SPARNET: BREAKING THE DETAIL–DIRECTION TRADE-OFF IN 3D MEDICAL STATE-SPACE MODELS WITH SCALE-PRESERVING RADIAL SCANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

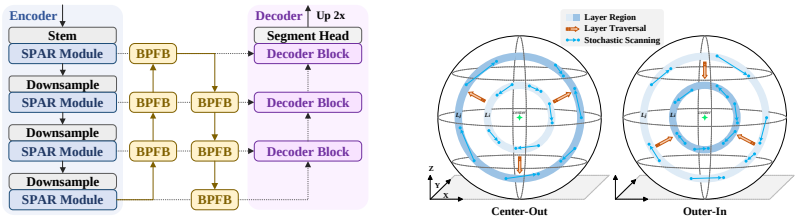
State-space models (SSMs) have emerged as an efficient alternative to Transformers for 3D medical image segmentation, yet how volumetric features are tokenized and serialized for these models remains largely overlooked. We identify two coupled limitations in existing 3D SSMs. First, fixed and coarse token scales cause detail loss at the boundaries of small or low-contrast structures. Second, Cartesian-axis serialization separates neighboring voxels and introduces directional bias into state propagation. Because tokenization determines what information is preserved and scan order determines how it propagates, we argue that the two must be addressed jointly. We present SPAR-Net, a Scale-Preserving Adaptive Radial Network built on this principle. A Scale-Adaptive Feature Extractor aligns fine-, intermediate, and coarse-scale features on a shared high-resolution token grid, which adapts the receptive field without discarding spatial detail. Stochastic Layer Scanning replaces axis-aligned scans with center-referenced radial layers traversed along complementary inward and outward paths. A Feature Refinement block restores volumetric correspondence after sequence modeling, and a Bidirectional Progressive Fusion Block exchanges semantic and boundary cues across encoder levels. Under a unified protocol, SPARNet consistently improves both Dice and HD95 over existing 3D segmentation models. It achieves 85.25% mean Dice and 12.20 HD95 on Synapse multi-organ CT, and 64.02% mean Dice and 17.15 HD95 on MSD Pancreas. The largest gains appear on boundary-sensitive organs such as the gallbladder and pancreas, and on pancreatic tumors. Analyses of token scale, scan order, feature fusion, and efficiency show that every component contributes. They also show that fine multi-scale tokenization and radial bidirectional propagation are complementary.

1 INTRODUCTION

Accurate 3D medical image segmentation must reconcile two competing requirements: preserving fine spatial detail for small, low-contrast structures such as vessels and lesions, while capturing long-range context for large anatomical regions. Convolutional encoder-decoder networks preserve local structure well, but their receptive fields expand gradually with depth (Çiçek et al., 2016; Isensee et al., 2021; Roy et al., 2023). Transformers provide direct global interactions, but dense self-attention incurs high computational cost on high-resolution 3D volumes (Hatamizadeh et al., 2022; 2021; Zhou et al., 2021; Chen et al., 2023). Selective state-space models (SSMs) offer an attractive alternative, propagating information through long sequences with linear complexity (Gu & Dao, 2023; Zhu et al., 2024; Liu et al., 2024b), motivating several Mamba-based architectures for medical image segmentation (Ma et al., 2024; Liu et al., 2024a; Xing et al., 2024; Chang et al., 2024; Xing et al., 2025).

Applying SSMs to 3D volumes introduces two closely related design choices: *token scale* and *token order*. A fixed patch size cannot simultaneously represent large anatomical structures and fine boundaries, while Cartesian serialization may separate neighboring voxels and bias information propagation along predefined axes. These choices are inherently coupled: tokenization determines what information is preserved, and scan order determines how it propagates through the recurrent state.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) SPAR-Net architecture. (b) Radial layer construction.

Figure 1: SPAR-Net overview and radial scanning: (a) four SPAR stages with bidirectional cross-scale fusion; (b) center-referenced layers traversed inward or outward.

