STRUCTURE-ENHANCED ATTENTIVE LEARNING FOR SPINE SEGMENTATION FROM ULTRASOUND VOLUME PROJECTION IMAGES

Rui Zhao [1,†], *Zixun Huang*[1,†] , *Tianshan Liu*[1], *Frank H. F. Leung*[1], *Sai Ho Ling*[3], *De Yang*[2],
Timothy Tin-Yan Lee[2], *Daniel P.K. Lun*[1], *Yong-Ping Zheng*[2], *Kin-Man Lam*[1,⋆]

[1] Department of Electronic and Information Engineering, [2]Department of Biomedical Engineering,
The Hong Kong Polytechnic University, Hong Kong, China
[3]School of Biomedical Engineering, University of Technology Sydney, NSW, Australia

ABSTRACT

Automatic spine segmentation, based on ultrasound volume projection imaging (VPI), is of great value in clinical applications to diagnose scoliosis in teenagers. In this paper, we propose a novel framework to improve the segmentation accuracy on spine images via structure-enhanced attentive learning. Since the spine bones contain strong prior knowledge of their shapes and positions in ultrasound VPI images, we propose to encode this information into the semantic representations in an attentive manner. We first revisit the self-attention mechanism in representation learning, and then present a strategy to introduce the structural knowledge into the key representation in self-attention. By this means, the network explores both the contextual and structural information in the learned features, and consequently improves the segmentation accuracy. We conduct various experiments to demonstrate that our proposed method achieves promising performance on spine image segmentation, which shows great potential in clinical diagnosis.

***Index Terms*—** Spine segmentation, Structure-enhanced attention, Ultrasound volume projection imaging.

1. INTRODUCTION

Ultrasound volume projection imaging (VPI) [1] is a recently proposed technique, which has shown a significant perspective in clinical applications for its harmlessness, efficiency, and flexibility. Automatic spine segmentation from ultrasound VPI images provides the basis for the intelligent diagnosis of scoliosis [2], by serving as a pre-analyzing step for the measurement of spine deformity. Recent studies on deep neural networks (DNNs) produced appealing results in computer vision tasks, including classification, detection, and segmentation. In terms of medical images, great efforts have been made to investigate effective backbone architectures [3, 4], learning algorithms [5, 6], and auxiliary supervisions [7, 8]. However, only limited exploration has been made to utilize the structural information of the different bones to enhance the semantic representations for spine segmentation. Therefore, in this paper, we propose a novel framework, based on structure-enhanced attentive learning, in order to enrich the feature representations with structural knowledge for more effective spine segmentation.

Numerous strategies have been proposed in the literature to improve the segmentation accuracy. Wang et al. [9] presented a knowledge-based method with adaptive thresholds for rib segmentation. Vania et al. [10] utilized the class redundancy as a soft penalty to regularize the segmentation learning from CT images. Quite recently, methods based on domain adaptation and style transfer have been widely studied in medical image segmentation. Liu et al. [8] presented WaveCT to address the appearance-shift problem in ultrasound image segmentation. Huang et al. [7] proposed an efficient regularization-based algorithm to tackle occlusion in VPI images for enhanced spine segmentation. Moreover, recent studies on the attention mechanism have also shown great potential in image segmentation. Lei et al. [11] proposed a deep attention fully convolution network to improve the segmentation on the prostate boundaries. Ding et al. [12] presented a hierarchical attention network for effective medical image segmentation. EM-Attention network [13] aggregated the EM algorithm into the attentive learning framework to enhance the semantic representations, which achieves state-of-the-art performance in natural image segmentation.

However, the aforementioned works take limited consideration of the strong prior knowledge on the structure of the spine bones when learning the semantic representations. To address this issue, we propose the structure-enhanced attention module (SEAM), and embed it into a segmentation network to enrich the learned features. Specifically, we first revisit the self-attention (Non-local) mechanism in representation learning, and then encode the structural information into the key representation in the self-attention module, which produces the structure-enhanced contextual representations. By this means, the resultant model can more effectively lo-

[†] The authors equally contribute to this paper.
[⋆] Corresponding author: Kin-Man Lam (enkmlam@polyu.edu.hk)

