VesMamba: 3D Pulmonary Vessel Segmentation from CT images via Mamba with Structural Perception and Scale-aware Filtering

Zhipeng Liu[1], Guilian Chen[1], Zheng Jiang[1], Huisi Wu[1]*, Jing Qin[2]
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
2400101053@mails.szu.edu.cn, hswu@szu.edu.cn

Abstract

Automated 3D pulmonary vessel segmentation from CT images is crucial for improving early screening and assessment of pulmonary vessel related diseases. However, it remains an extremely challenging task due to the complex and tree-like structures of vessels, large scale-variations, and the existence of highly similar tissues in the background. Existing segmentation models either cannot sufficiently capture long-range structural dependencies, which are of great importance in vessel segmentation, or are constrained by insufficient computational resources in clinical settings. In this paper, we propose VesMamba, a novel model for 3D pulmonary vessel segmentation that comprehensively addresses these challenges. Specifically, we first devise a spatial-gated structural perception (SSP) module, which employs Mamba to efficiently capture long-range dependencies. In SSP, we design dynamic spatial attention convolutions (DSAC) for dynamically learning the tree-like 3D vessel structures, providing Mamba with the spatial perception capability to better track the complicated topologies of vessels. Second, we propose an innovative bidirectional scale-aware filter (BSF) module to strengthen the representation capability of the encoder, facilitating our model to focus on vessels of different scales under noise. Moreover, we apply a mask-constrained decoder to further improve segmentation consistency and accuracy, which constrains the inference of adjacent low-layer decoders directly by high-layer masks. Extensive experiments on the public Parse22 and internal Lung79 datasets demonstrate that our method can achieve better performance than SOTAs. Code is available at https://github.com/Lzpbright/VesMamba.

1. Introduction

Pulmonary vessel diseases have become one of the most important causes threatening human health [14]. Computed tomography (CT) is the major imaging modality for the diagnosis and treatment planning of pulmonary vessel diseases. However, manual pulmonary vessel analysis based on CT images is time-consuming, labor-intensive, and experience-dependent. To this end, automated and accurate pulmonary vessel segmentation from CT images is of great clinical significance in practice. Yet, it is an extremely challenging task for the following reasons. First, pulmonary vessels present a very complicated tree-like anatomical structure with large variation in scales, as shown in Figure 1 (a-b). Second, there are many small-scale vessels with quite low contrast in CT images, while motion artifacts in CT images make them even more difficult to identify, as shown in Figure 1 (c-d). Third, it is difficult to discriminate overlapping vessels, as well as tissues with high similarity, such as the arteries and veins, as shown in Figure 1 (e-f).

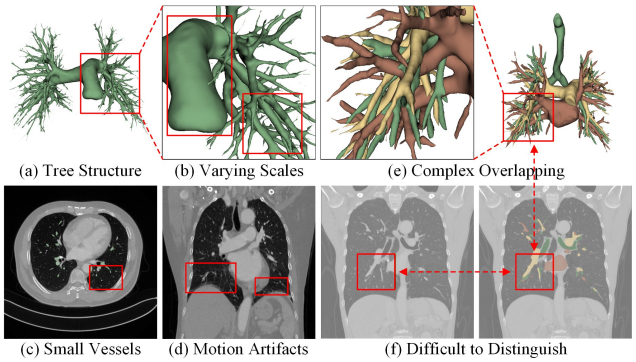

Figure 1. Challenges in 3D pulmonary vessel segmentation: (a) complex tree-like anatomical structure, (b) large variation in scales, (c) small vessels with low contrast, (d) motion artifacts in CT images, (e) complex overlapping, and (f) high similarity between artery and vein.

A lot of effort has been dedicated to addressing these challenges. Early investigations attempt to segment vessels via grayscale features [23, 33, 41], but they largely fail to address most of the aforementioned challenges. In

*Corresponding Author

recent years, many deep learning-based approaches have been proposed. CNN-based models are first applied for 3D vessel segmentation [12, 31]. Later, CNN-Transformer hybrid models [38] are proposed to employ transformers to extract vessel distribution information, and then combine the local features obtained from CNNs to improve segmentation accuracy. However, most of these methods ignore or do not sufficiently leverage the topological structure of vessels. To further improve performance, some studies introduce tree-like topological structures to guide the model in learning the semantic features of vessels [13], while other studies propose adaptive convolutions to simulate vessel topology [3, 25] or design special losses for vessel structures [29]. However, these methods are still incapable of efficiently modeling complicated and multi-scale vessels with reasonable computational resources in clinical settings, particularly under noise and complex backgrounds. Recently, Mamba [6], which is based on the state space model (SSM) [4], has shown significant advantages in sequence modeling with linear time complexity. Given its strong long-term modeling capability and computational efficiency, some studies redesign Mamba for 3D medical image segmentation [21, 40]. There are also a few studies that investigate how to employ multi-scale SSM for blood vessel segmentation [35]. However, while Mamba exhibits great potential in 3D pulmonary vessel segmentation, existing models do not sufficiently explore its advantages in this task, neither improving segmentation accuracy nor reducing computational burdens.

In this paper, to comprehensively address the above-mentioned challenges, we propose a novel 3D vessel segmentation model based on Mamba; we call it VesMamba. Specifically, we first incorporate Mamba into an innovative spatial-gated structural perception (SSP) module to efficiently capture the long-range topological dependencies of the vessels. In SSP, we devise a novel dynamic spatial attention convolution (DSAC) that dynamically adjusts directional weights to cope with the spatial anisotropy of 3D vessel data, ultimately highlighting features along the spatial direction of vessels. This enables Mamba to model long-term dependencies while also acquiring the spatial perception capability. Furthermore, we propose a bidirectional scale-aware filtering (BSF) module, which, considering the issues of feature interference and propagation, innovatively employs the bidirectionally fused features not as the output but for multi-scale noise filtering to enhance the representation capacity of encoder features at each layer. Finally, we design a mask-constrained decoder (MCDecoder) based on deep supervision, where high-layer masks are harnessed to directly constrain the inference of adjacent low-layer encoders to ensure the consistency of vessel segmentation while improving accuracy. Extensive experiments on a benchmarking public 3D pulmonary vessel dataset, Parse22 [20], and an in-house 3D pulmonary airway-artery-vein dataset, Lung79, demonstrate that our method achieves better performance than SOTA approaches. Our main contributions are summarized as follows.

