UMC-FM: Uncertainty-Guided Mamba-Consistent Feature Mixing for Semi-supervised Pulmonary Vessel Segmentation

Zhiqing Liang, Faizan Ahmad, Haomin Liang, Inayatul Haq, Liyilei Su, Rui Li(✉), and Bingding Huang(✉)

School of Artificial Intelligence, Shenzhen Technology University, Shenzhen 518118, China
{lirui,huangbingding}@sztu.edu.cn

Abstract. Accurate segmentation of 3D pulmonary vessels is critical for the diagnosis and surgical planning of vascular diseases. However, this remains challenging due to the scarcity of voxel-wise annotations and the complex, disconnected topology of distal vessels. Existing semi-supervised learning (SSL) methods, particularly mean-teacher frameworks, often suffer from confirmation bias and struggle to capture long-range vascular dependencies. To address these limitations, we propose a novel semi-supervised framework, uncertainty-guided mamba consistent feature mixing (UMC-FM). We introduce a hybrid architecture that embeds the visual state space (VSS) block into a mamba-based encoder-decoder network. This design enables the modeling of long-range vascular dependencies with linear complexity, successfully circumventing the inherent receptive field constraints of CNNs while avoiding the computational overhead of transformers. Furthermore, we introduce a dual-consistency strategy: (1) an uncertainty-guided region sampling that leverages teacher uncertainty to mix hard samples from unlabeled data actively, and (2) a manifold feature consistency loss that aligns the latent representations of mixed samples to learn domain-invariant vascular topology. Extensive experiments on the HiPaS and Parse datasets demonstrate that UMC-FM achieves state-of-the-art performance on both benchmarks. Our method significantly outperforms fully supervised baselines and transformer-based SSL methods, offering a robust and efficient solution for label-scarce volumetric pulmonary vessel segmentation.

Keywords: Pulmonary vessel segmentation · semi-supervised learning · uncertainty estimation · mamba

1 Introduction

Significant research has focused on addressing these challenges. To overcome the scarcity of labeled data, semi-supervised learning (SSL) has gained prominence for its ability to utilize large-scale unlabeled datasets. Among various SSL techniques [2], the mean teacher method, which has been extended in recent semi-supervised segmentation

D.-S. Huang et al. (Eds.): ICIC 2026, LNCS 16658, pp. 1–12, 2027.
https://doi.org/10.1007/978-981-92-3501-8_29

studies [3, 4]. Nevertheless, despite their effectiveness, these methods did not generalize well to vessel segmentation. This limitation primarily results from cognitive bias introduced by static supervision, a challenge intensified by the complexity of the vessel tree. Specifically, (i) pseudo-labels generated by a static teacher model are not consistently more accurate than those produced by the student model, which can misguide the student learning process; and (ii) if the student network acquires erroneous representations during early training, these errors may be transferred to the teacher model through the exponential moving average (EMA). With advances in training, inaccuracies are amplified within the mean-teacher framework, leading to segmentation results that deviate from the ground truth.

Existing approaches mainly include contrastive learning [5, 6], pseudo-labeling [7, 8], and consistency regularization [9–11]. Contrastive learning and pseudo-labeling methods often rely on a complex network to construct training pairs to refine pseudo-labels. In contrast, consistency regularization offers a comparatively simple learning framework and has demonstrated stable performance across diverse segmentation tasks. Consistency regularization methodologies can be broadly classified into two categories: single-model and mean-teacher networks. This optimizes a network by imposing consistency between its predictions on original and perturbed unlabeled samples [10, 12].

Due to the complex nature of 3D vessels, a number of efforts have been made to achieve a robust feature representation. For instance, DUNet [13] integrates deformable convolution [14] into a U-shaped architecture, enabling adaptive adjustment of the receptive field to accommodate variations in vessel scale and morphology. DSCNet [15] further extends deformable convolution by introducing a dynamic convolution guided by topological geometric constraints, which effectively captures the properties of tubular structures. MICNet [16] employs dense, hybrid, dilated convolution [17] within the connection layers to capture richer contextual information while preserving feature resolution. The state-of-the-art (SOTA) semi-supervised 3D vessel segmentation framework DiCo [1] introduces a dynamic collaborative network to alleviate cognitive bias in mean-teacher frameworks.

To address these limitations, we propose a novel semi-supervised framework, an uncertainty-guided mamba consistent feature mixing (UMC-FM) as shown in Fig. 1. Unlike standard consistency methods that rely on static supervision, our approach actively mitigates cognitive bias by integrating uncertainty-driven perturbations in both the input and latent spaces. The main contributions of this study are as follows:

1. We introduce mamba-based, encoder-decoder framework that integrates a VSS bottleneck of encoder-decoder.
2. We propose a dual-stage consistency strategy. In the input, we employ an uncertainty-guided region sampling to actively sample hard regions from unlabeled data. In the latent space, we enforce manifold consistency to ensure that the student network learns robust domain invariant representations.
3. To reduce the risk of the student model overfitting to incorrect pseudo labels, we propose an uncertainty-guided mamba-consistent feature mixing. This uses the teacher model uncertainty to guide the mixing of labeled and unlabeled data, which helps the model better handle complex pulmonary vessel structures.

