Frequency-enhanced Multi-granularity Context Network for Efficient Vertebrae Segmentation

Jian Shi[1], Tianqi You[1], Pingping Zhang[2(✉)], Hongli Zhang[3], Rui Xu[1], and Haojie Li[4]

[1] School of Software Technology & DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China
[2] School of Future Technology, School of Artificial Intelligence, Dalian University of Technology, Dalian, China
[3] The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[4] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China
zhpp@dlut.edu.cn

Abstract. Automated and accurate segmentation of individual vertebra in 3D CT and MRI images is essential for various clinical applications. Due to the limitations of current imaging techniques and the complexity of spinal structures, existing methods still struggle with reducing the impact of image blurring and distinguishing similar vertebrae. To alleviate these issues, we introduce a Frequency-enhanced Multi-granularity Context Network (FMC-Net) to improve the accuracy of vertebrae segmentation. Specifically, we first apply wavelet transform for lossless downsampling to reduce the feature distortion in blurred images. The decomposed high and low-frequency components are then processed separately. For the high-frequency components, we apply a High-frequency Feature Refinement (HFR) to amplify the prominence of key features and filter out noises, restoring fine-grained details in blurred images. For the low-frequency components, we use a Multi-granularity State Space Model (MG-SSM) to aggregate feature representations with different receptive fields, extracting spatially-varying contexts while capturing long-range dependencies with linear complexity. The utilization of multi-granularity contexts is essential for distinguishing similar vertebrae and improving segmentation accuracy. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches on both CT and MRI vertebrae segmentation datasets. The source code is publicly available at https://github.com/anaanaa/FMCNet.

Keywords: Vertebrae Segmentation · Wavelet Transform · Frequency Feature · Multi-granularity · State Space Model.

1 Introduction

The spine is vital to the body, supporting the organ function and serving as the core of the neural pathways. Any injury to the spine can impair neural conduction

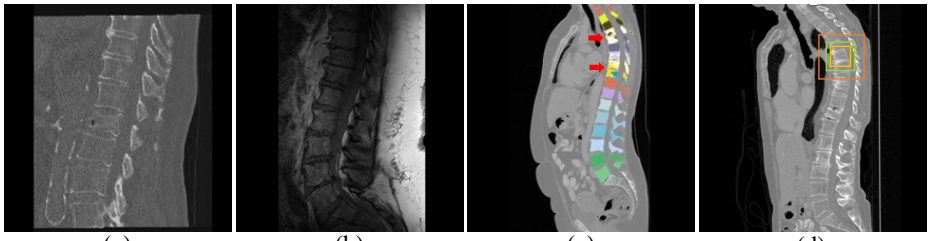

Fig. 1. Challenging examples in vertebrae segmentation. (a) and (b) present examples of image blurring caused by current imaging techniques in CT and MRI modalities, respectively. (c) presents examples of the challenges in distinguishing similar vertebrae. (d) highlights the importance of multi-granularity contextual information contained in the features of different receptive fields for distinguishing similar vertebrae.

and lead to associated disorders [1]. Automatic segmentation and identification of individual vertebra in the spine are crucial for the treatment of orthopedic diseases, as they play a key role in preoperative diagnosis, treatment planning, surgical guidance, and postoperative assessment [2–4]. Although significant efforts have been made to achieve highly accurate vertebrae segmentation, the performance still remains suboptimal due to the limitations of current imaging techniques and the complexity of spinal structures [5]. As illustrated in Fig. 1, there are two main challenges in this task: (1) The intrinsic properties of different 3D medical imaging modalities, such as high noise levels and low contrast in CT scans, and non-isotropic spatial resolution in MRI images, lead to image blurring and distortion. Examples of these phenomena in CT and MRI images are in illustrated Fig. 1 (a) and Fig. 1 (b), respectively. (2) The high similarity in shape and structure between adjacent vertebrae presents a significant challenge for accurate segmentation. Fig. 1 (c) demonstrates the impact of homogeneous structures on vertebrae segmentation, showing that partial voxel predictions for adjacent vertebrae are incorrectly assigned the same label.