- We propose a novel VesMamba for 3D pulmonary vessel segmentation from CT images, where Mamba is innovatively adapted for efficient long-range dependency modeling and spatial perception to capture the structural and topological features of 3D vessels.
- We further propose a BSF module to suppress noise in encoder features at different scales while highlighting vessels in complicated background, thereby producing more robust features to counteract noise interference.
- We extensively evaluate the proposed VesMamba on a public dataset and an in-house dataset, achieving better performance than SOTAs on both datasets.

2. Related Work

2.1. 3D Vessel Segmentation

Early 3D vessel segmentation methods rely on thresholding [33], morphological operations [23], and region growing [41], but fail to segment low-contrast small vessels. With the development of deep learning, some CNN-based models [12, 31] (CNNs) achieve preliminary 3D vessel segmentation. However, CNNs are limited in capturing global features, inevitably susceptible to background noise. Therefore, transformers [5] are developed in CNN-Transformer models [38] to improve segmentation performance. Moreover, to address the characteristics of vessel morphology, DSCNet [25] simulates vessel morphology by using adaptive convolution kernels and topological continuity loss for vessel structures. However, it is prone to background interference and struggles with the task of segmenting non-tubular objects.

2.2. Mamba-based Medical Image Segmentation

Recently, state space models (SSMs) [4] represented by Mamba have attracted widespread attention. SSMs are originally proposed for sequence processing, but their efficient long-range dependency modeling capability has also shown significant potential in vision tasks. With VMamba [19] pioneering the application of Mamba to vision tasks, some Mamba-based methods are developed for medical image segmentation. UMamba [21] designs a hybrid CNN-SSM block to capture long-range dependencies within 3D medical images. However, due to the characteristics of recursive state updates in the standard Mamba, it cannot directly perceive adjacent local features. In addition, VisionMamba [42] proposes a bidirectional modeling method that is not fully explored in 3D segmentation tasks. SegMamba [40] proposes a 3D medical image segmentation framework with its Triple-Space Mamba (ToM) module,

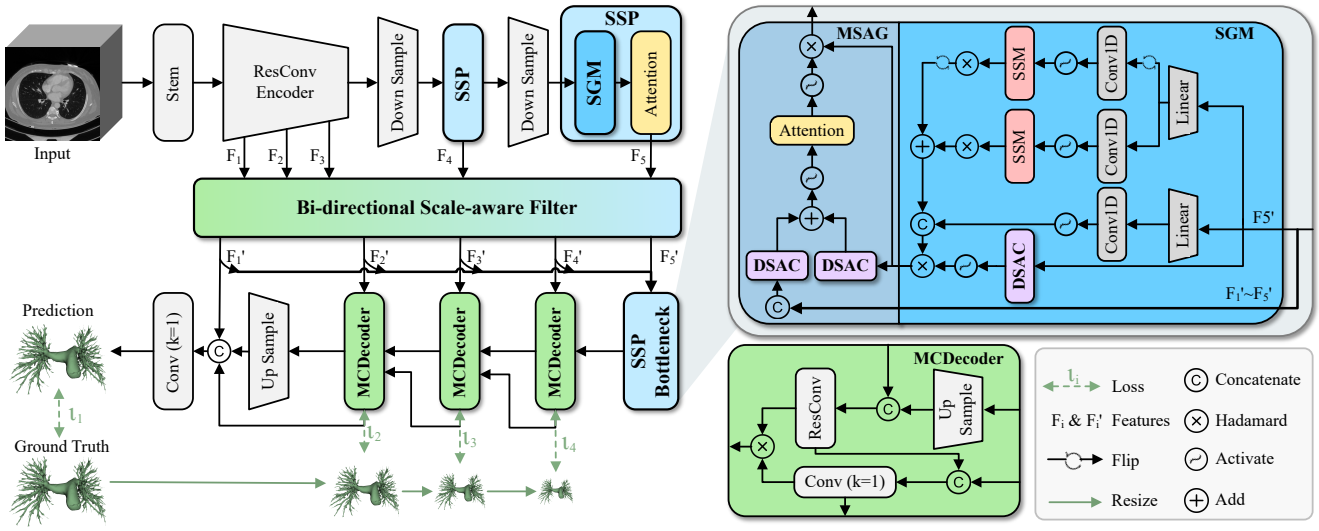

Figure 2. Overview of VesMamba, which primarily consists of two innovative modules: spatial-gated structural perception (SSP) module and bidirectional scale-aware filter (BSF) module. SSP is composed of a spatial gate mamba (SGM) and a self-attention mechanism, enabling efficient feature extraction with spatial perception. BSF fuses encoder features of each layer and filters out noise at different scales from the encoder features to obtain more robust feature representations. Additionally, a mask-constrained decoder (MCDecoder) directly constrains the adjacent low-layer predictions inference by using contour and position information in the high-layer masks, further improving the segmentation consistency and accuracy.

which is able to model in three orthogonal directions. However, such a triple-scan mechanism significantly increases the computational overhead, and it is difficult to segment small vessels due to insufficient local feature extraction.

Based on Mamba, some efficient 3D medical segmentation models [15, 36] are proposed. However, standard Mamba cannot meet the spatial perception requirements for visual tasks [42]. Although some studies design various multidirectional scans of Mamba [17, 19, 28] to address this issue, stacking scanning strategies may increase computational overhead with suboptimal performance improvements [7, 39, 43]. In contrast, we propose a dynamic spatial attention convolution, endowing Mamba with powerful 3D spatial perception capabilities.