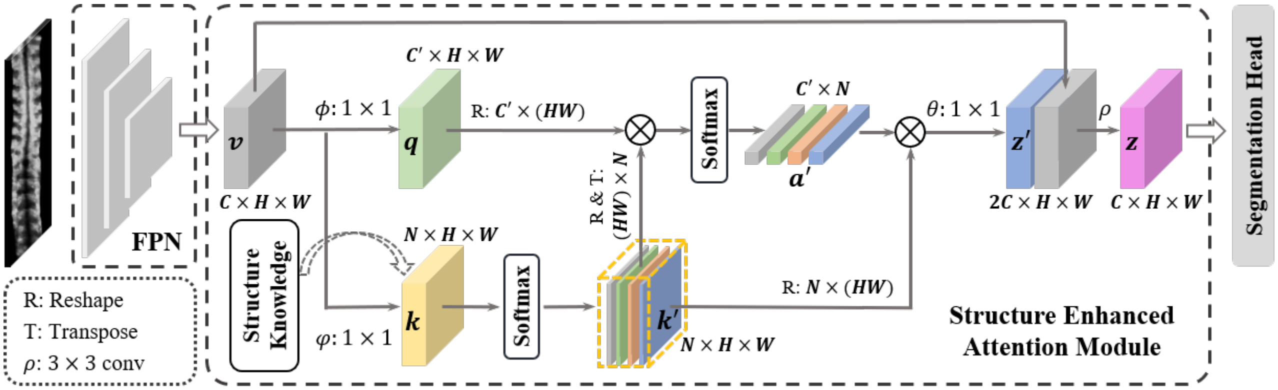

Fig. 1. Overview of the proposed framework with the structure-enhanced attention module for spine segmentation.

calize and recognize the spine bones in ultrasound images.

2. METHODOLOGY

In this section, we introduce the details of the proposed framework, including the structure-enhanced attention module (SEAM) and the learning criteria that we adopt to promote the learning of the structural knowledge.

2.1. Structure-Enhanced Attentive Learning

Self-attention (Non-local) mechanism has been widely used in representation learning to investigate the global dependency on feature maps. The non-local module is formulated as follows:

$$\hat{\boldsymbol{v}}_i = \frac{1}{\mathcal{S}(\boldsymbol{v}_i)} \sum_j f(\boldsymbol{q}_i, \boldsymbol{k}_j) g(\boldsymbol{v}_j), \tag{1}$$

where $\boldsymbol{v}$ denotes a feature map. $\mathcal{S}(\cdot)$ is the normalization factor. $f(\cdot,\cdot)$ explores the correlation between the i-th and the j-th feature vectors in the query $\boldsymbol{q}$ and the key $\boldsymbol{k}$ representations of the signal respectively, and produces a weight matrix implying the self correlation of the input signal; $g(\cdot)$ generates another representation of the input signal $\boldsymbol{v}$. Eq. (1) indicates that the self-attention module is a process of re-estimating the input signal $\boldsymbol{v}$ by a linear combination of all its elements. From this perspective, $g(\boldsymbol{v})$ defines the bases of a space, and $f(\boldsymbol{q}_i, \boldsymbol{k}_j)$ defines the coefficients for the reconstruction of the original signal in that space. Based on this observation, EM-attention [13] employed the expectation-maximization (EM) algorithm to optimize the bases and the coefficients in an unsupervised manner. The EM algorithm serves as a clustering approach, which aims to reconstruct $\boldsymbol{v}$ in a space spanned by the learned attention maps (bases) with less redundancy.

However, different from natural image segmentation, the spine bones contain much stronger prior knowledge on the categories and the structure, because the bone features have a relatively uniform position and shape in different ultrasound images. Therefore, the attention maps in our proposed structure-enhanced attention module can be learned under the supervision of the ground-truth segments. Firstly, there are three different bones in a spine image, i.e. lumbar, thoracic, and rib. Thus, we only need four attention maps, i.e. three foreground attention maps and one background attention map, to re-estimate the input features. Secondly, the attention maps should contain the structural information, i.e. the shape and the position, of the different bones to facilitate the reconstruction of the semantic representations. To this end, we propose the structure-enhanced attention module (SEAM), which is illustrated in Fig. 1.