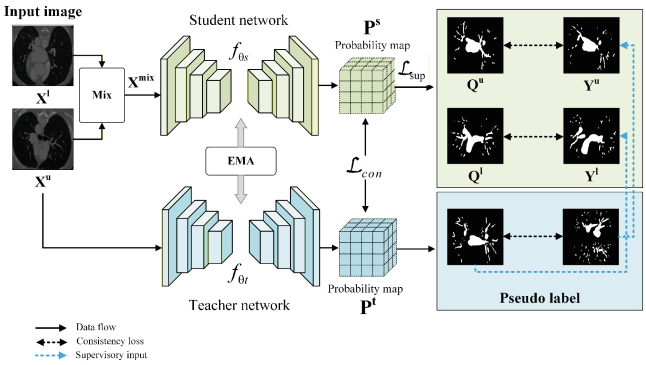


Fig. 1. Overview of the proposed semi-supervised segmentation framework. The model employs an uncertainty-guided mixing module to process labeled and unlabeled inputs.

2 Method

Our roposed approach follows a dual-phase consistency principle to address the challenge of disconnected vascular trees. Phase 1 utilizes uncertainty-guided region sampling for semi-supervised learning to extract complex features using a targeted entropy-based cropping method. Phase 2 enforces feature-manifold consistency within a VSS-based mamba bottleneck, ensuring that global vascular connectivity is preserved in the latent space. Finally, these phases collectively facilitate end-to-end training of the proposed UMC-FM network, as shown in Fig. 2.

Our base model employs a mean-teacher framework, adapted for semi-supervised pulmonary vessel segmentation. Specifically, the framework comprises a teacher network $\mathcal{F}_t(\mathbf{X}^u; \Theta_t)$ and a student network $\mathcal{F}_s\left(\mathbf{X}^{\text{mix}}; \Theta_s\right)$, both utilizing our proposed mamba-based encoder-decoder architecture to capture long-range vascular dependencies. The teacher and Student networks are optimized using exponential moving average (EMA) and stochastic gradient descent (SGD), respectively. This dual-stage network strategy enhances model performance and ensures training stability in the semi-supervised setting. The training procedure consists of three distinct steps. First, we pre-train the UMC-FM explicitly on labeled data $\mathcal{D}^l$ to initialize the weights and stabilize the mamba bottleneck. Second, the pre-trained model initializes the teacher network, which then generates voxel-wise entropy maps and pseudo labels for the unlabeled data $\mathcal{D}^u$. These steps guide the uncertainty-guided region sampling in stage 1 and manifold consistency stage 2 processes. During each iteration, the student network parameters Θ_s are optimized using SGD to minimize the total loss $\mathcal{L}_{\text{total}}$, which combines supervised, unsupervised, and feature consistency loss. Finally, the teacher network parameters Θ_t are updated using the EMA of the student network weights to maintain a stable, temporally ensembled vessels target.

2.1 Uncertainty-Guided Region Sampling

Typical semi-supervised learning methods rely on random cropping; however, this strategy is suboptimal for sparse anatomical structures such as pulmonary vessels. To address

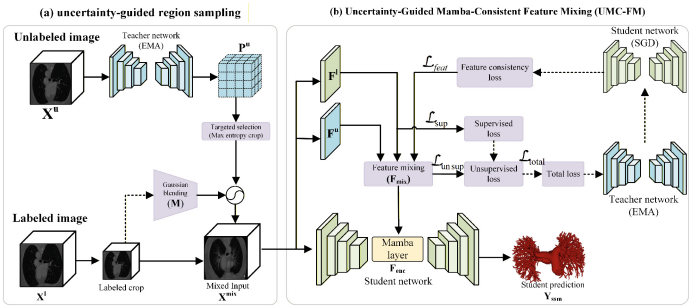


Fig. 2. Overview of the proposed uncertainty-guided mamba consistent feature mixing (UMC-FM) framework. The framework consists of two main components: (a) uncertainty- guided region sampling and (b) Mamba-consistent feature mixing.

this limitation, stage 1 introduces a targeted sampling strategy that estimates uncertainty from the teacher network to actively identify and prioritize anatomically ambiguous regions.

Voxel-Wise Entropy Mapping: Let $\mathbf{X}^u \in \mathbb{R}^{C_{in} \times W \times H \times D}$ be an unlabeled input volume. We perform a forward pass through the teacher network $\mathcal{F}_t$ to generate the probability map $\mathbf{P}^u$. We quantify the model epistemic uncertainty using the voxel-wise shannon entropy $\mathbf{E}$:

$$\mathbf{E}(\mathbf{v}) = -\sum_{c=1}^{C_{\text{out}}} \mathbf{P}_c^u(\mathbf{v}) \log \mathbf{P}_c^u(\mathbf{v}) \tag{1}$$

where $\mathbf{v}$ represents the voxel coordinate vector.

Targeted Centroid Optimization: We define a local sampling window S_{crop} to solve the optimal centroid vector $\mathbf{c}^*$ that maximizes the integral of entropy within the window, effectively sampling the rich anatomical features:

$$\mathbf{c}^* = \underset{\mathbf{c}}{\text{argmax}} \sum_{\mathbf{v} \in \Omega_{\mathbf{c}}} \mathbf{E}(\mathbf{v}) \tag{2}$$

where $\Omega_{\mathbf{c}}$ is the local region centered at $\mathbf{c}$.