3. Method

3.1. Overview

Figure 2 presents an overview of the VesMamba architecture, which mainly consists of four parts: the encoder, the bidirectional scale-aware filter (BSF) module, the spatial-gated structural perception (SSP) bottleneck, and the decoder. Specifically, given an input $I \in \mathbb{R}^{1\times H\times W\times D}$, a stem layer [30] extracts preliminary features $F_0 \in \mathbb{R}^{32\times H\times W\times D}$. The five-layer encoder processes F_0 through successive blocks to extract multi-scale encoder features $F_i \in \mathbb{R}^{(32\cdot 2^i)\times \frac{H}{2^i}\times \frac{W}{2^i}\times \frac{D}{2^i}}$ for $i \in 1, ..., 5$. The first three encoder layers employ residual convolutions to extract low-layer features, while the last two layers use the SSP module to extract high-layer features. SSP consists of a spatial gate mamba (SGM) block and a self-attention [34] block, where the former enables efficient feature extraction with dynamic spatial perception and the latter further captures long-range dependencies [8]. In addition, BSF bidirectionally fuses encoder features to aggregate both local and global information. The fused features highlight vessel structures by filtering noise at different scales in F_i, and the obtained enhanced encoder features are denoted as F_i'. In the SSP bottleneck, the self-attention is replaced by the multi-scale spatial attention gate (MSAG), which guides the model to focus more on vessels of different scales. Finally, F_i' and the high-level features of the SSP bottleneck are fed into the decoder, where high-layer masks containing vessel contours and position information are used to directly constrain adjacent low-layer prediction inference, further improving segmentation robustness and consistency.

3.2. Spatial-gated Structural Perception Module

For deploying Mamba on 3D vessel segmentation, 3D data should be converted into sequences [21], which may destroy spatial locality, leading to the absence of spatial perception. To address this, we propose the spatial-gated structural perception (SSP) module. As shown in Figure 2, SSP comprises a spatial gate mamba (SGM) and a self-attention block. SGM has a top-down architecture with upper, middle, and lower layers according to the input branches. As-

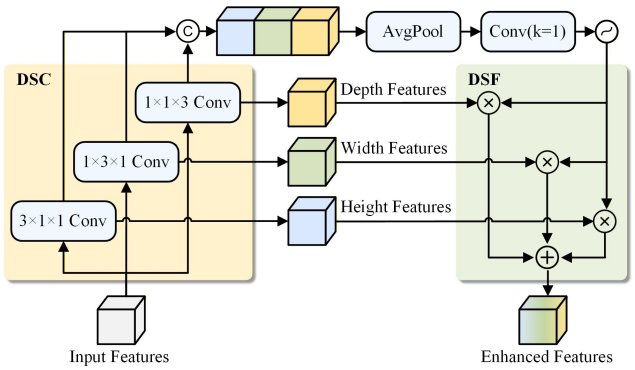

Figure 3. Illustration of DSAC, which consists of directional separation convolution (DSC) and dynamic spatial fusion (DSF).

suming C input channels, the upper and middle layers each receive $\frac{C}{2}$ channels, while the lower layer receives all C channels. The upper layer employs the state space model (SSM) with a bidirectional scanning strategy to comprehensively model long-range dependencies, where the obtained bidirectional features are then aligned and added together. In addition, the middle layer uses convolution to extract local features, which are combined with the upper layer output to aggregate both local and global information, hence highlighting blood vessels of different scales. We use SMM and $\mathrm{SSM_b}$ to denote the two branches of SSM in the upper layer, $\mathcal{M}_i$ to represent the operation of 1D convolution and Silu activation, and Flip to indicate the reverse operation on sequence. The corresponding output of the upper and middle layers s_1 can be formulated as:

$$\begin{aligned} x'_{1b}, x'_1, x'_2 &= \mathcal{M}_1(\mathrm{Flip}(x_1)), \mathcal{M}_2(x_1), \mathcal{M}_3(x_2), \\ s_1 &= \mathrm{Concat}(\mathrm{Flip}(\mathrm{SSM}_b(x'_{1b})) + \mathrm{SSM}(x'_1), x'_2) \end{aligned} \tag{1}$$

where x_1 and x_2 are the $\frac{C}{2}$ inputs after linear projection. x'_{1b} is the reversed sequence.

Due to spatial anisotropy in 3D vessel data, targets exhibit long-strip morphology parallel to vessels and elliptical morphology perpendicular to vessels. Traditional convolutions perform the same operation in each spatial direction, which fails to focus on the anisotropic features. Therefore, at the bottom layer of SGM, we propose the dynamic spatial attention convolution (DSAC). As shown in Figure 3, we first use direction-separated convolutions to extract three spatial anisotropic features, each with C channels. These features are concatenated into a single $3C$ feature. Then, the space of the concatenated features is compressed into a shape of $3C \times 1 \times 1 \times 1$ by an average pooling operation. Subsequent convolution reduces the number of channels to 3, followed by a Softmax function to generate directional weights. Finally, the spatial anisotropic features are fused according to the directional weights to obtain the output s_2.