Existing vertebrae segmentation methods typically employ an encoder-decoder structure and have achieved commendable results [6, 7]. However, previous methods often overlook the critical impact of downsampling techniques. The max-pooling and bilinear interpolation are commonly utilized, frequently resulting in a substantial loss of fine-grained details during the downsampling process. This loss is especially harmful for blurred images, as retaining fine-grained details is essential for accurate segmentation. Besides, vertebrae segmentation requires sufficient contextual information to distinguish vertebrae. Existing Convolutional Neural Networks (CNNs) rely on local convolutional kernels to extract features, which limits their ability to extract global and geometric features. Although Transformer-based methods [8, 9] excel in capturing long-range dependencies, their self-attention mechanism has a high computational complexity. Recently, the State Space Model (SSM) [10, 11] could capture the long-range dependencies with linear complexity, providing a promising solution to the limitations of CNNs

and Transformers. While SSM lacks consideration of spatially-varying contexts, this limitation reduces its sensitivity to multi-scale contextual information in similar vertebrae, impairing its ability to distinguish them.

To mitigate the aforementioned issues, we introduce a Frequency-enhanced Multi-granularity Context Network (FMC-Net) for accurate vertebrae segmentation. To begin with, we replace traditional downsampling methods with wavelet transform to reduce the loss of detailed information in blurred images. The wavelet transform achieves downsampling by decomposing the features into high and low-frequency components, which are then processed separately in the subsequent steps. High-frequency components contain a wealth of image details, textures, and edge information, but they also include noises. Therefore, we employ a High-frequency Feature Refinement (HFR) to amplify the prominence of key features while filtering out noises, effectively restoring details and textures in blurred images. Low-frequency components contain the most image content, we propose a Multi-granularity State Space Model (MG-SSM) to comprehensively extract low-frequency information. Specifically, MG-SSM contains granularity-aware SSMs for spatial perception. These SSMs extract information at different granularities to distinguish similar vertebrae. We evaluate the proposed method on both CT and MRI vertebrae segmentation datasets. Extensive experimental results demonstrate that our approach achieves state-of-the-art performance.

Our main contributions can be summarized as follows:

• We propose FMC-Net that leverages wavelet transform to avoid information loss and enhances high and low-frequency features to address the challenges of image blurring and the high similarity of vertebral structures.

• For the high-frequency components, we introduce the HFR to restore fine-grained details. For the low-frequency components, we employ the MG-SSM to capture spatially-varying long-range dependencies.

• Experiments on VERSE2019 and LUMBAR datasets demonstrate the superior performance of FMC-Net compared to other state-of-the-art approaches.

2 Method

The Overall Architecture: As shown in Fig. 2, given a 3D input volume $I \in \mathbb{R}^{D\times H\times W}$, D, H and W represent the depth, height, and width of the volume, respectively. FMC-Net aims to achieve accurate vertebrae segmentation by employing a wavelet-based encoder-decoder structure integrated with HFR and MG-SSM. Firstly, in the encoder stage, wavelet transform enables lossless downsampling, separating the features into high-frequency and low-frequency components. The HFR and MG-SSM are respectively applied to these components, yielding enhanced feature representations. These enhanced components are fused as the final features at each stage, effectively capturing both semantic and detailed information. In the decoder stage, Wavelet Transform Upsampling (WTU) employs wavelet transform-based feature fusion during upsampling, enabling the decoder to effectively leverage the information from the encoder.