We address both challenges with **SPAR-Net**, a **S**cale-**P**reserving **A**daptive **R**adial Network for 3D medical image segmentation. The name reflects its design. *Scale-preserving* means that multi-scale context is aligned onto a shared high-resolution token grid rather than fused only after spatial detail has been discarded. *Adaptive* refers to the joint use of small, medium, and large receptive-field streams for structures of different spatial extent. *Radial* refers to center-referenced sequence construction, which replaces a fixed Cartesian preference with complementary inward and outward propagation. SPAR-Net implements this principle through four components. The Scale-Adaptive Feature Extractor (SAFE) forms parallel patch streams and aligns them on a common fine grid. Stochastic Layer Scanning (SLS) partitions the grid into radial layers and constructs complementary center-directed paths for state-space modeling. The Feature Refinement (FR) block restores the processed tokens to a structured 3D feature map. Finally, the Bidirectional Progressive Fusion Block (BPFB) exchanges semantic and boundary information across adjacent encoder levels. Together, these components preserve boundary evidence before sequence modeling and maintain spatial correspondence after it.

We evaluate SPAR-Net on Synapse multi-organ CT segmentation and MSD Pancreas tumor segmentation (Landman et al., 2015; Antonelli et al., 2022). Under a unified training and evaluation protocol, SPAR-Net consistently improves both Dice and HD95, with particularly clear gains on the gallbladder, pancreas, and pancreatic tumors. Component, scan-order, patch-scale, efficiency, and plug-in analyses converge on a consistent explanation: preserving a fine multi-scale token grid improves representation quality, while radial bidirectional propagation improves boundary continuity.

Our contributions are threefold:

- We propose SPAR-Net, a state-space architecture that jointly addresses multi-scale representation and sequence construction through scale-preserving tokenization and radial propagation.
- We introduce a Scale-Adaptive Feature Extractor that aligns multi-scale features on a shared high-resolution token grid, together with Stochastic Layer Scanning that constructs complementary inward and outward radial sequences for state-space modeling.
- We perform comprehensive analyses of token scale, scan order, feature fusion, efficiency, and architectural components, demonstrating consistent improvements over existing 3D segmentation models.

2 Related Work

3D Medical Image Segmentation: CNN-based encoder-decoder models, including 3D U-Net (Çiçek et al., 2016), SegResNet (Myronenko, 2018), nnU-Net (Isensee et al., 2021), 3D UX-Net (Lee et al., 2022), and MedNeXt (Roy et al., 2023), remain strong volumetric baselines because convolutions preserve local spatial structure. Transformer architectures such as UNETR (Hatamizadeh et al., 2022), Swin UNETR (Hatamizadeh et al., 2021), nnFormer (Zhou et al., 2021), and 3D TransUNet (Chen et al., 2023) expand the receptive field through self-attention. Their global interactions are useful, but the memory and computation of dense attention motivate sequence models with more favorable scaling.

State-Space Models and Spatial Serialization: Mamba introduces input-dependent selective state-space updates with linear sequence complexity (Gu & Dao, 2023). Vision Mamba and VMamba

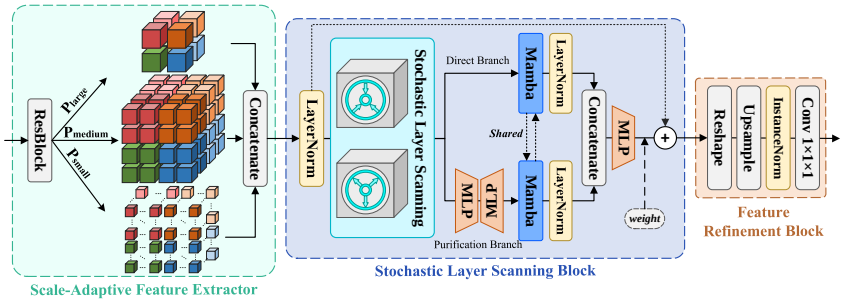


Figure 2: The SPAR module. SAFE aligns multi-scale patch streams on a shared fine grid, SLS performs complementary radial state-space modeling, and FR restores the processed sequence to a structured 3D feature map.

adapt this mechanism to images through directional scanning (Zhu et al., 2024; Liu et al., 2024b). Medical variants, including U-Mamba (Ma et al., 2024), Swin-UMamba (Liu et al., 2024a), SegMamba (Xing et al., 2024), EM-Net (Chang et al., 2024), and SegMamba-V2 (Xing et al., 2025), demonstrate the value of state-space modeling for volumetric segmentation. Nevertheless, serialization remains a structural design choice. Axis-aligned flattening is efficient but encodes a fixed directional preference and does not explicitly capture the 3D spatial relationships among tokens. SPAR-Net instead constructs radial layers before state propagation and evaluates alternative linear, spiral, stochastic, and paired traversal rules in the same backbone.