$$\begin{aligned} x, y, z &= \mathrm{Conv}_1(X), \mathrm{Conv}_2(X), \mathrm{Conv}_3(X), \\ w_x, w_y, w_z &= \sigma(\mathrm{Conv}_4(\mathrm{Pool}(\mathrm{Concat}(x, y, z)))), \\ s_2 &= w_x \cdot x + w_y \cdot y + w_z \cdot z \end{aligned} \tag{2}$$

where $\{\mathrm{Conv_i}, i = 1, 2, 3, 4\}$ represents convolutions with kernel sizes of $1\times1\times3$, $1\times3\times1$, $3\times1\times1$, and $1\times1\times1$, respectively. w_x, w_y, and w_z are the directional weights, Pool represents average pooling, and σ is the Softmax function. DSAC effectively characterizes both vessel distribution and spatial anisotropy. Furthermore, the concatenated features are processed by the bottom layer using a gating mechanism, endowing SGM with the powerful spatial perception capability. The final output of SGM s_3 is defined as:

$$s_3 = s_1 \odot \varepsilon(s_2) \tag{3}$$

where $\odot$ represents the Hadamard product, and ε is the Sigmoid function. Finally, a self-attention block is applied to enhance the long-range dependency modeling of SGM, which is expressed as:

$$s_4 = \mathrm{Attn}(s_3) \tag{4}$$

where Attn is the self-attention block. For the SSP bottleneck, the self-attention is replaced with a multi-scale spatial attention gate (MSAG), where enhanced encoder features are used to further constrain the output of the high-layer SGM, enabling the model to focus more on vessel features at different scales. The output of MSAG s_5 can be formulated as:

$$\begin{aligned} F &= \mathrm{Concat}(\mathrm{Resize}(F'_1 \sim F'_5)), \\ F' &= \delta(\mathrm{DSAC}(F) + \mathrm{DSAC}(s_3)), \\ s_5 &= s_3 \odot \varepsilon(\mathrm{Attn}(F')) \end{aligned} \tag{5}$$

where $F'_1 \sim F'_5$ represent enhanced encoder features. δ and ε are ReLU and Sigmoid functions, respectively. Resize operation aligns the size of $F'_1 \sim F'_5$ with s_3.

3.3. Bidirectional Scale-aware Filter Module

Fusing vessel details and contour information contained in multi-layer encoder features is able to further improve model performance on 3D pulmonary vessel segmentation tasks. Common methods include feature pyramid networks [16] and Path Aggregation Networks (PAN) [18]. However, if low-layer and high-layer features are extracted in different manners, it may induce significant feature distribution gaps, and directly fusing these features may cause mutual interference [9]. In addition, the excessive convolutions in PAN cost large computational resources. Therefore, we propose a bidirectional scale-aware filter (BSF) module.

As shown in Figure 4, BSF adopts a bidirectional fusion approach to ensure that the fused features contain both local and global information. First, the high-layer semantic

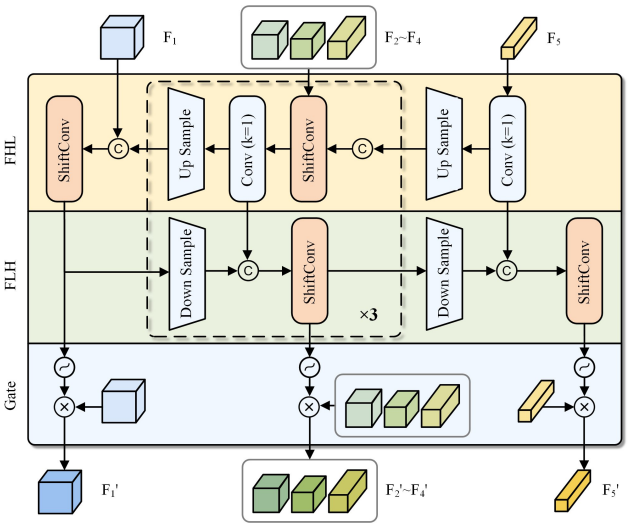

Figure 4. Illustration of BSF. ShiftConv is Depth-shift convolution. FHL denotes high-to-low layer fusion, while FLH is low-to-high layer fusion.

encoder features containing vessel contour information in F_i are transmitted layer by layer to the low layers. After obtaining the intermediate fusion features, the vessel detail information is transmitted layer by layer to the high layers. Meanwhile, in order to reduce mutual interference between features and ensure information transmission, we concatenate the features instead of element-wise addition in the fusion process. Moreover, we employ depth-shifted convolution (ShiftConv) [37], in which feature interaction in the depth direction is only performed through data movement. This allows for appropriately extracting 3D data features with the efficiency of 2D convolution. The fusion methods can be formulated as follows:

$$\begin{aligned} \mathcal{F}^{HL}(F_i, M_{i+1}) &= \text{Conv}(\mathcal{S}(F_i, \text{Up}(M_{i+1}))), \\ \mathcal{F}^{LH}(M'_{i-1}, M_i) &= \mathcal{S}(M_i, \text{Down}(M'_{i-1})) \end{aligned} \tag{6}$$

where $\mathcal{F}^{HL}$ represents fusion from high layer to low layer, and $\mathcal{F}^{LH}$ represents fusion from low layer to high layer. $\mathcal{S}$ represents fusion using ShiftConv after concatenation. F_i represents the encoder features. M_i and M'_i represent the intermediate and final fusion features, respectively. Conv represents convolution with a kernel size of $1 \times 1 \times 1$. Up and Down represent upsampling and downsampling, respectively. The feature fusion process can be expressed as:

$$\begin{aligned} M_1, M_5 &= \mathcal{S}(F_i, \text{Up}(M_2)), \text{Conv}(F_5), \\ M_2 \sim M_4 &= \mathcal{F}^{HL}(F_2 \sim F_4, M_3 \sim M_5), \\ M'_1, M'_2 \sim M'_5 &= M_1, \mathcal{F}^{LH}(M'_1 \sim M'_4, M_5 \sim M_2) \end{aligned} \tag{7}$$

The final fusion features M'_i contain both local and global information, while the low-layer and high-layer decoders are mainly responsible for extracting vessel details and vessel contour information [16], respectively. If the fusion features are directly used for decoding, local information may interfere with high-layer decoding processes, while global information could disrupt low-layer encoders [1, 26]. Therefore, we employ a gating mechanism to filter out noise within the encoder features F_i at different scales, to further highlight vessel information and enhance the feature representations. This process can be defined as:

$$F'_i = \varepsilon(M'_i) \odot F_i \tag{8}$$

where F'_i represents the enhanced encoder features, ε is the Sigmoid function, and $\odot$ represents the Hadamard product.