Gaussian Soft-Blending: To integrate the sampled region without introducing high-frequency edge artifacts, we employ a Gaussian blending kernel M. The final mixed Input $\mathbf{X}^{\text{mix}}$ is synthesized by blending a labeled patch $\mathbf{X}^l_{\text{crop}}$ into the high-uncertainty map of $\mathbf{X}^u$:

$$\mathbf{X}^{\text{mix}} = \mathbf{X}^l_{\text{crop}} \odot \mathbf{M} + \mathbf{X}^u \odot (\mathbf{1} - \mathbf{M}) \tag{3}$$

The blending kernel $\mathbf{M}$ is parameterized by a 3D Gaussian filter with a kernel size of $K \times K \times K$ and a standard deviation of σ. In our implementation, we empirically set $K = 11$ and $\sigma = 5.0$.

2.2 Manifold Consistency

To overcome sequential-structure disruption from patch pasting and preserve coherent feature representation, we enforce latent-space linearity, thereby maintaining manifold feature consistency at the mamba VSS model bottleneck. We construct the target feature map $\mathbf{F}_{\text{mix}}$ representing the latent representation. Let $\phi_s(\cdot)$ be the student encoder mapping to the bottleneck dimensions. We mix the latent features of the labeled $\mathbf{F}^l$ and Unlabeled $\mathbf{F}^u$ images using the down-sampled mask $\mathscr{M}$:

$$\mathbf{F}_{\text{mix}} = \phi_s\left(\mathbf{X}^l\right) \odot \mathcal{M} + \phi_s\left(\mathbf{X}^u\right) \odot (\mathbf{1} - \mathcal{M}) \tag{4}$$

where $\mathbf{F}^l = \phi_t(\mathbf{X}^l)$, and $\mathbf{F}^u = \phi_t(\mathbf{X}^u)$ are the features of labeled and unlabeled images, respectively.

Furthermore, we pass the mixed image $\mathbf{X}^{\text{mix}}$ (from Stage 1) through the student network encoder ϕ_s to obtain $\mathbf{F}_{\text{student}} = \phi_s(\mathbf{X}^{\text{mix}})$. To ensure that the internal state of the mamba layer captures a linear combination of anatomical features instead of edge artifacts, we minimize the feature discrepancy loss as follows:

$$\mathcal{L}_{\text{fd}} = \|\mathbf{F}_{\text{student}} - \mathbf{F}_{\text{mix}}\|_2^2 \tag{5}$$

2.3 Global-Local Representation Learning Using Hybrid-Mamba

While CNNs excel at local texture extraction, they struggle with modeling global dependencies. We replace the deepest convolutional bottleneck with a mamba layer to model vascular topology with linear complexity. Let the encoder feature map at the bottleneck is $\mathbf{F}_{\text{enc}} \in \mathbb{R}^{C_b \times w' \times h' \times d'}$. We flatten the spatial dimensions to create a sequence $\mathbf{S} \in \mathbb{R}^{L \times C_b}$. The backbone adopts a symmetric encoder and a decoder architecture. The network takes as input a volumetric patch $\mathrm{X} \in \mathbb{R}^{C_{\text{in}} \times 128 \times 192 \times 192}$. The encoder comprises four downsampling stages implemented using $3 \times 3 \times 3$ convolutions and max-pooling, progressively extracting hierarchical features and producing a bottleneck representation $\mathbf{F}_{\text{enc}} \in \mathbb{R}^{512 \times 8 \times 12 \times 12}$. To enable sequence modeling with the mamba layer, the spatial dimensions of $\mathbf{F}_{\text{enc}}$ are flattened into a sequence of $L = 1{,}152$ tokens, forming $\mathbf{S}_{\text{in}} \in \mathbb{R}^{L \times 512}$. We placed the VSS block placed at the bottleneck, which operates on the flattened sequence S_{in}. The VSS block integrates a selective space model with a gated design to regulate information flow. Given the input sequence, we first apply Layer Normalization, $\tilde{\mathbf{S}} = \mathbf{LN}(\mathbf{S}_{\text{in}})$, followed by a dual-branch linear projection. Specifically, a gating branch produces $\mathbf{Z} = \sigma\left(\text{Linear}_z\left(\tilde{\mathbf{S}}\right)\right)$, where $\sigma(\cdot)$ denotes the SiLU activation, while the main branch generates $\mathbf{X}$= Linear $_x(\mathbf{S})$. To capture local continuity, the main branch is processed by a depth-wise 1D convolution with kernel size 3:

$$\mathbf{X}' = \text{Conv1D}(\mathbf{X}) \odot \sigma\left(\text{Conv1D}(\mathbf{X})\right) \tag{6}$$

The resulting sequence is then passed through the selective SSM to model long-range dependencies and produce global features $\mathbf{Y}_{\text{ssm}}$. Finally, the global features are modulated by the gating signal and projected back to the original feature dimension via a residual connection:

$$\mathbf{S}_{\text{out}} = \text{Linear}_{\text{out}}(\mathbf{Y}_{\text{ssm}} \odot \mathbf{Z}) + \mathbf{S}_{\text{in}} \tag{7}$$

The output sequence is reshaped back to $\mathbf{S}_{\text{out}} = \mathbb{R}^{C_b \times d' \times h' \times w'}$ and forwarded to the decoder.

