orientation Field Regression Head Network

Orientation field regression head network is trained to predict a set of vector fields, which represent the relative spatial relationships between neighboring landmarks. Taking segmentation and orientation field regression as auxiliary objectives simultaneously, the model can preserve both location and orientation information across the anatomical tree structures.

The orientation fields are constructed based on the branch segmentation. Consider a single branch p_k and two adjacent landmarks $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, for each voxel $\boldsymbol{x}$ on the branch, the groundtruth value of the orientation field $\boldsymbol{O}_k^*$ is a 3D unit vector $\boldsymbol{v}$ from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$; for other voxels, the vector is zero-valued. Given an input volume with N_p branches, $\boldsymbol{O}_k^*$ is defined as:

$$\boldsymbol{O}_k^*(\boldsymbol{x}) = \begin{cases} \boldsymbol{v}, & \text{if } \boldsymbol{x} \text{ on branch } p_k, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where $\boldsymbol{v} = (\boldsymbol{x}_j - \boldsymbol{x}_i)/\left\|\boldsymbol{x}_j - \boldsymbol{x}_i\right\|_2$ is the 3D unit vector in the direction of the branch. It should be noted that orientations in the fields are all from upper bifurcations to the lower bifurcations, namely, extending from parent nodes to the child nodes in the tree structure. Particularly, we do not define orientation field on the trunk, since it may have large curvature change, e.g. the aorta.

Similarly, a weighted L_2 loss function is applied for this head network. The final objective function is formulated by the linear combination of all losses:

$$\mathcal{L} = \mathcal{L}_{heat} + \alpha\mathcal{L}_{seg} + \beta\mathcal{L}_{ori} \tag{3}$$

where the hyper parameters α and β are adjusted to make the different components having the same scale.

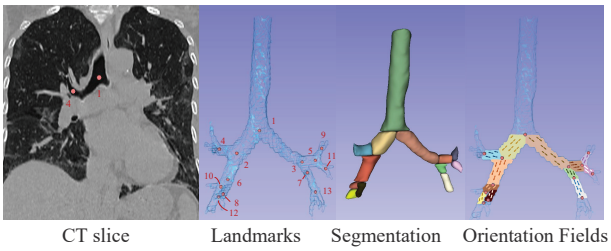
During inference, both landmark detection and segmentation head networks are utilized for final prediction. We argue that the landmark prediction should be located on the anatomical tree structure. So, the predicted heatmaps are filtered by segmentation prediction firstly, discarding the voxels classified as the background. Then, the final predicted position for each landmark is chosen to be the voxel with the maximum probability in the corresponding filtered heatmap.

3. EXPERIMENTS

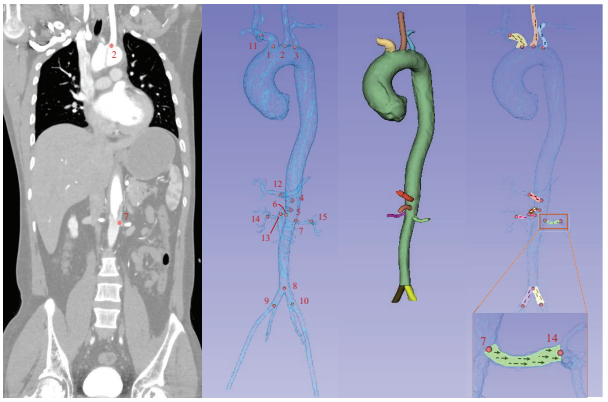
3.1. Data and Implementation Details

We trained and validated our multi-task learning framework on two types of anatomical tree structure: airway in chest CT volumes and aorta in aortic CTA volumes. As shown in Fig. 2, 13 airway landmarks, 15 aortic landmarks and the whole segmentation of these two anatomical tree structures were annotated by an expert. 13 branch regions of airway and 10 branch regions of aorta were spilt accordingly. Then, the orientation fields were computed based on the branches.