Multi-Scale Representation and Feature Fusion: Scale variation has been addressed through hierarchical representations (Heidari et al., 2023), adaptive volumetric decomposition (Wang et al., 2023), boundary-aware learning (Pham et al., 2025), and dynamic feature fusion (Yang et al., 2026). These approaches improve scale coverage or boundary recovery, but multi-scale representation and state-space serialization are usually designed separately. SPAR-Net connects them in a single information flow: SAFE adapts receptive-field scale while preserving the finest token grid, SLS defines radial propagation on that grid, and BPFB reconciles semantic and boundary information across encoder levels.

3 METHOD

Overview: Accurate 3D medical image segmentation requires fine spatial detail to delineate small or low-contrast structures, while also capturing broad contextual information for larger anatomical regions. To address these complementary requirements, SPAR-Net adopts a four-stage encoder–decoder architecture, as shown in Fig. 1. Each encoder stage employs a SPAR module to extract multi-scale features and model long-range dependencies, while a convolutional decoder progressively integrates the encoded representations to generate dense voxel-wise predictions.

Within each encoder stage, the SPAR module comprises three main components, illustrated in Fig. 2. The Scale-Preserving Adaptive Feature Extractor (SAFE) first constructs a multi-scale representation on a shared fine-resolution token grid. Stochastic Layer Scanning (SLS) then captures long-range dependencies along complementary radial paths, after which the Feature Refinement (FR) block reconstructs the processed tokens into a structured 3D feature volume. Together, these components preserve fine boundary information during tokenization and maintain spatial correspondence after state-space modeling. The Bidirectional Progressive Fusion Block (BPFB) further exchanges complementary semantic and boundary information between adjacent encoder levels before the features are passed to the decoder. The following sections describe each component in detail.

Stage Configuration: Table 1 summarizes the encoder configuration used in all experiments. The first two stages use three SAFE streams, whereas the deepest stages omit the largest stream because their feature grids are too small to support it effectively. The SLS depth is fixed across stages to isolate the effects of token scale and scan order from stage-specific architectural tuning.

Table 1: Stage configuration. Patch entries list SAFE streams; d is SLS depth and Δr is radial-layer width.

Stage	Channels	Synapse patches	MSD patches	$d/\Delta r$
1	48	$[2, 4, 8]$	$[2, 4, 8]$	$3/3.0$
2	96	$[2, 4, 8]$	$[1, 2, 4]$	$3/3.0$
3	192	$[1, 2, 4]$	$[1, 2, 4]$	$3/3.0$
4	384	$[1, 2]$	$[1, 2]$	$3/3.0$

3.1 Scale-Adaptive Feature Extractor (SAFE)

Most 3D segmentation networks rely on fixed-size tokens for all anatomical structures. This is suboptimal for multi-organ segmentation, where small organs require fine boundary precision while large organs require broader semantic context. To address this limitation, SAFE performs multi-scale tokenization and aligns all feature streams onto a shared fine grid, as shown in Figure 2.

Let $F \in \mathbb{R}^{C\times D\times H\times W}$ be an input feature map, where C is the number of channels and D, H, W denote the depth, height, and width of the 3D feature volume. SAFE divides F into three parallel patch streams with patch sizes $\{p_s, p_m, p_l\}$, corresponding to small, medium, and large receptive fields. The small-patch stream preserves fine boundary details, the medium-patch stream captures intermediate structural patterns, and the large-patch stream provides coarse semantic context for large volumetric regions.

To unify spatial resolution, the medium- and large-patch streams are upsampled to the finest grid defined by p_s using 3D pixel shuffle Shi et al. (2016). The aligned streams are then concatenated along the channel dimension and normalized:

$$F_{tok} = \mathrm{LN}([F_{p_s}, F^{\uparrow}_{p_m}, F^{\uparrow}_{p_l}]), \tag{1}$$

where $\uparrow$ denotes pixel-shuffle upsampling and LN denotes layer normalization. The resulting F_{tok} is a unified token representation that preserves both fine-scale boundary cues and large-scale semantic information for subsequent layer-aware sequence modeling.

3.2 Stochastic Layer Scanning (SLS)

State-space models such as Mamba Gu & Dao (2023) capture long-range dependencies with linear complexity, but applying them to 3D volumes requires converting spatial features into one-dimensional sequences. Existing 3D extensions often use fixed Cartesian-axis scanning, which can disrupt local spatial continuity and introduce directional bias because nearby voxels in 3D may become distant in the serialized sequence.