The VSS block dynamically adapts its parameters based on the input. Unlike standard recurrent or convolutional operators that apply fixed weights across positions, VSS conditions its state-space dynamics on the current pulmonary vessels. We model the flattened vessel feature sequence $x(t)$ as a continuous driving of a latent state $h(t)$, as follow:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \tag{8}$$

where $\mathbf{A}$ controls how long the vessel path is preserved, effectively tracing the global structure. B is the pixels projection which enable discrete voxel-wise processing. The continuous system is discretized using a learned timescale Δ, yielding the recurrence.

$$\mathbf{h}_k = \overline{\mathbf{A}}\mathbf{h}_{k-1} + \overline{\mathbf{B}}\mathbf{x}_k \tag{9}$$

where $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ derived from the continuous matrices using standard zeroorder. Crucially, VSS block forms a linear time-invariant by making the state-space parameters input-dependent.

2.4 Loss Function

The student network $\mathcal{F}_s$ is trained end-to-end using a composite objective function that simultaneously optimizes for segmentation accuracy on labeled data, predictive consistency on mixed data, and manifold alignment in the latent mamba space. The total loss $\mathcal{L}_{\text{total}}$ is a weighted sum of three components:

$$\mathcal{L}_{\text{tatal}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{unsup}}\mathcal{L}_{\text{unsup}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} \tag{10}$$

where λ_{unsup} and λ_{feat} are weighted ramp-up factors that gradually increase the contribution of the unsupervised and feature consistency terms during training, avoiding instability in early epochs.

Supervised Loss: This component ensures the model learns precise anatomical segmentation from the available annotated data. Let $(\mathbf{X}^l, \mathbf{Y}^l)$ represent a batch of labeled inputs and their corresponding ground truth one-hot masks, where $Y^l \in \{0, 1\}^{C \times D \times H \times W}$ and $C = 3$ (background, artery, and vein). Let $\mathbf{P}^l = \mathcal{F}_s(\mathbf{X}^l)$ be the softmax probability maps predicted by the student network. To handle class imbalance inherent in vascular segmentation, we employ a hybrid loss combining categorical cross-entropy $\mathcal{L}_{\text{ce}}$ and soft dice loss $\mathcal{L}_{\text{dice}}$:

$$\mathcal{L}_{\text{sup}} = \frac{1}{2}\left(\mathcal{L}_{ce}\left(\mathbf{P}^l, \mathbf{Y}^l\right) + \mathcal{L}_{\text{dice}}\left(\mathbf{P}^l, \mathbf{Y}^l\right)\right) \tag{11}$$

$$\mathcal{L}_{ce} = -\frac{1}{N_{\text{vox}}}\sum_{v}\sum_{c=1}^{C}\mathbf{Y}_c^l(v)\log\left(\mathbf{P}_c^l(v)\right) \tag{12}$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{v}\sum_{c=1}^{C}\mathbf{P}_c^l(v)\mathbf{Y}_c^l(v)}{\sum_{v}\sum_{c=1}^{C}\left(\mathbf{P}_c^l(v)^2 + \mathbf{Y}_c^l(v)^2 + \varepsilon\right)} \tag{13}$$

where v denotes the voxel index, N_{vox} is the total number of voxels, and ϵ is a smoothing term to prevent division by zero.

Unsupervised Loss: This loss ensures that the student network predictions on mixed inputs remain consistent with the teacher network's pseudo-labels, leveraging the uncertainty-guided mixing from Stage I.

Let X^{mix} be the batch of mixed images generated using Gaussian soft blending. We obtain the student prediction $\mathbf{P}_s^{\text{mix}} = \mathcal{F}_s\left(\mathbf{X}^{\text{mix}}\right)$ and the teacher prediction $\mathbf{P}_t^{\text{mix}} = \mathcal{F}_t\left(\mathbf{X}^{\text{mix}}\right)$. We utilize mean squared error (MSE) on the soft probability maps to encourage smooth consistency:

$$\mathcal{L}_{\text{unsup}} = \frac{1}{N_{\text{vox}}}\sum_{v}\left\|\mathbf{P}_s^{\text{mix}}(v) - \left(\mathbf{P}_t^{\text{mix}}(v)\right)\right\|_2^2 \tag{14}$$

Feature Consistency Loss: This is the critical component of the latent space of the mamba bottleneck. It ensures that the features of mixed inputs $\mathbf{X}^{\text{mix}}$ linearly map to the mixing of their constituent latent features.

$$\mathcal{L}_{\text{feat}} = \|\mathbf{F}_{\text{s}} - \mathbf{F}_{\text{mix}}\|_2^2 \tag{15}$$

where $\mathbf{F}_{\text{s}} = \phi_s(\mathbf{X}^{\text{mix}})$ represents the bottleneck features of the student, and F_{mix} is the synthesized target feature map derived from the teacher network of labeled and unlabeled images.