Chest CT dataset was randomly divided into 91 training scans from 56 patients and 24 test scans from 15 patients. All



(a) Airway Annotation



(b) Aorta Annotation

Fig. 2. Illustration of annotations of airway and aorta on chest CT and aorta CTA.

patients have COVID-19 and some of them have two scans at different times. Note that scans from the same person was divided into either training set or test set. Input volumes were cropped to 144×112×192 along the lung area obtained by a simple lung segmentation network firstly, which is based on U-Net [13] and pretrained on the same training set. The proposed network was implemented in TensorFlow and trained using Adam with a learning rate of 0.0001 for 200 epochs.

Aorta CTA dataset contains 48 scans with mean shape of 366×366×630. All patients have aortic dissection, which is a life-threatening vascular disease with very high mortality rate [16]. Accurate localization of aortic landmarks is obviously a much more challenging task due to its low resolution, complex organ distribution and interference of true and false lumen. According to the average distribution of landmarks on the training set, we divided the input volume into three sub-regions along longitudinal axis, and a multi-task network was trained for each sub-region separately. Considering limited data at hand, we applied data augmentation by random translation for 4 times. 4-fold cross-validation was performed and the rest of the training process was similar.

3.2. Results and Comparison

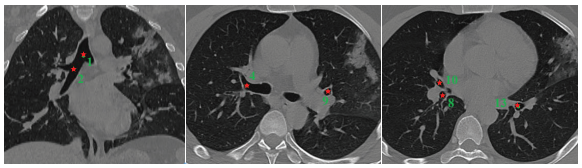
Fig. 3 shows some typical results of the proposed method on two datasets, where the groundtruth and predicted landmarks

Dataset	Landmark Index							
Airway	1	2	3	4	5	6	7	8
	1.76±0.91	2.01±1.30	2.14±0.95	1.29±0.58	1.72±0.63	1.94±1.34	1.86±0.96	2.12±1.24
	9	10	11	12	13	Mean		
	2.10±1.31	2.24±1.09	2.52±1.51	2.48±1.39	3.42±2.31	2.12±1.36		
Aorta	1	2	3	4	5	6	7	8
	2.56±1.38	1.99±2.29	2.01±1.79	1.55±0.91	1.27±0.84	1.47±1.34	2.69±5.68	1.63±1.08
	9	10	11	12	13	14	15	Mean
	2.24±3.67	2.98±4.44	2.84±3.66	2.87±5.25	4.79±5.82	5.78±7.73	7.16±9.15	2.92±5.05

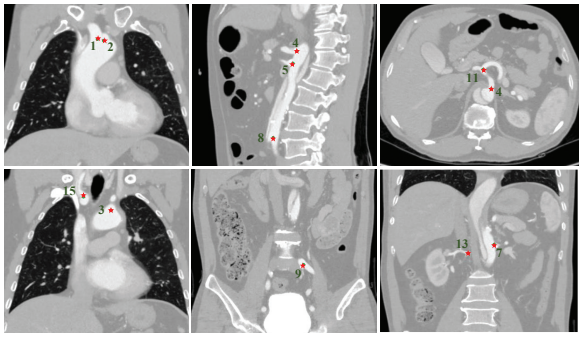
Table 1. Mean with standard deviation of Euclidean distances (in mm) between the groundtruth and predicted landmarks on the airway and aorta datasets.

Model	Aortic Landmark Index															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
Heatmap	3.27	3.45	2.45	1.59	**1.19**	1.73	2.78	1.66	2.69	4.73	4.26	3.26	7.33	9.68	15.61	4.38
Heat+Seg	2.99	3.11	2.21	1.61	1.25	1.80	2.77	1.71	**2.19**	4.45	3.52	2.96	5.26	6.18	12.89	3.66
Full	**2.56**	**1.99**	**2.01**	**1.55**	1.27	**1.47**	**2.69**	**1.63**	2.24	**2.98**	**2.84**	**2.87**	**4.79**	**5.78**	**7.16**	**2.92**

Table 2. Comparison of different network architectures on the aorta dataset. Performance is measured using MRE (in mm).



(a) Airway Landmarks



(b) Aorta Landmarks

Fig. 3. Typical results on two datasets, where the green dots are the groundtruth landmarks and the red stars represent the predicted positions.

are marked in green and red respectively. It is obvious that our framework is effective for bifurcation landmark detection, since the groundtruth and predicted positions are highly consistent even on tiny branches.