Given a feature map $\boldsymbol{v} \in \mathbb{R}^{C\times H\times W}$, we generate the query and the key representations, i.e. $\boldsymbol{q} \in \mathbb{R}^{C'\times H\times W}$ and $\boldsymbol{k} \in \mathbb{R}^{N\times H\times W}$ respectively, as follows:

$$\boldsymbol{q} = \phi(\boldsymbol{v});\ \boldsymbol{k} = \varphi(\boldsymbol{v}), \tag{2}$$

where ϕ and φ represent the convolutional mapping for the query and key representations respectively; C, H, and W denote the channel number, height, and width of $\boldsymbol{v}$ respectively; C' is less than C to reduce the computational complexity; N denotes the number of categories, which is 4 in our task. In SEAM, we introduce the structural knowledge into the key representation $\boldsymbol{k}$. Thus, we consider the structural penalty on $\boldsymbol{k}$ (see Sec. 2.2), and produce the structure-enhanced representation $\boldsymbol{k}'$ with the softmax mapping as follows:

$$\boldsymbol{k}' = Softmax(\boldsymbol{k}),\ \text{with}\ \boldsymbol{k}' \in \mathbb{R}^{N\times H\times W}. \tag{3}$$

Then we compute the correlation between the elements in $\boldsymbol{k}'$ and $\boldsymbol{q}$ to generate the attentive weight matrix $\boldsymbol{a}'$ as follows:

$$\boldsymbol{a}' = Softmax(\boldsymbol{q} \times \boldsymbol{k}'^{T}),\ \text{with}\ \boldsymbol{a}' \in \mathbb{R}^{C'\times N}, \tag{4}$$

where the softmax layer functions as the normalization in Eq (1). We reconstruct the signal by a combination of the elements in $\boldsymbol{k}'$ with the weight matrix $\boldsymbol{a}'$, as follows:

$$\boldsymbol{z}' = \theta(\boldsymbol{a}' \times \boldsymbol{k}'),\ \text{with}\ \boldsymbol{z}' \in \mathbb{R}^{C\times H\times W}, \tag{5}$$

where $\boldsymbol{z}'$ denotes the re-estimated structure-enhanced features, and θ refers to a convolutional mapping. By this means, the structural knowledge of the different spine bones is fully

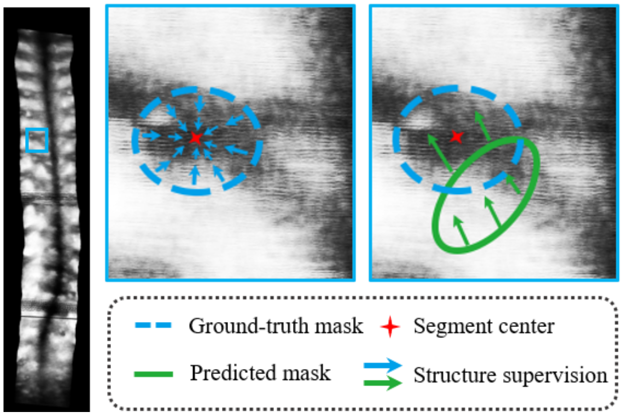

Fig. 2. Illustration of the structure supervision in the representation learning. The center regression task forces the learning of both the shape and the location of the segmentation mask.

explored in the re-estimated features, because the features are directly synthesized with the structure-regularized representations $\boldsymbol{k}'$. With the supervision on the key representation, the proposed SEAM can be regarded as an extension of the Synthesizer [14] in a structure-based manner.

To stabilize the learning process, we establish the residual connection as $\boldsymbol{z} = \rho([\boldsymbol{z}', \boldsymbol{v}])$, where $\boldsymbol{z} \in \mathbb{R}^{C \times H \times W}$ denotes the output representation; ρ refers to a convolutional mapping; $[\cdot]$ represents channel concatenation for feature fusion.

2.2. Learning Criteria

To effectively encode the structural knowledge in SEAM, we employ the similar penalty in SA-SSD [15] for center regression. Given a training pair $\{\boldsymbol{x}, \boldsymbol{y}\}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input observation and its ground-truth segments respectively, the network outputs both the key representation $\boldsymbol{k}$ in SEAM and the predicted segment mask $\hat{\boldsymbol{y}}$. To achieve the structure-enhanced attentive learning, we first utilize the category information to penalize $\boldsymbol{k}$ as follows:

$$\mathcal{L}^{\boldsymbol{k}}_{cls} = \frac{1}{M} \sum_{i=1}^{M} CE(\boldsymbol{k}_i, \boldsymbol{y}_i), \tag{6}$$

where CE refers to the standard Cross Entropy loss, and M denotes the number of elements in $\boldsymbol{k}$. By this means, each channel in the key representation can describe the features of one foreground segment or the background. Then, the three foreground attention channels are selected to perform pixel-wise center regression as follows:

$$\mathcal{L}^{\boldsymbol{k}}_{reg} = \frac{1}{M_{fg}} \sum_{i=1}^{M} \mathit{Smooth}\text{-}\ell_1(\Delta\hat{\boldsymbol{c}} - \Delta\boldsymbol{c}) \cdot \mathbb{1}[\boldsymbol{k}_i \neq 0], \tag{7}$$
$$\text{with } \Delta\hat{\boldsymbol{c}} = \boldsymbol{p}(\boldsymbol{k}_i) - \boldsymbol{p}(\hat{\boldsymbol{c}}); \Delta\boldsymbol{c} = \boldsymbol{p}(\boldsymbol{k}_i) - \boldsymbol{p}(\boldsymbol{c}),$$

where $\Delta\hat{\boldsymbol{c}}$ and $\Delta\boldsymbol{c}$ are the offsets between the pixel and the center of its corresponding estimated $\hat{\boldsymbol{c}}$ and ground-truth $\boldsymbol{c}$ segment respectively; $\boldsymbol{p}(\cdot)$ is the position function to obtain the normalized vertical and horizontal coordinates of the point; M_{fg} refers to the number of pixels belonging to the foreground segments; $\mathbb{1}[condition]$ is a conditional function, which is equal to 1 if the condition that the feature point $\boldsymbol{k}_i$ describes a foreground segment is satisfied, or otherwise 0. This center regression regularization not only forces the network to learn the shape of each segment, but also indicates the mask shift when localizing the foreground objects, as illustrated in Fig. 2. Finally, we consider the segmentation loss on the predicted pixel-wise label $\hat{\boldsymbol{y}}$ as follows:

$$\mathcal{L}_{seg} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{\#\text{cls}} \boldsymbol{y}_i \log(\hat{\boldsymbol{y}}_i), \tag{8}$$

where #cls refers to the number of classes. The overall objective function is formulated as:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda(\mathcal{L}^{\boldsymbol{k}}_{cls} + \beta \mathcal{L}^{\boldsymbol{k}}_{reg}), \tag{9}$$

where λ and β are the hyperparameters controlling the trade-off between the loss terms.

3. EXPERIMENT

3.1. Dataset

We collected 109 ultrasound VPI images from 109 subjects with different degrees of spine deformity. Each VPI image is generated by projecting a whole spine 3D ultrasound sequence into a 2D coronal plane. The ground-truth segments were manually annotated by ultrasound experts. We randomly split the dataset into a training branch and a testing branch of 80 and 29 samples, respectively. All images were rescaled to 1024×256. In the training phase, patches of size 512×256 were densely extracted from the resized training samples. Random flip and rotation were employed for data augmentation. In the testing phase, each query sample was first rescaled to 1024×256, and then fed to the segmentation model to produce the segmentation mask, which was then resized to the original resolution for assessment.

3.2. Implementation details

We establish the proposed framework with PyTorch [16] and MMSegmentation [17]. The backbone is built with the feature pyramid network (FPN) based on the same settings in [18]. We adopt FPN because it can fuse the multi-scale information of the image to promote segmentation. The segmentation head refers to the last convolutional layer to produce a four-channel tensor, indicating the segmentation predictions. In SEAM, all the convolutional kernels are of size 1×1 with padding 0, except for the last residual mapping ρ, where 3×3 filters with padding 1 are used. During training, we build a mini-batch with 12 training samples. The learning rate is initialized to 10^{-4} and gradually decreased to 5×10^{-6} with the cosine annealing strategy [19]. We adopt Adam [20] to