3 Experiments and Results

To evaluate the efficacy of our proposed framework in handling complex pulmonary vascular topologies, we utilize the HiPaS dataset [18] and Parse dataset [19].

3.1 Data Pre-processing

The 3D image is first reoriented to the right-anterior-inferior coordinate system and anisotropically resampled to a target spacing of (0.65, 0.65, 1.0) mm to preserve high in-plane resolution, standardize slice thickness, and avoid interpolation artifacts along the axial direction. Intensity values are then clipped to the lung window $[-1000, 600]$ HU to suppress irrelevant structures and linearly normalized to [0, 1]. For training, large-context volumetric patches of size $1 \times 192 \times 192 \times 128$ are randomly extracted, enabling the model to capture long-range vascular continuity while employing balanced foreground sampling to alleviate class imbalance.

3.2 Comparisons with the State-of-the-Art Methods

To comprehensively verify the superiority of UMC-FM, we conduct quantitative comparisons with three categories of SOTA methods on both HiPaS and Parse datasets: (1) fully supervised CNN baselines UNet with different labeled data ratios; (2) Transformer-based SSL methods SwinUNETR and UNETR) with 10% labeled 90% unlabeled data; (3) the latest SSL model DiCo [1] with the same semi-supervised setting 10% labeled 90% unlabeled data. Performance is evaluated using three key metrics: higher DSC values, higher Jaccard values and lower HD_{95} and Average Surface Distance (ASD) values, which comprehensively reflect the volumetric overlap and boundary delineation accuracy of vascular segmentation.

Table 1 presents UMC-FM achieves the best performance across all metrics, with a DSC of 77.60 $\pm$ 3.83%, HD_{95} of 6.62 $\pm$ 0.27 voxels and ASD of 2.48 $\pm$ 1.89 voxels. Notably, our method outperforms the latest SSL model DiCo by a large margin 12.73 DSC points, with DiCo suffering from severe boundary errors HD_{95} 18.37 $\pm$ 1.91 voxels due to its ineffective handling of vascular topology. Compared with fully supervised UNet trained with 100% labeled data, UMC-FM still achieves a 2.58 DSC point improvement, demonstrating its excellent data efficiency. Transformer-based methods SwinUNETR and UNETR exhibit poor performance of DSC less than 61%, due to their high computational overhead and poor generalization on label-scarce vascular data. Even the semi-supervised VNet is outperformed by UMC-FM by 5.04 DSC points, confirming the effectiveness of our Mamba-based architecture and dual-consistency strategy.

Table 1. Performance comparison of different methods on the HiPaS dataset.

Method	Scan used		Metrics			
	Labeled	Unlabeled	Dice (%) ↑	Jaccard (%) ↑	HD_{95} (voxel) ↓	ASD (voxel) ↓
UNet	5%	95%	57.83 $\pm$ 0.90	53.17 $\pm$ 1.03	14.63 $\pm$ 1.29	9.64 $\pm$ 0.17
	10%	90%	69.81 $\pm$ 2.63	61.18 $\pm$ 1.19	10.77 $\pm$ 0.53	6.81 $\pm$ 0.67
	100%	0	75.02 $\pm$ 2.12	71.97 $\pm$ 1.29	7.23 $\pm$ 0.89	5.17 $\pm$ 0.25
SwinUNETR	10%	90%	60.71 $\pm$ 1.45	44.49 $\pm$ 2.89	16.42 $\pm$ 0.58	4.05 $\pm$ 2.63
UNETR			58.29 $\pm$ 2.14	41.74 $\pm$ 3.61	23.723 $\pm$ 0.63	5.72 $\pm$ 1.24
VNet			72.56 $\pm$ 3.36	58.99 $\pm$ 2.56	9.76 $\pm$ 1.61	2.95 $\pm$ 0.67
DiCo			64.87 $\pm$ 5.37	48.36 $\pm$ 2.13	18.37 $\pm$ 1.91	6.64 $\pm$ 0.45
Proposed			**77.60 $\pm$ 3.83**	**64.93 $\pm$ 3.38**	**6.62 $\pm$ 0.27**	**2.48 $\pm$ 1.89**

Table 2 presents that the superior performance of UMC-FM is further verified on the Parse dataset, where our method again achieves SOTA across all metrics, with a DSC of 77.81 $\pm$ 1.21, HD_{95} of 12.07 $\pm$ 0.11 voxels, and ASD of 3.23 $\pm$ 0.93 voxels. It surpasses the latest semi-supervised model DiCo by a large margin of 12.73 DSC points. This outperforms the semi-supervised VNet by 2.20 DSC points and the fully supervised UNet by 0.57 points, while exhibiting the lowest boundary errors.

Table 2. Performance comparison of different methods on the Parse dataset.