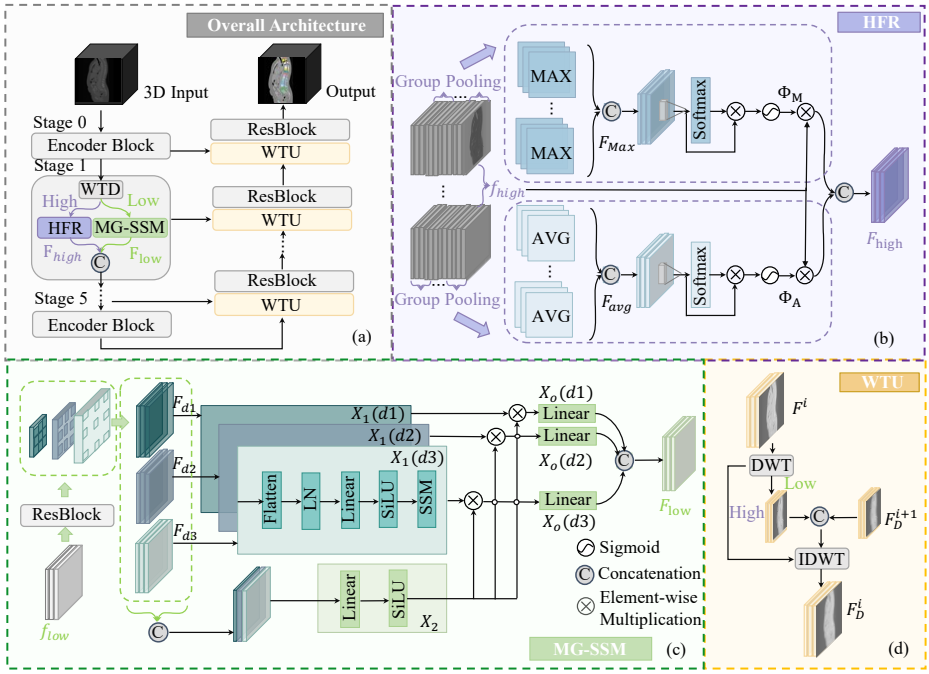

Fig. 2. (a) is the overall architecture of the proposed FMC-Net. (b) and (c) are HFR and MG-SSM, which enhance the high-frequency and low-frequency components, respectively. (d) is WTU, which utilizes features from different scales of the encoder while performing upsampling in the decoder.

2.1 Wavelet Transform-based Sampling

Wavelet Transform Downsampling. Inspired by Guo *et al.* [12], we implement Wavelet Transform Downsampling (WTD) using the Discrete Wavelet Transform (DWT) [13] to alleviate the irreversible distortion and information loss caused by downsampling. For the feature $F^i \in \mathbb{R}^{\left(\frac{D}{2^i} \times \frac{H}{2^i} \times \frac{W}{2^i}\right)}$ of the i-th $(i \in (0, 1, \ldots, 5)$) stage encoder, we employ DWT to decompose the feature F^i into eight sub-wavelet bands, i.e., low-frequency component F^i_{lll}, and high-frequency components $F^i_{\{llh,\ldots,hhh\}}$ in horizontal, vertical, and depth dimensions (which are x-axis, y-axis and z-axis):

$$\left\{F^i_{lll}, F^i_{llh}, F^i_{lhl}, F^i_{lhh}, F^i_{hll}, F^i_{hlh}, F^i_{hhl}, F^i_{hhh}, \right\} = DWT(F^i), \tag{1}$$

where $F^i_{\{lll,\ldots,hhh\}} \in \mathbb{R}^{\left(\frac{D}{2^{i+1}} \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}\right)}$. As low-frequency components contain the majority of the essential information, while high-frequency components capture intricate details, textures, and edge information, we focus on enhancing both of these components:

$$F^i_{low} = \mathcal{S}\left(F^i_{lll}\right), \tag{2}$$

$$F^i_{high} = \mathcal{H}\left(F^i_{\{llh,\ldots,hhh\}}\right), \tag{3}$$

where F^i_{low} and F^i_{high} denote the enhanced low-frequency and high-frequency features, respectively. $\mathcal{S}$ denotes MG-SSM, and $\mathcal{H}$ denotes HFR. Then the enhanced feature $F^{i+1} \in \mathbb{R}^{\left(\frac{D}{2^{i+1}} \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}\right)}$ is derived by concatenating F^i_{low} and F^i_{high}, which contains both semantic and fine-grained detail information.