3.4. Mask-constrained Decoder

As shown in Figure 2, to further enhance the guiding role of deep supervision on the model, we propose a Mask-constrained Decoder (MCDecoder). The high-layer masks contain both vessel contour and position information, while the low-layer masks are rich in local details. As the differences between adjacent masks are relatively small, we use the high-layer masks as the input for the adjacent low-layer decoder, providing the low-layer decoder with the contour and position information of the vessel. This directly constrains the prediction of the low-layer masks, further improving segmentation consistency and accuracy. The decoding process can be formulated as follows:

$$\begin{aligned} T_i &= \text{ResConv}(\text{Concat}(F'_i, \text{Up}(T'_{i+1}))), \\ Mask_i &= \text{Conv}(\text{Concat}(Mask_{i+1}, T_i)), \\ T'_i &= T_i \odot Mask_i, \\ Mask_1 &= \text{Conv}(\text{Concat}(F'_1, \text{Up}(T'_2), Mask_2)) \end{aligned} \tag{9}$$

where $\{T_i, i = 2, 3, 4\}$ denotes the intermediate features of the decoder, T'_i represents the feature outputs, and F'_i denote the enhanced encoder features. ResConv is the residual convolution. $Mask_i$ are the mask outputs for deep supervision. Up stands for the upsampling operation. $Mask_1$ is the final segmentation result.

3.5. Loss Functions

For deep supervision, we use binary cross-entropy (BCE) loss and Dice loss to calculate the stage loss, then weight and aggregate all stage losses to calculate the total loss:

$$\mathcal{L} = \sum_{i=1}^{4} k_i \cdot (\mathcal{L}_{BCE} + \mathcal{L}_{Dice}) \tag{10}$$

where $\mathcal{L}_{BCE}$ and $\mathcal{L}_{Dice}$ are BCE loss and Dice loss, respectively. k_i represents the weight of different layers, which is obtained by normalizing the ratio of the mask size to the input size at each stage.

Table 1. Quantitative comparisons with different state-of-the-art methods on the Lung79 dataset, which includes three categories of airway, artery, and vein.

Method	Airway				Artery				Vein			
	Dice↑	clDice↑	HD95↓	NSD↑	Dice↑	clDice↑	HD95↓	NSD↑	Dice↑	clDice↑	HD95↓	NSD↑
3D-UNet	84.07	90.49	3.47	93.29	83.87	80.70	8.05	87.61	84.11	83.21	6.82	87.55
nnUNet	84.15	91.23	3.50	93.61	84.27	82.00	7.80	88.41	85.07	83.64	6.17	88.55
DSCNet	77.62	75.42	25.39	85.11	77.09	69.22	12.71	77.88	78.23	72.52	25.63	78.23
SegFormer3D	76.47	70.66	10.08	80.42	80.27	70.89	10.41	82.16	81.56	74.45	8.71	82.78
UNETR++	82.04	88.31	4.02	91.62	82.12	78.60	8.97	86.83	83.62	81.79	6.73	87.29
Swin-UNETR	81.12	86.05	3.88	91.60	80.05	71.26	10.96	81.46	81.11	74.90	9.67	82.13
UMamba	84.63	90.80	3.34	93.85	83.96	81.33	7.68	88.78	84.89	83.61	6.01	88.72
SegMamba	83.38	89.90	3.32	92.89	83.45	80.00	8.31	87.18	84.16	82.01	6.98	87.44
LKM-UNet	84.28	91.18	3.94	93.63	84.74	82.13	7.23	88.77	84.91	83.56	6.29	88.61
Ours	**84.85**	**91.46**	**3.10**	**94.08**	**84.96**	**82.44**	**7.01**	**89.03**	**85.76**	**84.24**	**5.71**	**88.86**

Table 2. Quantitative comparisons with different state-of-the-art methods on the Parse22 dataset.

Method	Parse22			
	Dice↑	clDice↑	HD95↓	NSD↑
3D-UNet	85.25	82.93	4.96	90.65
nnUNet	84.20	80.85	6.21	89.11
DSCNet	76.63	63.92	12.27	76.19
SegFormer3D	77.69	60.06	10.70	78.45
UNETR++	85.27	84.02	4.61	91.36
Swin-UNETR	79.96	67.18	10.78	81.08
UMamba	85.94	86.64	3.69	92.74
SegMamba	85.23	84.21	4.99	91.32
LKM-UNet	86.21	87.14	3.31	92.95
Ours	**86.65**	**87.46**	**2.98**	**93.21**

4. Experiment

4.1. Datasets and Evaluation Metrics

To evaluate the performance of our model, we select two 3D pulmonary CT datasets:

- **Parse22:** A public dataset for segmenting complex pulmonary arteries from CT images. Image sizes range from $512 \times 512 \times 228$ to $512 \times 512 \times 376$, with a total of 200 cases. This study uses 100 of the publicly available cases. The segmentation targets vary in size.
- **Lung79:** An in-house dataset for segmenting three complex structures (arteries, veins, and airways) from CT images. Image sizes range from $512 \times 512 \times 209$ to $512 \times 512 \times 657$, with a total of 79 cases. The segmentation targets are multiple and highly similar to each other.

This study uses four metrics: Dice, clDice [29], HD95 [10], and NSD [22] to fully evaluate model performance.

4.2. Implementation Details

We implement VesMamba based on the UMamba framework, and all experiments are conducted on a single NVIDIA A100 PCIe 40GB GPU. The SGD optimizer is used, with an initial learning rate set to 1e-2. The PolyLRScheduler is employed as the scheduler, and the network is trained for 200 epochs. The batch size is 2. The Parse22 and Lung79 datasets are split into training, validation, and test sets in the ratios of 64:16:20 and 55:8:16, respectively.