To reduce this dependence on axis-aligned serialization, we introduce Stochastic Layer Scanning (SLS). As shown in Figure 1b, SLS groups tokens into center-referenced radial layers and scans them through inward or outward traversal paths. The reference point is the geometric center of the current crop or sliding-window region, not an organ-specific anatomical centroid. This makes the scan rule available during both training and inference without requiring ground-truth anatomical priors.

Let L_i denote the i-th radial layer obtained by quantizing the radius r into uniform bins, where Δr controls the interval between adjacent layers. For each token at coordinate (z, y, x), its radial distance is computed relative to the crop center and assigned to layer L_i by $i = \lfloor r/\Delta r \rfloor$. This produces a near–mid–far organization of tokens within the processed 3D crop.

Within each layer L_i, tokens are randomly permuted to introduce stochastic traversal diversity. Across layers, SLS follows either inward traversal, from peripheral layers to the center, or outward traversal, from the center to peripheral layers. The resulting one-dimensional index sequence is precomputed and used as a permutation operator in the SSM encoder, while the inverse permutation restores the canonical spatial layout after sequence modeling. This design preserves efficient sequence processing while exposing the encoder to spatially structured but non-deterministic scan patterns.

We use two complementary SLS strategies. **Global Layer Scanning** traverses radial layers across the full crop to preserve long-range organ continuity. **Regional Rotational Layer Scanning** applies localized stochastic traversal with cyclic shifts and random rotations, improving directional diversity for small or highly variable structures. Their empirical comparison is reported in the experimental section.

3.2.1 DUAL-BRANCH STATE-SPACE MODELING:

A limitation of long-range state-space modeling in 3D segmentation is that global semantic aggregation may suppress boundary-sensitive, high-frequency cues. Although SLS provides a more spatially coherent token order, a single-path Mamba update may still struggle to preserve both global context and fine structural detail. To mitigate this imbalance, we adopt a dual-branch design that separates detail-preserving modeling from semantic purification. This design is conceptually related to EM-Net Chang et al. (2024), but both branches in our model operate on layer-aware token sequences rather than axis-flattened sequences.

After SLS, the token sequence $F_{tok} \in \mathbb{R}^{N\times C}$ is treated as $X \in \mathbb{R}^{N\times C}$, where N is the number of tokens and C is the channel dimension. The sequence is processed by two complementary Mamba branches with shared weights.

Direct branch. The direct branch operates on the uncompressed token sequence to preserve fine anatomical details:

$$Z_{dir} = \mathrm{LN}(\mathrm{Mamba}(\mathrm{LN}(X)))\,. \tag{2}$$

Because this branch does not compress the representation, it helps retain high-frequency structural cues.

Purification branch. The purification branch focuses on semantic abstraction by filtering redundant channels:

$$Z_{pur} = \mathrm{LN}\big(\mathrm{Mamba}\big(\mathrm{LN}\big(\mathrm{MLP}_{\uparrow}(\mathrm{MLP}_{\downarrow}(X))\big)\big)\big)\,, \tag{3}$$

where $\mathrm{MLP}_{\downarrow}$ compresses the channel dimension and $\mathrm{MLP}_{\uparrow}$ restores it. This suppresses redundant responses while retaining global semantic cues.

Fusion and residual integration. The two branch outputs are concatenated and normalized, and the fused representation is obtained as $Z_{fused} = \mathrm{MLP}(\mathrm{LN}([Z_{dir}, Z_{pur}]))$. The final output is then computed by $Z_{out} = \alpha Z_{fused} + X$, where α is a trainable scalar controlling the refinement strength. This residual formulation stabilizes optimization, preserves identity information, and avoids over-correction.

3.3 FEATURE REFINEMENT (FR)

The SLS block models long-range dependencies in the token domain, but dense segmentation requires a structured volumetric feature map. The FR Block therefore transforms the token sequence back into 3D space and restores voxel-wise correspondence.

Given $Z_{out} \in \mathbb{R}^{N\times C}$, we first rearrange it into a compact 3D grid whose spatial resolution matches the finest patch granularity p_s defined in SAFE. A 3D pixel-shuffle operation is then applied to recover the full-resolution spatial layout:

$$G_{up} = \mathrm{PixelShuffle}_{3D}(\mathrm{Reshape}(Z_{out}))\,. \tag{4}$$

The upsampled feature is then normalized and projected through a pointwise convolution, i.e., $F_{out} = \mathrm{Conv}_{1\times1\times1}(\mathrm{IN}(G_{up}))$, where IN denotes instance normalization. The resulting $F_{out} \in \mathbb{R}^{C_{out}\times D\times H\times W}$ is a reconstructed 3D feature map that restores spatial alignment and consolidates the multi-scale information encoded by SAFE and SLS.