Wavelet Transform Upsampling. To fully leverage the features from the encoder, we propose Wavelet Transform Upsampling (WTU), which effectively utilizes features from different scales of the encoder while performing upsampling in the decoder. For the features in the encoder $F^i \in \mathbb{R}^{\left(\frac{D}{2^i} \times \frac{H}{2^i} \times \frac{W}{2^i}\right)}$, we apply DWT to halve the resolution and decompose them into low-frequency and high-frequency components, F^i_{lll}, $F^i_{\{llh,\ldots,hhh\}} \in \mathbb{R}^{\left(\frac{D}{2^{i+1}} \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}\right)}$. The low-frequency component F^i_{lll} is then concatenated with the features $F^{i+1}_D \in \mathbb{R}^{\left(\frac{D}{2^{i+1}} \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}\right)}$ from the decoder to obtain the feature F_{fuse}. Finally, both the high-frequency components $F^i_{\{llh,\ldots,hhh\}}$ and F_{fuse} are input into the inverse Discrete Wavelet Transform (IDWT) to generate the upsampled feature $F^i_D \in \mathbb{R}^{\left(\frac{D}{2^i} \times \frac{H}{2^i} \times \frac{W}{2^i}\right)}$.

2.2 High-frequency Feature Refinement

High-frequency components capture detailed textures, but also contain noises. We employ HFR to effectively leverage high-frequency information. The high-frequency components f_{high} (comprising $F_{llh}, \ldots, F_{hhh}$) are processed through two distinct paths: one focuses on amplifying the prominence of key features, and the other filters out noises, thereby facilitating the restoration of fine-grained details and textures in blurred images. Both paths employ spatial attention mechanisms to dynamically modulate different frequency components during training. In the amplification path, the spatial attention mechanism Φ_M can be formulated as:

$$F_{\text{Max}} = \mathcal{F}^{7C\to C}_{3\times 3}([\text{GroupMax}(F_{llh}), \ldots, \text{GroupMax}(F_{hhh})]), \tag{4}$$

$$\Phi_M = \sigma(\text{Softmax}(F_{\text{Max}}) \otimes F_{\text{Max}}), \tag{5}$$

where σ denotes the Sigmoid function, $[\cdot]$ is the concatenation operation, and GroupMax is the group max-pooling operation. Each high-frequency component is divided into 2^i groups for pooling, and i denotes the number of encoder stages. After pooling, each component undergoes a dimension transformation that reduces the channel dimension from 2^i to C. $\mathcal{F}^{7C\to C}_{3\times 3}$ is a 3×3 convolutional layer that transforms the dimension of the concatenated high-frequency feature from $7C$ to C. The concatenation operation aggregates the amplified features of each high-frequency component after max-pooling operations, while the Softmax operation computes the contribution of each component to the spatial attention map. Meanwhile, the denoising path follows the same structure as the amplification path, except that the average pooling is used instead of max-pooling.

Similarly, a spatial attention map Φ_A is generated, which smooths the features of each component and effectively filters out noises. The final enhanced high-frequency features can be obtained by the following formula:

$$F_{high} = [\Phi_M \otimes F_{\{llh,\ldots,hhh\}}, \Phi_A \otimes F_{\{llh,\ldots,hhh\}}]. \tag{6}$$

Here, the spatial attention map Φ_M is multiplied with each high-frequency component to emphasize the key features, while Φ_A performs a similar operation to filter out noises from each component. Finally, the amplified and denoised components are concatenated to obtain the enhanced high-frequency component.

2.3 Multi-Granularity State Space Model

For low-frequency component f_{low}, we propose MG-SSM to effectively utilize the rich information for distinguishing similar vertebrae. Firstly, we employ a residual block to extract local features. Then we use three parallel 3D depth-wise convolutional layers with different dilation rates to capture deep features with different scales of receptive fields. Subsequently, these features (F_{d1}, F_{d2}, F_{d3}) are respectively fed into three Visual State Space Modules (VSSMs) to learn multi-scale contexts. This process can be defined as following:

$$\begin{aligned} F_{dj} &= \mathrm{DWConv}_{dj}(\mathrm{ResBlock}(F_{lll})), j \in (1,2,3), \\ X_1(dj) &= \mathrm{SSM}\left(\phi\left(\mathrm{Linear}\left(\mathrm{LN}(F_{dj})\right)\right)\right), \end{aligned} \tag{7}$$