4.3. Comparison with State-of-the-art Methods

We compare our method with nine state-of-the-art 3D medical segmentation competitors on the Parse22 and Lung79 datasets. The comparison methods can be categorized into three types: convolutional-based methods including 3D-UNet [2], nnUNet [11], and DSCNet [25]; Transformer-based methods including SegFormer3D [24], UNETR++ [27], Swin-UNETR [32]; and Mamba-based methods including UMamba [21], SegMamba [40], and LKM-UNet [35]. Table 2 shows the quantitative comparison results between our method and the above methods on the Parse22 dataset. The outstanding performance of the Mamba-based approach demonstrates the superiority of the Mamba architecture for this task. Moreover, our method outperforms other Mamba-based methods across all metrics. Compared to the baseline model UMamba, our method improves the Dice and clDice scores by 0.71% and 0.82%, respectively, on the Parse22 dataset, demonstrating that our model learns better the overall distribution of vessels while ensuring vessel continuity. The improvements in HD95 and

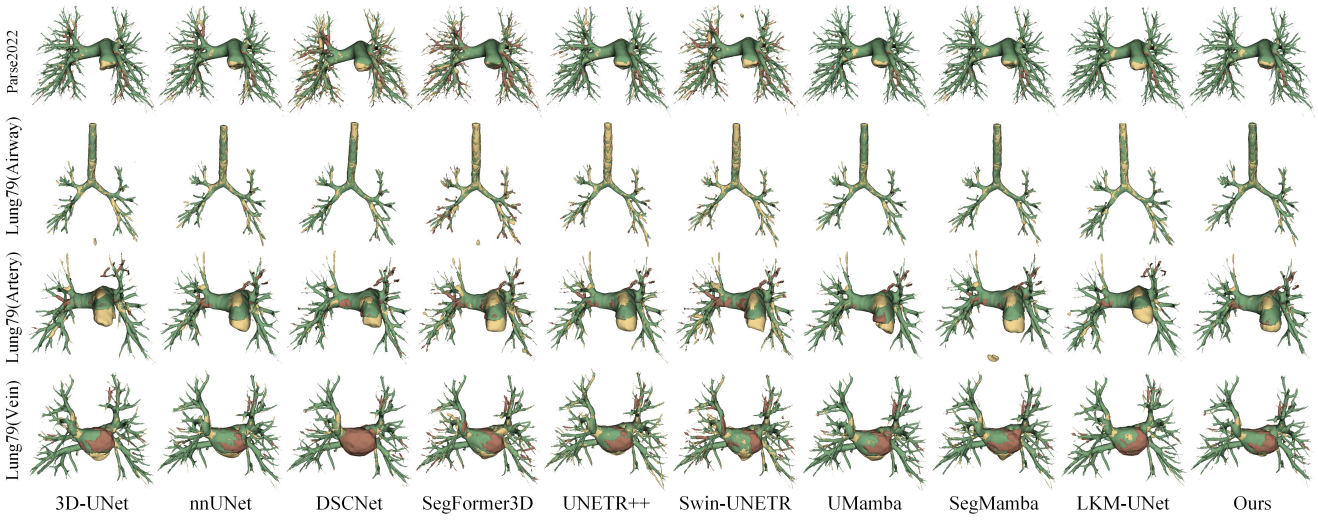

Figure 5. Visual comparison of challenging cases with state-of-the-art methods on the Parse22 and Lung79 datasets. Green, yellow, and red represent true positives, false positives, and false negatives, respectively.

Table 3. Ablation studies of SSP, BSF and MCDecoder on the Parse22 and Lung79 datasets, comparing Dice scores.

SSP	BSF	MCDecoder	Parse22	Lung79
			85.94	84.49
✓			86.42	84.95
	✓		86.35	84.87
✓	✓		86.55	85.06
✓	✓	✓	**86.65**	**85.19**

NSD indicate that our model exhibits superior spatial similarity while focusing more on boundary information. In addition, Table 1 presents the quantitative comparison results on the in-house dataset Lung79. On this more challenging segmentation task, our model still achieves excellent performance across three segmentation targets compared to other methods, fully demonstrating the robustness of our method in 3D vessel segmentation tasks.

Figure 5 presents the comparison results between our method and other state-of-the-art methods on Parse22 and Lung79. This intuitively demonstrates the robustness of our method in segmenting targets of different scales. Compared with other methods, our method yields significantly fewer missegmented and under-segmented parts, while the continuity of small branches is also well preserved.

4.4. Ablation Studies

We conduct all ablation experiments on the Parse22 and Lung79 datasets. As shown in Table 3, adding the SSP module results in the greatest performance improvement on both datasets, indicating that SSP effectively learns vessel features. As shown in Figure 6, compared to the baseline, the model using SSP is able to focus more on the vessel regions, significantly increasing attention to large-scale vessels while alleviating interference from non-vessel areas. Next, after adding the BSF module, the Dice scores of both datasets improve steadily, indicating the enhancing effect of BSF on encoder features. As shown in Figure 7, the model combined with BSF better segments vessels of different scales, with a significant reduction in false positive and false negative regions compared to the baseline. In addition, after adding both the SSP and BSF modules, the model performance exceeds that of a single module. Finally, when further combined with MCDecoder, the model achieves optimal performance. These ablation studies demonstrate that the proposed modules cooperate effectively, allowing the model to possess robust spatial perception capability, efficiently learn the 3D vessel distribution, and significantly enhance vessel features.