Method	Scan used		Metrics			
	Labeled	Unlabeled	Dice (%) ↑	Jaccard (%) ↑	HD_{95} (voxel) ↓	ASD (voxel) ↓
UNet	5%	95%	61.17 ± 1.70	59.54 ± 1.87	16.76 ± 2.67	5.91 ± 1.17
	10%	90%	73.95 ± 1.54	62.79 ± 0.25	15.47 ± 0.95	5.73 ± 0.81
	100%	0	77.24 ± 3.56	63.19 ± 1.87	12.44 ± 1.97	3.47 ± 1.50
SwinUNETR	10%	90%	56.15 ± 1.27	39.20 ± 0.16	16.99 ± 0.78	5.97 ± 0.31
UNETR			52.06 ± 1.17	35.73 ± 4.27	17.46 ± 0.87	6.98 ± 1.77
VNet			75.61 ± 4.96	61.01 ± 2.23	12.56 ± 1.60	3.42 ± 1.53
DiCo			38.50 ± 2.13	23.21 ± 3.61	27.69 ± 1.60	9.31 ± 069
Proposed			**77.81 ± 1.21**	**63.82 ± 4.83**	**12.07 ± 0.11**	**3.23 ± 0.93**

Table 3 summarizes the component-wise validation of the proposed framework, illustrating the progressive performance gains. Integrating the mamba with the encoder-decoder yields a clear improvement in segmentation accuracy.

Table 3. The proposed UMC-FM framework on the HiPaS dataset reports the impact of different backbone architectures, sampling strategies, and the feature consistency loss L_{feat} on segmentation performance.

Experiment	Backbone	Components		DSC (%)↑	HD_{95} (mm)↓	ASD (voxel) ↓
		Sampling	L_{feat}			
Baseline	V-Net	Random	×	72.56	9.76	2.95
Exp. 1	Mamba	Random	×	74.85	8.42	2.71
Exp. 2	Mamba	Uncertainty	×	76.40	7.35	2.60
Proposed	Mamba	Uncertainty	✓	**77.60**	**6.62**	**2.48**

3.3 Qualitative Results

To provide an intuitive assessment of the segmentation results, we provide a comparative 3D visualization in Fig. 3. The transformer-based baselines SwinUNETR and UNETR exhibit structural artifacts. Specifically, UNETR suffers from severe anatomical hallucinations and disconnected artifacts, failing to reconstruct the cohesive vascular tree. While pure CNN-based methods, UNet and VNet, capture the primary pulmonary trunk effectively, they struggle to preserve the continuity of fine distal branches, resulting in a sparse and fragmented vascular representation compared to the ground-truth.

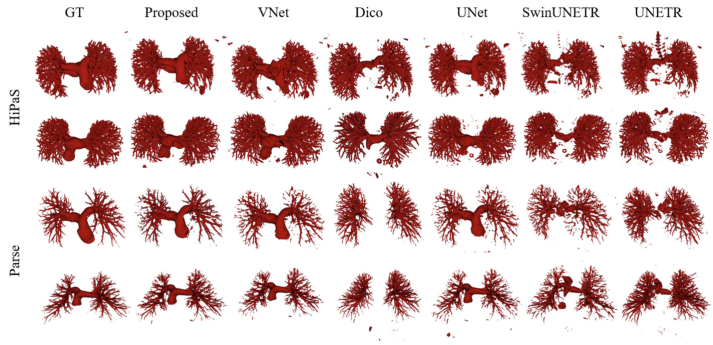


Fig. 3. The visual comparison of 3D pulmonary vessel segmentation on the HiPaS and Parse dataset.

The slice-level qualitative results are illustrated in Fig. 4. Notable failure cases are observed in SwinUNETR and UNETR, which produce substantial over-segmentation by incorrectly classifying large non-vascular regions as pulmonary vessels. This suggests limited precision in separating vessel boundaries from the surrounding background. In contrast, UNet and VNet largely avoid severe false positives but exhibit under-segmentation, frequently missing fine vascular structures. These CNN-based methods demonstrate structural discontinuities at challenging bifurcation regions.

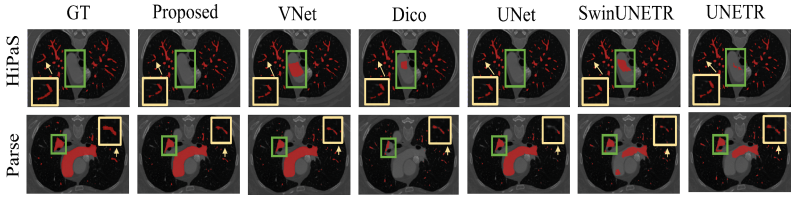


Fig. 4. The visual comparison of 2D pulmonary vessel segmentation results on HiPaS and PARSE datasets. The yellow arrows denote challenging peripheral vessels, and the green box indicates false-positive regions.

Figure 5 presents a quantitative comparison of segmentation performance on the HiPaS dataset. The proposed framework achieves a superior mean DSC of 77.60%, outperforming the fully supervised UNet baseline 75.02% and significantly surpassing transformer-based architectures such as SwinUNETR 60.71% and UNETR 58.29%. In terms of boundary delineation, our method yields the lowest error rates, with a 95% mean HD_{95} of 6.62 voxels and an ASD of 2.48 voxels. This indicates that a visual state-based Mamba backbone with uncertainty-guided consistency not only enhances volumetric overlap but also more effectively preserves topological correctness.