where dj represents the j-th dilation rate, ϕ denotes the Sigmoid Linear Unit (SiLU) activation function [14], and $X_1(dj)$ is the first branch of each VSSM. Next, F_{d1}, F_{d2}, and F_{d3} are concatenated to generate multi-scale features while simultaneously capturing the inter-scale relationships. And the concatenated feature is then fed into the second branch of VSSM:

$$X_2 = \phi(\mathrm{Conv}([F_{d1}, F_{d2}, F_{d3}]), \tag{8}$$

The features from both branches are combined to produce the output:

$$X_o(dj) = \mathrm{Linear}(X_1(dj) \otimes X_2), \tag{9}$$

where X_2 is shared among three VSSMs. Finally, each VSSM output is passed through a linear layer, followed by concatenation to generate the final enhanced low-frequency component. MG-SSM addresses the limitations of a single SSM in capturing spatially-varying features and provides multi-granularity contextual information for distinguishing similar vertebrae.

3 Experiments

Datasets and Metrics. We evaluate the effectiveness of our approach on two public vertebrae segmentation datasets. The VERSE2019 dataset [15] consists of 160 CT scans with 26 segmentation classes, covering vertebrae from the cervical

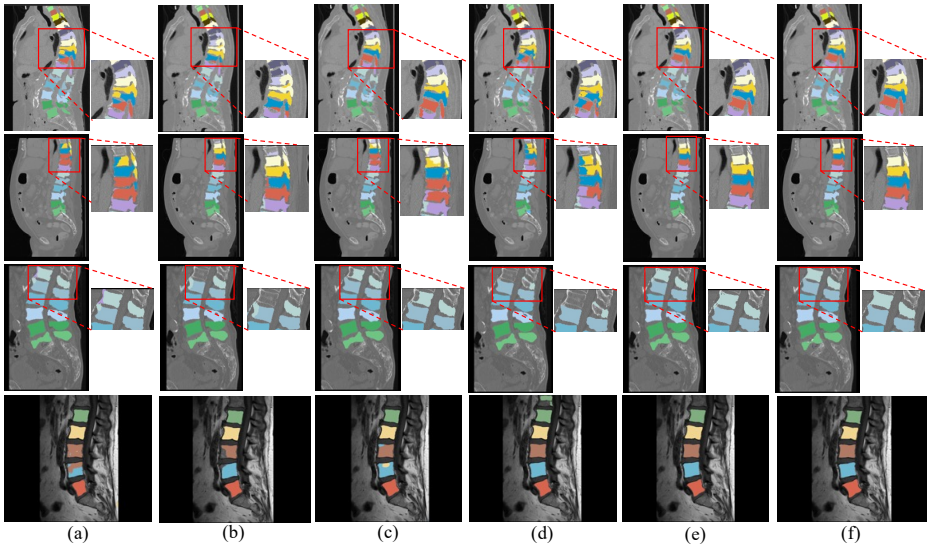

Fig. 3. Qualitative results with different methods. The top three rows display segmentation results from the VERSE2019 dataset, covering cervical, thoracic, and lumbar regions with various regions and views, whereas the bottom row shows results from the LUMBAR dataset. (a) nn-UNet, (b) UMamba, (c) MambaClnix, (d) Segmamba, (e) FMC-Net (Ours), (f) Ground Truth.

to the lumbar regions. The LUMBAR dataset [16] contains 156 MRI scans from multiple scanner models and vendors, including T1-weighted, contrast-enhanced T1-weighted, and T2-weighted sequences. It provides segmentation labels for lumbar vertebrae and intervertebral discs. We use only the lumbar vertebrae labels and conduct experiments on T1-weighted images due to their superior quality. Following the previous works [15, 21], we use two standard metrics to evaluate segmentation performance: Dice Similarity Coefficient (DSC) and 95th percentile of Hausdorff Distance (HD95).

Implementation Details. The proposed method is implemented in PyTorch and runs on an NVIDIA-A800-SXM4-80GB GPU. It is integrated within the nnU-Net framework for automatic data processing and augmentation, and employs a combined loss of weighted cross-entropy and Dice losses at each decoder stage. The model is trained on the VERSE2019 dataset with a patch size of (128, 160, 96) for 1000 epochs and on the LUMBAR dataset with a patch size of (32, 256, 160) for 600 epochs, both using a learning rate of 0.01.