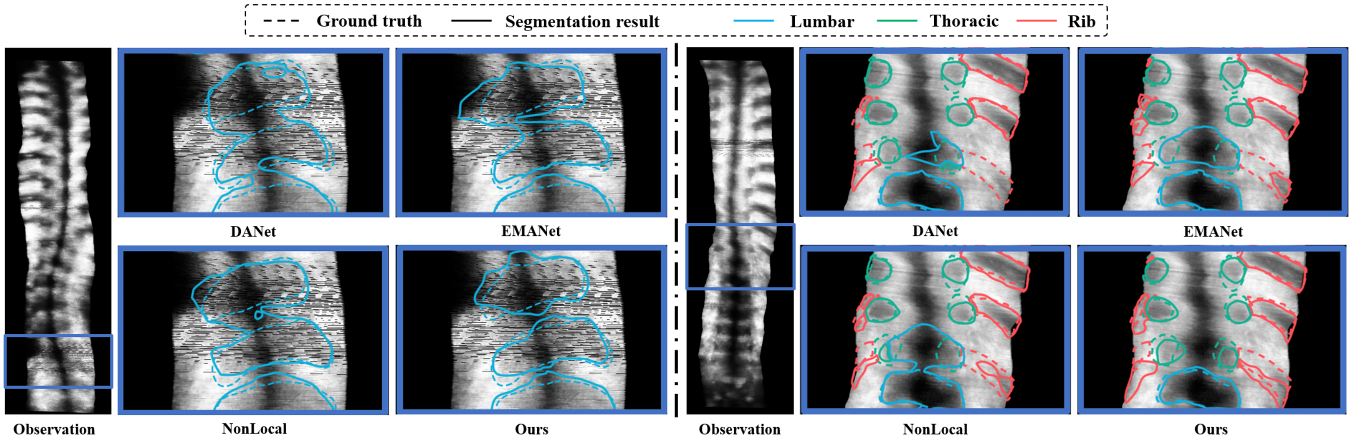

Fig. 3. Visualization of the predicted segments based on the different attention-based segmentation methods.

Table 1. Quantitative segmentation results, where D: Dice score (%), J: Jaccard index (%), and R: Runtime (s).

Methods	Lumbar		Thoracic		Rib		*Ave.*		R
	D	J	D	J	D	J	D	J	
Vanilla [18]	85.69	75.29	76.42	62.12	78.02	64.24	80.04	67.21	0.32
UNet [3]	82.21	70.26	74.70	59.94	77.37	63.46	78.09	64.56	0.25
PPMU [4]	84.58	73.68	76.55	62.22	78.21	64.48	79.78	66.79	0.61
RSNU [7]	85.85	75.52	77.45	63.39	79.26	65.92	80.86	68.28	0.32
WaveCT [8]	86.59	76.58	75.91	61.36	78.49	64.82	80.33	67.58	0.67
DANet [21]	83.75	72.60	75.93	61.47	77.48	63.53	79.05	65.86	0.32
EMANet [13]	84.73	73.82	77.68	63.64	79.02	65.54	80.48	67.66	0.46
NonLocal [22]	85.22	74.63	76.81	62.61	78.72	65.19	80.02	67.47	0.63
∼ w/o AL	85.68	75.24	77.94	64.02	79.79	66.61	81.14	68.62	0.32
∼ w/o SS	86.54	76.52	78.10	64.24	79.22	65.83	81.29	68.87	0.37
Ours	**87.09**	**77.45**	**78.32**	**64.58**	**80.16**	**67.02**	**81.85**	**69.68**	0.37

optimize the objective function defined in Eq. (9) with the hyperparameters, λ and β, empirically set to 0.4 and 0.5 respectively. We train the network on a Nvidia GTX 2080 Ti GPU, and it takes about 7 hours to train up the model.

3.3. Results

Quantitative segmentation results: To validate the proposed framework for spine segmentation, we compare our method with other state-of-the-art segmentation algorithms on ultrasound images. The results are tabulated in Table 1. Specifically, we consider the benchmark methods for medical image segmentation, i.e. the vanilla FPN model [18], UNet [3], and PPMU [4]; the recently proposed methods for ultrasound image segmentation, i.e. RSNU [7] and WaveCT [11]; and the state-of-the-art attention-based methods, i.e. DANet [21], EMANet [13], and NonLocal [22]. All methods are established based on the same settings as in Sec. 3.2, and we also build their models with the capacity equal to, or larger than, the proposed method. It can be seen that our proposed method outperforms all the benchmark methods [18, 3, 4] by a large margin. Compared to the methods designed for ultrasound images [7, 11], we can observe a significant improvement of over 1% in terms of both the Dice score and Jaccard index. More importantly, our proposed method also surpasses EMANet [13] by about 1.5% and 2% on Dice score and Jaccard index respectively, which demonstrates the benefit of introducing the structure supervision in an attentive manner for spine segmentation. We further visualize two segmentation results in Fig. 3, and compare them with the other attention-based algorithms. It can be seen that the structural knowledge from SEAM facilitates the segmentation by accurately locating the lumbars and preserving the shape of the each spine bone.

Ablation study: To perform a comprehensive study on the proposed method, we investigate different designs in our framework. Specifically, we explore the effect from the attentive learning (AL) and the structure supervision (SS). The model, without attentive learning, directly introduces the structure supervision to the features extracted from the FPN backbone, and the model, without structure supervision, is trained with $\beta = 0$. We can observe from Table 1 that the attention mechanism contributes greatly to the segmentation results. It enhances the contextual information between different segments, which leads to an improvement of about 0.7%. The structure supervision also promotes the segmentation by learning the shape and the position knowledge, which gains a Dice improvement of about 0.55%.