Furthermore, we utilized the mean radial error (MRE, in mm) as our metric to evaluate the proposed method quantitatively, which is defined by $MRE = (\sum_{i=1}^{n} R_i)/n$ where n refers to the number of test data and R indicates the Euclidean distance between the grundtruth and predicted landmark position. The obtained results, shown in Table 1, demonstrate that our method achieves excellent performance on both airway and aorta datasets. The landmarks with larger number have larger error, since they are located at tiny branches which may have large variations and brings higher detection difficulty.

To investigate the effectiveness of the proposed multi-task framework, we compared the performance of three different network architectures on the aorta dataset: the U-Net with only landmark detection head network, with the landmark detection and branch segmentation head networks, and with all three head networks. As shown in Table 2, adding each of them successively brings better results. Especially for difficult situations (e.g., landmark No. 13 and 14), the full network architecture shows a noticeable improvement, with the help of spatial constraints. The performance of some landmarks shows a slight decrease (e.g., landmark No. 5 and 9), which may be due to the interference of suboptimal branch segmentation or orientation field prediction.

4. CONCLUSION

In this work, we have developed an end-to-end multi-task network to automatically and accurately detect bifurcation landmarks in anatomical tree structures. We show that incorporating the structural prior by applying branch segmentation as auxiliary objective can guide the network to learn more discriminative features. Explicitly exploiting spatial relationships among landmarks by regressing orientation fields further increases the accuracy of landmark detection. Experiments on airway and aorta datasets demonstrate that our proposed method achieves promising performance. We believe that the accuracy of landmark detection will be further improved with larger training datasets and finer adjustments. Future work could be to extend our framework to other anatomical tree structures.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data in Wuhan Union Hospital. The need for informed consent was waived by the institutional review board of Wuhan Union Hospital.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61976121 and 82071921.

7. REFERENCES

[1] S Kevin Zhou, "Discriminative anatomy detection: Classification vs regression," *Pattern Recognition Letters*, vol. 43, pp. 25–38, 2014.

[2] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler, "Regressing heatmaps for multiple landmark localization using CNNs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 230–238.

[3] 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*. Springer, 2019, pp. 540–548.

[4] Hansang Lee, Minseok Park, and Junmo Kim, "Cephalometric landmark detection in dental X-ray images using convolutional neural networks," in *Medical Imaging 2017: Computer-Aided Diagnosis*. International Society for Optics and Photonics, 2017, vol. 10134, p. 101341W.

[5] Yaozong Gao and Dinggang Shen, "Context-aware anatomical landmark detection: Application to deformable model initialization in prostate CT images," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2014, pp. 165–173.

[6] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[8] Jun Zhang, Mingxia Liu, Li Wang, Si Chen, Peng Yuan, Jianfu Li, Steve Guo-Fang Shen, Zhen Tang, Ken-Chung Chen, James J Xia, et al., "Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization," *Medical Image Analysis*, vol. 60, pp. 101621, 2020.

[9] Jinming Duan, Ghalib Bello, Jo Schlemper, Wenjia Bai, Timothy JW Dawes, Carlo Biffi, Antonio de Marvao, Georgia Doumoud, Declan P O'Regan, and Daniel Rueckert, "Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2151–2164, 2019.

[10] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen, "Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1195–1206, 2018.

[11] Feng Pan, Tianhe Ye, Peng Sun, Shan Gui, Bo Liang, Lingli Li, Dandan Zheng, Jiazheng Wang, Richard L Hesketh, Lian Yang, et al., "Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia," *Radiology*, 2020.

[12] Michelle A McMahon and Christopher A Squirrell, "Multidetector CT of aortic dissection: A pictorial review," *Radiographics*, vol. 30, no. 2, pp. 445–460, 2010.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[14] 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*, 2016, pp. 770–778.

[15] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[16] Antonio Pepe, Jianning Li, Malte Rolf-Pissarczyk, Christina Gsaxner, Xiaojun Chen, Gerhard A Holzapfel, and Jan Egger, "Detection, segmentation, simulation and visualization of aortic dissections: A review," *Medical Image Analysis*, vol. 65, pp. 101773, 2020.