Comparison with State-of-the-art Methods. We compare the proposed model with three categories of methods to evaluate its effectiveness. The first category includes CNN-based methods, such as nn-UNet [17]. The second category includes Transformer-based approaches, comprising Swin-UNetR [18] and

Table 1. Comparison of different methods on VERSE2019 and LUMBAR datasets. Cer., Tho., Lum. denote cervical, thoracic, and lumbar vertebrae, respectively.

Methods	DSC (%)					HD95 (mm) ↓				
	VERSE2019				LUMBAR	VERSE2019				LUMBAR
	Cer.	Tho.	Lum.	Mean	Lum.	Cer.	Tho.	Lum.	Mean	Lum.
nn-UNet	77.89	74.77	70.95	74.73	61.76	**1.21**	5.12	**3.94**	3.74	39.38
CoTR	74.56	67.15	64.57	68.61	72.55	7.93	10.15	11.39	9.83	31.87
Swin-UNetR	79.30	56.90	60.90	64.73	47.75	2.12	17.15	9.07	11.00	97.28
UMamba	83.54	79.34	71.01	78.28	72.47	1.60	**3.13**	4.75	3.09	27.59
SegMamba	73.90	74.63	65.43	72.22	78.17	5.70	4.86	11.70	6.73	17.53
MambaClinix	83.94	71.26	67.93	74.01	55.79	1.91	7.80	5.54	5.61	29.09
Ours	**85.46**	**80.74**	**71.83**	**79.92**	**79.13**	1.36	3.32	4.47	**3.04**	**17.13**

Table 2. Ablation experiments with different components on LUMBAR dataset.

Baseline	DWT-Sample	HFR	MG-SSM	DSC%	HD95
✓				71.09	42.77
✓	✓			71.75	22.12
✓	✓	✓		75.30	24.52
✓	✓		✓	76.56	18.67
✓			✓	75.94	16.72
✓	✓	✓	✓	79.13	17.13

CoTR [19]. The third category consists of Mamba-based methods, including UMamba [20], SegMamba [21] and MambaClinix [22]. he evaluation results on the VERSE2019 and LUMBAR datasets are presented in Tab. 1. The VERSE2019 dataset is very challenging due to the inclusion of cervical, thoracic, and lumbar regions with varying fields of view. Although the performance improvements of our method over existing approaches may be modest in certain regions, it still achieves a new state-of-the-art. Compared with the best-performing Mamba-based method, SegMamba, our approach improves the mean DSC per class by 1.64%. On the LUMBAR dataset, which contains only lumbar vertebrae, our method demonstrates a more substantial performance gain, achieving 79.13% in DSC and 17.13 in HD95. Fig. 3 shows representative visualizations from both datasets for qualitative comparison. The visualizations demonstrate that our method has a significant advantage in distinguishing similar vertebrae, showcasing its robustness and effectiveness in the vertebrae segmentation task.

Ablation Study. To evaluate the effectiveness of each component within the proposed model, we conduct ablation studies on the LUMBAR dataset. Tab. 2 illustrates the contribution of each component to the performance. Specifically, we employ a standard U-shaped model with additive residual blocks as the baseline. Subsequently, we incrementally integrate the DWT-Sample, HFR, and MG-

SSM to evaluate the effectiveness of each component. Here, DWT-Sample refers to the combination of WTD and WTU, which are utilized for downsampling and upsampling during the encoder and decoder stages, respectively. The results demonstrate that each component significantly enhances the overall performance, confirming the effectiveness of each individual component.

4 Conclusion

In this paper, we propose a novel vertebrae segmentation network, FMC-Net, which effectively addresses the challenges of image blurring and the difficulty in distinguishing similar vertebrae. We take advantage of the wavelet transform to prevent the information loss, decomposing the features into low-frequency and high-frequency components for feature enhancement. HFR refines the high-frequency components, successfully restoring fine-grained details and alleviating image blurring. MG-SSM extracts spatially-varying contexts and long-range dependencies, which enable the full utilization of the rich information contained in the low-frequency components. Extensive experiments demonstrate that FMC-Net outperforms state-of-the-art approaches on both CT and MRI vertebrae segmentation datasets.