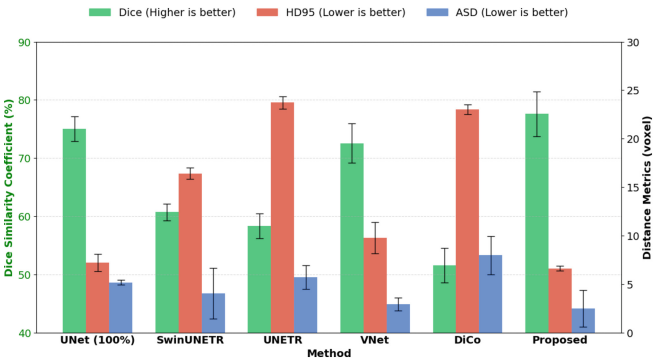


Fig. 5. The average Performance metrics comparison of proposed UMC-FM method.

4 Conclusion

In this work, we introduce UMC-FM, a robust semi-supervised framework designed to overcome the limitations of data scarcity and topological complexity in 3D pulmonary vessel segmentation. By synergizing a visual state space (Mamba) backbone with an uncertainty-guided dual-stage consistency, our method efficiently captures global vascular dependencies under limited annotation settings. Evaluation on the HiPaS and Parse datasets demonstrates that UMC-FM consistently establishes new performance benchmarks against typical CNN and transformer architectures. Furthermore, our method significantly improves the structural continuity of fine-grained distal pulmonary vascular branches. Importantly, UMC-FM handles extreme morphological variations-such as severe stenoses or occlusions that typically cause topological disconnections in CNNs-by leveraging Mamba global contextual awareness to inferentially bridge structural gaps, while our uncertainty-guided mixing explicitly aligns these highly ambiguous regions. Ultimately, our proposed UMC-FM method exceeds the segmentation performance of a fully supervised UNet baseline trained with 100% label data, highlighting its profound data efficiency and robustness in clinical settings.

Acknowledgments. This study was supported by the Shenzhen Science and Technology Program (KJZD20240903095605007, JCYJ20250604145033042) and the Shenzhen Medical Research Fund (D250402003), and Natural Science Foundation of Top Talent of SZTU (Grant No. GDRC202213).

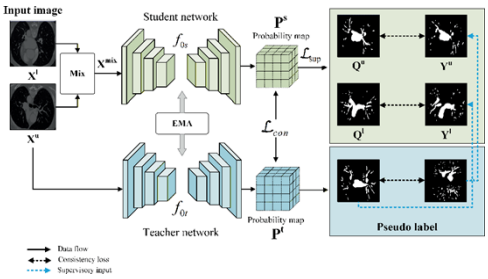
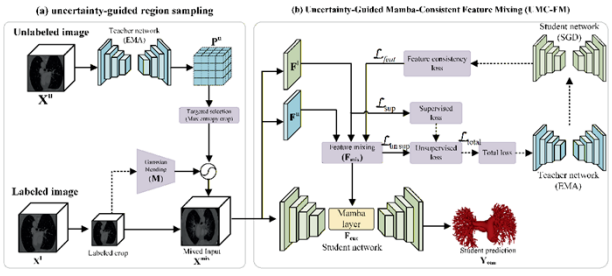
Disclosure of Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

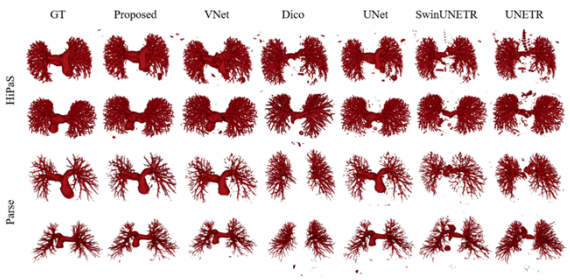
References

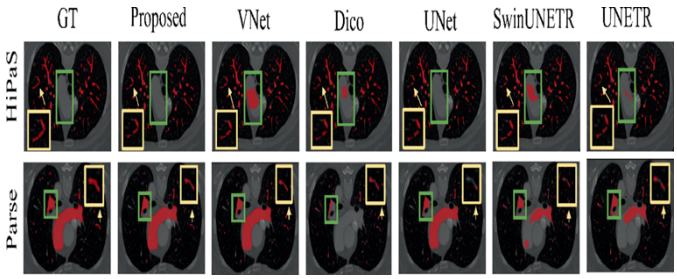
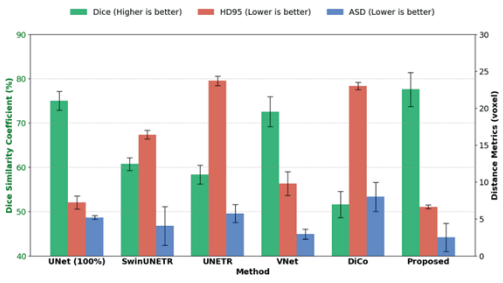
1. Xu, J., Chen, X., Zhang, L.: Learning dynamic collaborative network for semi-supervised 3D vessel segmentation. In: In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10445–10454. IEEE (2025)