Additionally, to further analyze the effectiveness of the intra-module designs, we conduct detailed ablation experiments on the Parse22 and Lung79 datasets for SSP, BSF, and MCDecoder modules respectively. Table 4 shows the ablation studies results. Conv_1 represents replacing DSAC in all SSP modules with convolution, and Atten means replacing MSAG at the SSP bottleneck with a self-attention mechanism. Add represents using element-wise addition to replace the connection operation in BSF, Conv_2 represents using convolution to replace ShiftConv in BSF, and w/o gate denotes without the gating mechanism in BSF, directly using the fusion features as output. constraint_1 denotes constraints on the current decoder features output, while constraint_2 represents constraints on the position and

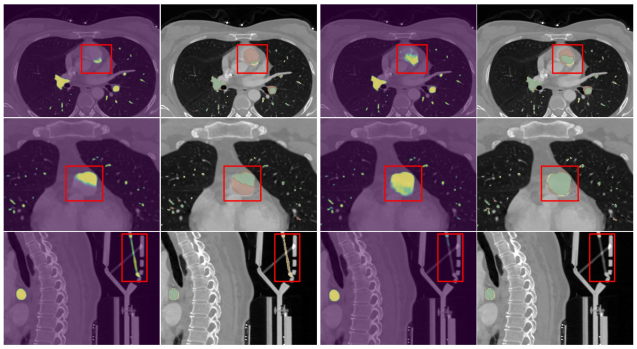

Figure 6. Visual comparison of the results of the SSP module ablation study on Parse22 dataset. Green, yellow, and red represent true positives, false positives, and false negatives, respectively.

Table 4. Ablation studies of the design of SSP (1), BSF (2), and MCDecoder (3) modules on the Parse22 and Lung79 datasets.

	Method	Parse22		Lung79	
		Dice↑	clDice↑	Dice↑	clDice↑
(1)	$Conv_1$ + Atten	86.25	86.89	84.75	85.40
	DSAC + Atten	86.38	87.15	84.81	85.66
	$Conv_1$ + MSAG	86.32	87.09	84.76	85.53
	SSP	**86.42**	**87.38**	**84.95**	**85.78**
(2)	Add + $Conv_2$	84.32	84.66	82.98	83.87
	Concat + $Conv_2$	85.75	87.11	84.42	85.32
	Add + ShiftConv	86.06	86.97	84.78	85.68
	w/o gate	85.84	86.48	84.38	85.62
	BSF	**86.35**	**87.26**	**84.87**	**85.89**
(3)	w/o $constraint_1$	86.24	87.11	84.67	85.48
	w/o $constraint_2$	86.20	87.07	84.74	85.51
	MCDecoder	**86.29**	**87.18**	**84.76**	**85.64**

contour information of the adjacent shallow decoder. According to these results, the designs in these modules collectively enable SSP, BSF and MCDecoder modules to achieve optimal performance.

4.5. Discussions and Limitations

Figure 8 shows that our method achieves a balance between performance and efficiency, where LKM-UNet performs similarly to our method, but our method requires only about a quarter of the computational effort. In addition, our method still has some limitations, as shown in Figure 9. Compared to the normal segmentation target (a), segmentation targets with excessive vascular branching (b-c) and abnormal and uneven shapes (d-e) may limit our method.

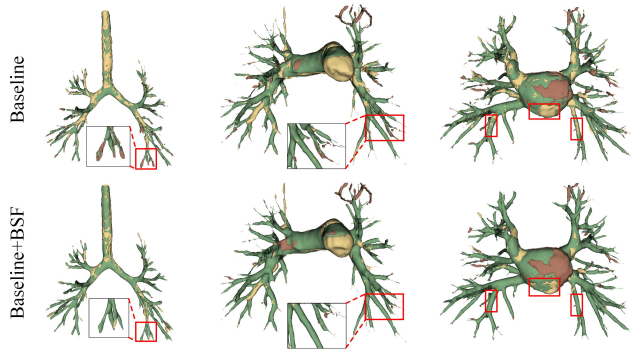

Figure 7. Visual comparison of the results of the BSF module ablation study on Lung79 dataset. Green, yellow, and red represent true positives, false positives, and false negatives, respectively.

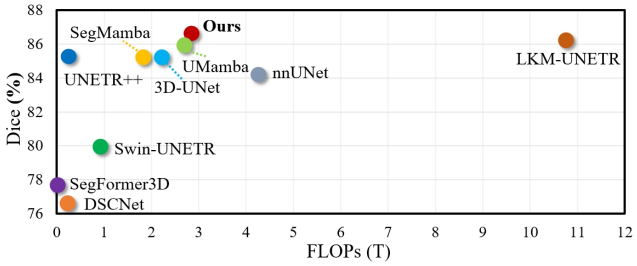

Figure 8. Performance-efficiency comparison with other state-of-the-art methods on Parse22.

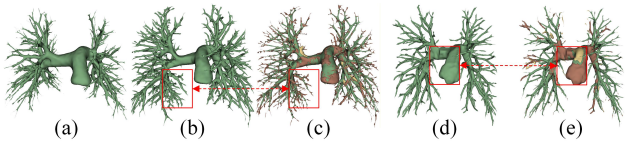

Figure 9. Failure cases. Green, yellow, and red represent true positives, false positives, and false negatives, respectively.

5. Conclusion

We propose a novel 3D pulmonary vessel segmentation model in this paper, named VesMamba, which primarily comprises two innovative modules. Firstly, we propose an SSP module, which combines SGM and self-attention. In SGM, DSAC dynamically learns the 3D vessel distribution, endowing SGM with powerful spatial perception capabilities. The self-attention is then used to further enhance SSP's ability to capture long-range dependencies. Additionally, we introduce a BSF module, which uses bidirectionally fused features to filter out noise of different scales in the encoder features, thereby increasing the focus on vessels. Finally, we propose a MCDecoder to ensure the accuracy and consistency of predictions. Benefiting from these modules, VesMamba achieves outstanding performance in experiments on both public and in-house datasets.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (No. 62273241), Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), the Shenzhen Research Foundation for Basic Research, China (No. JCYJ20250604181940054), and the grant of Innovation and Technology Fund under Innovation and Technology Support Programme (project no ITS/202/23).