Acknowledgments. This study was supported by the Fundamental Research Funds for the Central Universities with No. DUT24YG135 and DUT23YG232.

Disclosure of Interests. All authors declare that they have no conflicts of interest.

References

1. Wang, H., Song, Q., Yin, R., Ma, R.: B-spine: Learning B-spline curve representation for robust and interpretable spinal curvature estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, No. 6, pp. 5381-5389 (2024).
2. Pang, S., Pang, C., Zhao, L., Chen, Y., Su, Z., Zhou, Y., Feng, Q.: SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. IEEE Transactions on Medical Imaging, 40(1), 262-273 (2020).
3. Lessmann, N., Van Ginneken, B., De Jong, P. A., Išgum, I.: Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. Medical Image Analysis, 53, 142-155 (2019).
4. Wu, H., Zhang, J., Fang, Y., Liu, Z., Wang, N., Cui, Z., Shen, D.: Multi-view vertebra localization and identification from ct images. In: Medical Image Computing and Computer Assisted Intervention, pp. 136-145 (2023).
5. Tao, R., Liu, W., Zheng, G.: Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers. Medical Image Analysis, 75, 102258 (2022).
6. Shi, W., Xu, T., Yang, H., Xi, Y., Du, Y., Li, J., Li, J.: Attention gate based dual-pathway network for vertebra segmentation of X-ray spine images. IEEE Journal of Biomedical and Health Informatics, 26(8), 3976-3987 (2022).

7. Masuzawa, N., Kitamura, Y., Nakamura, K., Iizuka, S., Simo-Serra, E.: Automatic segmentation, localization, and identification of vertebrae in 3D CT images using cascaded convolutional neural networks. In: Medical Image Computing and Computer Assisted Intervention, pp. 681-690 (2020).
8. You, X., Gu, Y., Liu, Y., Lu, S., Tang, X., Yang, J.: EG-Trans3DUNet: a single-staged transformer-based model for accurate vertebrae segmentation from spinal CT images. In: International Symposium on Biomedical Imaging, pp. 1-5 (2022).
9. You, X., Gu, Y., Liu, Y., Lu, S., Tang, X., Yang, J.: VerteFormer: A single-staged Transformer network for vertebrae segmentation from CT images with arbitrary field of views. Medical Physics, 50(10), 6296-6318 (2023).
10. Smith, J. T., Warrington, A., Linderman, S. W.: Simplified state space layers for sequence modeling. In: International Conference on Learning Representations, (2022).
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling. (2024).
12. Xu, G., Liao, W., Zhang, X., Li, C., He, X., Wu, X.: Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. Pattern Recognition, 143, 109819 (2023).
13. Li, Q., Shen, L., Guo, S., Lai, Z.: Wavelet integrated CNNs for noise-robust image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7245-7254 (2020).
14. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks, 107: 3-11 (2018).
15. Sekuboyina, A., Husseini, M. E., Bayat, A., Löffler, M., Liebl, H., Li, H., Kirschke, J. S.: VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. Medical Image Analysis, 73, 102166 (2021).
16. Khalil, Y. A., Becherucci, E. A., Kirschke, J. S., Karampinos, D. C., Breeuwer, M., Baum, T., Sollmann, N.: Multi-scanner and multi-modal lumbar vertebral body and intervertebral disc segmentation database. Scientific Data, 9(1), 97 (2022).
17. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Maier-Hein, K. H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203-211 (2021).
18. Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730-20740 (2022).
19. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention, pp. 171-180 (2021).
20. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024).
21. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 578-588 (2024).
22. Bian, C., Xia, N., Yang, X., Wang, F., Wang, F., Wei, B., Dong, Q.: Mambaclinix: Hierarchical gated convolution and mamba-based u-net for enhanced 3d medical image segmentation. arXiv preprint arXiv:2409.12533 (2024).