2. A. Tarvainen and H. Valpola, “Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results,” (2017).
3. Chen, D., Bai, Y., Shen, W., Li, Q., Yu, L., Wang, Y.: MagicNet: semi-supervised multi-organ segmentation via magic-cube partition and recovery. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 23869–23878 (2022)
4. Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K.: Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE Trans. Med. Imaging. **42**, 1265–1277 (2022)
5. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, in NIPS ‘20. Curran Associates Inc., Red Hook (2020)
6. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 34–43 (2022)
7. Shi, Y., et al.: Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging. **41**(3), 608–620 (2021)
8. Wang, X., et al.: SSA-net: spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning. Med. Image Anal. **79**, 102459 (2022)
9. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3D semantic segmentation for medical images. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4—8, 2020, Proceedings, Part I 23, 2020, pp. 552–561 (2020)
10. Luo, X., Chen, J., Song, T., Chen, Y., Wang, G., Zhang, S.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conference on Artificial Intelligence (2020)
11. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Adv. Neural Inf. Process. Syst. **29** (2016)
12. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12671–12681 (2020)
13. Jin, Q., Meng, Z.-P., Pham, T.D., Chen, Q., Wei, L., Su, R.: DUNet: a deformable network for retinal vessel segmentation. Knowl. Based Syst. **178**, 149–162 (2018)
14. Dai, J., et al.: Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 764–773 (2017)
15. Qi, Y., He, Y., Qi, X., Zhang, Y., Yang, G.: Dynamic Snake convolution based on topological geometric constraints for tubular structure segmentation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6047–6056 (2023)
16. Wang, J., Zhou, L., Yuan, Z., Wang, H., Shi, C.: MIC-net: multi-scale integrated context network for automatic retinal vessel segmentation in fundus image. Math. Biosci. Eng. **20**(4), 6912–6931 (2023)
17. Yu, F., Koltun, V., Funkhouser, T.A.: Dilated residual networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 636–644 (2017)
18. Chu, Y., et al.: Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences. Nat. Commun. **16** (2024)
19. Luo, G. et al.: Efficient automatic segmentation for multi-level pulmonary arteries: the PARSE challenge. arXiv preprint (2024)

Alternative Texts for Your Images, Please Check and Correct them if Required

Page no	Fig/Photo	Thumbnail	Alt-text Description
3	Fig1		A diagram illustrating a semi-supervised learning process for image segmentation using student and teacher neural networks. Two input images, one labeled and one unlabeled, are combined and fed into the student network, producing a probability map. The unlabeled image also goes directly to the teacher network, which generates its own probability map. The student and teacher networks exchange information through an exponential moving average (EMA) to improve learning. The student's output is compared to the labeled data using supervised loss, while the student and teacher outputs are compared using consistency loss. On the right, two sets of black-and-white medical images show the comparison between predicted segmentations and ground truth labels for both labeled and pseudo-labeled data, highlighting how the model refines its predictions. Arrows indicate data flow, consistency loss, and supervisory input, emphasizing the interaction between networks and the use of both labeled and unlabeled data to enhance segmentation accuracy.
4	Fig2		A flow chart compares two methods for processing 3D medical images to improve tumor segmentation. On the left, the first method uses a single 3D image input that passes through a 3D convolutional neural network, followed by a 3D decoder, producing a segmented tumor

Page no	Fig/Photo	Thumbnail	Alt-text Description
			output. On the right, the second method splits the 3D image into multiple 2D slices, processes each slice through a 2D convolutional neural network and decoder, then combines the results to form the final 3D tumor segmentation. Arrows show the flow from input images through processing steps to output. The chart highlights the difference between using full 3D data at once versus slice-by-slice 2D processing, illustrating approaches to improve accuracy in identifying tumors in medical scans.
10	Fig3		Comparison of 3D lung airway models from two datasets, HiPas and Parse, showing ground truth and results from seven different methods: Proposed, VNet, Dico, UNet, SwinUNETR, and UNETR. Each row represents a dataset, with the first two rows for HiPas and the last two for Parse. The images display detailed branching structures in dark red, illustrating how closely each method replicates the ground truth. The Proposed method shows airway structures most similar to the ground truth, with clearer and more complete branches, while other methods vary in accuracy and completeness. This comparison highlights the effectiveness of the Proposed method in accurately reconstructing lung airways from medical imaging data.

Page no	Fig/Photo	Thumbnail	Alt-text Description
10	Fig4		Two rows of lung scan images compare different methods for highlighting blood vessels in red. The top row shows scans from the HiPaas dataset, and the bottom row shows scans from the Parse dataset. Each column represents a different method: GT (ground truth), Proposed, VNet, Dico, UNet, SwinUNETR, and UNETR. The green boxes highlight a specific lung area, and the yellow boxes zoom in on smaller vessel details. The Proposed method closely matches the ground truth in both datasets, showing clear and accurate vessel structures, while other methods show varying levels of detail and accuracy. This comparison highlights the effectiveness of the Proposed method in accurately segmenting lung blood vessels.
11	Fig5		Comparison of six methods based on three performance measures: Dice similarity coefficient (higher values are better, shown in green), HD95 distance metric (lower values are better, shown in red), and ASD distance metric (lower values are better, shown in blue). The proposed method achieves the highest Dice score around 77%, a low HD95 value near 6, and the lowest ASD value close to 2, indicating better overall accuracy and precision compared to other methods. UNet and VNet also show relatively high Dice scores but have higher HD95 and ASD values. This chart highlights the proposed method as the most effective in balancing accuracy and distance metrics.