4. CONCLUSION

In this paper, we have proposed a novel framework to introduce the structural knowledge into the semantic representations for more effective spine segmentation from ultrasound volume projection images. To efficiently encode both the contextual and structural information into the learned semantic representations, we present a structure-enhanced attention module, and integrate it with a segmentation network. Extensive experimental results show that the proposed method outperforms other state-of-the-art segmentation algorithms, making it a potential solution to clinical scoliosis diagnosis.

5. ACKNOWLEDGEMENT

This study was partially supported by the Research Impact Fund (R5017-18) of Hong Kong Research Grant Council.

6. REFERENCES

[1] Chung-Wai James Cheung, Guang-Quan Zhou, Siu-Yin Law, Tak-Man Mak, Ka-Lee Lai, and Yong-Ping Zheng, "Ultrasound volume projection imaging for assessment of scoliosis," *IEEE transactions on medical imaging*, vol. 34, no. 8, pp. 1760–1768, 2015.

[2] J. COBB, "Outline for the study of scoliosis," *Instr Course Lect AAOS*, vol. 5, pp. 261–275, 1948.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[4] Ahmed H Shahin, Karim Amer, and Mustafa A Elattar, "Deep convolutional encoder-decoders with aggregated multi-resolution skip connections for skin lesion segmentation," in *ISBI*. IEEE, 2019, pp. 451–454.

[5] H. Xu, S. Geng, Y. Qiao, K. Xu, and Y. Gu, "Combining cgan and mil for hotspot segmentation in bone scintigraphy," in *ICASSP*, 2020, pp. 1404–1408.

[6] Cheng Chen, Qi Dou, Hao Chen, and Pheng-Ann Heng, "Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation," in *International workshop on machine learning in medical imaging*. Springer, 2018, pp. 143–151.

[7] Zixun Huang Li-Wen Wang, Frank H. F. Leung, Sunetra Banerjee, De Yang, Timothy Lee, Juan Lyu, Sai Ho Ling, and Yong-Ping Zheng, "Bone feature segmentation in ultrasound spine image with robustness to speckle and regular occlusion noise," in *SMC*, 2020.

[8] Z. Liu, X. Yang, R. Gao, S. Liu, H. Dou, S. He, Y. Huang, Y. Huang, H. Luo, Y. Zhang, Y. Xiong, and D. Ni, "Remove appearance shift for ultrasound image segmentation via fast and universal style transfer," in *ISBI*, 2020, pp. 1824–1828.

[9] Qiang Wang, Qingqing Chang, Yu Qiao, Yuyuan Zhu, Gang Huang, and Jie Yang, "Knowledge-based segmentation of spine and ribs from bone scintigraphy," in *Neural Information Processing*, 2011, pp. 241–248.

[10] Malinda Vania, Dawit Mureja, and Deukhee Lee, "Automatic spine segmentation from ct images using convolutional neural network via redundant generation of class labels," *Journal of Computational Design and Engineering*, vol. 6, no. 2, pp. 224 – 232, 2019.

[11] Yang Lei, Xue Dong, Zhen Tian, Yingzi Liu, Sibo Tian, Tonghe Wang, Xiaojun Jiang, Pretesh Patel, Ashesh B Jani, Hui Mao, et al., "Ct prostate segmentation based on synthetic mri-aided deep attention fully convolution network," *Medical physics*, vol. 47, no. 2, pp. 530–540, 2020.

[12] Fei Ding, Gang Yang, Jinlu Liu, Jun Wu, Dayong Ding, Jie Xv, Gangwei Cheng, and Xirong Li, "Hierarchical attention networks for medical image segmentation," *arXiv preprint arXiv:1911.08777*, 2019.

[13] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu, "Expectation-maximization attention networks for semantic segmentation," in *ICCV*, 2019, pp. 9167–9176.

[14] Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng, "Synthesizer: Rethinking self-attention in transformer models," *arXiv preprint arXiv:2005.00743*, 2020.

[15] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang, "Structure aware single-stage 3d object detection from point cloud," in *CVPR*, 2020.

[16] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," in *NIPS Workshop*, 2017.

[17] Jiarui Xu et al., "Mmsegmentation," in *https://github.com/open-mmlab/mmsegmentation*, 2020.

[18] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar, "Panoptic feature pyramid networks," in *CVPR*, Jun 2019.

[19] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[20] Diederik P. and Jimmy Ba Kingma, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[21] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.