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432, Cham, 2016. Springer International Publishing. 6

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[4] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 10041–10071. PMLR, 2024. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[7] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28124–28133, 2025. 3

[8] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25261–25270, 2025. 3

[9] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5557–5566, 2023. 4

[10] Daniel P. Huttenlocher, Gregory A. Klanderman, and William Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 15:850–863, 1993. 6

[11] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 6

[12] Titinunt Kitrungrotsakul, Xian-Hua Han, Yutaro Iwamoto, Lanfen Lin, Amir Hossein Foruzan, Wei Xiong, and Yen-Wei Chen. Vesselnet: A deep convolutional neural network with multi pathways for robust hepatic vessel segmentation. *Computerized Medical Imaging and Graphics*, 75: 74–83, 2019. 2

[13] Bin Kong, Xin Wang, Junjie Bai, Yi Lu, Feng Gao, Kunlin Cao, Jun Xia, Qi Song, and Youbing Yin. Learning tree-structured representation for 3d coronary artery segmentation. *Computerized Medical Imaging and Graphics*, 80: 101688, 2020. 2

[14] Peter J Leary, Megan Lindstrom, Catherine O Johnson, et al. Global, regional, and national burden of pulmonary arterial hypertension, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Respiratory Medicine*, 13(1):69–79, 2025. 1

[15] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Chengwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation, 2024. 3

[16] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 5

[17] Hui Liu, Chen Jia, Fan Shi, Xu Cheng, and Shengyong Chen. Scsegamba: Lightweight structure-aware vision mamba for crack segmentation in structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29406–29416, 2025. 3

[18] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[19] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 103031–103063. Curran Associates, Inc., 2024. 2, 3

[20] Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, Pengxin Yu, Enyou Liu, Hongrong Wei, Na Wang, Jia Guo, Huiqi Li, Zhao Zhang, Ziwei Zhao, Na Gao, Nan An, Ashkan Pakzad, Bojidar Rangelov, Jiaqi Dou, Song Tian, Zeyu Liu, Yi Wang, Ampatishan Sivalingam, Kumaradevan Punithakumar, Zhaowen Qiu, and Xin Gao. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge, 2024. 2

[21] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation, 2024. 2, 3, 6

[22] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Hughes, Joseph R. Ledsam, and Olaf Ronneberger. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy, 2021. 6

[23] M. Orkisz, M. Hernández Hoyos, V. Pérez Romanello, C. Pérez Romanello, J.C. Prieto, and C. Revol-Muller. Segmentation of the pulmonary vascular trees in 3d ct images using variational region-growing. *Innovation and Research in BioMedical engineering (IRBM)*, 35(1):11–19, 2014. 1, 2

[24] Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: An efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4981–4988, 2024. 6

[25] Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6070–6079, 2023. 2, 6

[26] 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 (MICCAI)*, pages 234–241. Springer, 2015. 5

[27] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: Delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 43(9):3377–3390, 2024. 6

[28] Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. Groupmamba: Efficient group-based visual state space model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14912–14922, 2025. 3

[29] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. cldice - a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16560–16569, 2021. 2, 6

[30] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 4277–4285, 2017. 3

[31] Wenjun Tan, Luyu Zhou, Xiaoshuo Li, Xiaoyu Yang, Yufei Chen, and Jinzhu Yang. Automated vessel segmentation in lung ct and cta images via deep neural networks. *Journal of X-ray science and technology*, 29(6):1123–1137, 2021. 2

[32] Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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 (CVPR)*, pages 20730–20740, 2022. 6

[33] Evelien Van Dongen and Bram van Ginneken. Automatic segmentation of pulmonary vasculature in thoracic ct scans with local thresholding and airway wall removal. In *2010 IEEE international symposium on biomedical imaging: From nano to macro*, pages 668–671. IEEE, 2010. 1, 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2017. 3

[35] Jinhong Wang, Jintai Chen, Danny Chen, and Jian Wu. Lkm-unet: Large kernel vision mamba unet for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 360–370. Springer, 2024. 2, 6

[36] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024. 3

[37] Boqian Wu, Qiao Xiao, Shiwei Liu, Lu Yin, Mykola Pechenizkiy, Decebal Constantin Mocanu, Maurice van Keulen, and Elena Mocanu. E2enet: Dynamic sparse feature fusion for accurate and efficient 3d medical image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 118483–118512. Curran Associates, Inc., 2024. 5

[38] Yanan Wu, Shouliang Qi, Meihuan Wang, Shuiqing Zhao, Haowen Pang, Jiaxuan Xu, Long Bai, and Hongliang Ren. Transformer-based 3d u-net for pulmonary vessel segmentation and artery-vein separation from ct images. *Medical & Biological Engineering & Computing*, 61(10):2649–2663, 2023. 2

[39] Chaodong Xiao, Minghan Li, zhengqiang ZHANG, Deyu Meng, and Lei Zhang. Spatial-mamba: Effective visual state space models via structure-aware state fusion. In *International Conference on Representation Learning (ICLR)*, pages 73777–73795, 2025. 3

[40] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 578–588, Cham, 2024. Springer Nature Switzerland. 2, 6

[41] Yajun Xu, Zhendong Mao, Chunxiao Liu, and Bin Wang. Pulmonary vessel segmentation via stage-wise convolutional networks with orientation-based region growing optimization. *IEEE Access*, 6:71296–71305, 2018. 1, 2

[42] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 2, 3

[43] Qinfeng Zhu, Yuan Fang, Yuanzhi Cai, Cheng Chen, and Lei Fan. Rethinking scanning strategies with vision mamba in semantic segmentation of remote sensing imagery: An experimental study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:18223–18234, 2024. 3