3.4 BIDIRECTIONAL PROGRESSIVE FUSION BLOCK (BPFB)

Encoder–decoder segmentation networks often struggle to balance global semantics and fine structural precision. Deeper features capture holistic context but lose detailed boundaries, whereas shallower features preserve boundary cues but lack semantic abstraction. To address this discrepancy, BPFB facilitates two-way feature exchange across encoder levels, as illustrated in Figure 1a. Unlike conventional unidirectional skip connections, BPFB propagates information both top–down and

Table 2: Results on Synapse. All methods are trained and evaluated using the same data split, preprocessing, optimization, checkpoint selection, and sliding-window inference protocol. Values are reported as mean ± standard deviation across repeated runs. SPAR-Net [8R-Rot] uses eight-region rotational SLS.

Method	Spl.	L.Kid.	R.Kid.	Gall.	Liv.	Sto.	Aor.	Panc.	Dice↑	HD95↓
3D U-Net (Çiçek et al., 2016)	89.22 ±1.18	85.79 ±0.18	84.99 ±0.19	61.32 ±3.10	95.39 ±0.07	78.44 ±0.54	90.41 ±0.33	65.19 ±0.52	81.35 ±0.59	21.95 ±2.83
SegResNet (Myronenko, 2018)	91.61 ±1.53	84.60 ±1.47	85.51 ±0.28	63.76 ±4.89	95.15 ±0.35	76.09 ±0.67	90.49 ±1.22	65.09 ±3.34	81.54 ±0.09	33.92 ±3.23
3D UX-Net (Lee et al., 2022)	90.48 ±1.24	86.07 ±0.42	84.27 ±0.71	63.58 ±1.91	95.58 ±0.29	75.19 ±1.26	89.24 ±0.57	64.18 ±0.97	81.07 ±0.17	34.41 ±2.08
MedNeXt (Roy et al., 2023)	87.05 ±1.36	85.67 ±0.67	84.96 ±1.39	65.25 ±1.76	<u>95.77</u> ±0.22	78.18 ±1.74	89.78 ±0.68	70.07 ±2.09	82.09 ±0.20	32.64 ±6.05
UNETR (Hatamizadeh et al., 2022)	88.63 ±0.42	82.70 ±0.30	82.19 ±0.25	62.01 ±1.10	93.34 ±0.34	66.52 ±1.84	84.20 ±0.33	54.48 ±1.05	76.76 ±0.46	38.05 ±4.63
Swin UNETR (Hatamizadeh et al., 2021)	89.00 ±1.73	85.44 ±0.29	83.96 ±1.17	64.61 ±1.89	95.19 ±0.28	75.37 ±0.88	89.81 ±0.77	65.46 ±0.74	81.11 ±0.18	36.61 ±7.09
SegMamba (Xing et al., 2024)	90.10 ±1.80	86.63 ±0.84	85.73 ±0.30	66.76 ±2.82	94.70 ±1.02	77.13 ±0.89	91.22 ±0.22	69.51 ±0.77	82.72 ±0.50	30.69 ±8.11
U-Mamba (Ma et al., 2024)	91.85 ±1.13	85.88 ±0.50	86.17 ±0.37	62.67 ±0.86	95.76 ±0.18	78.92 ±2.81	<u>91.69</u> ±0.23	68.46 ±1.56	82.68 ±0.15	31.50 ±2.32
SegMamba-V2 (Xing et al., 2025)	90.92 ±0.65	86.31 ±0.77	85.82 ±1.03	68.45 ±0.66	95.51 ±0.63	77.52 ±1.32	91.68 ±0.28	71.15 ±2.19	83.42 ±0.44	28.34 ±9.56
SPAR-Net	**93.79** ±0.91	<u>86.65</u> ±0.08	**86.44** ±0.15	**71.10** ±0.36	95.64 ±0.26	<u>82.46</u> ±2.10	91.55 ±0.35	<u>73.19</u> ±0.68	<u>85.10</u> ±0.21	<u>12.99</u> ±3.29
SPAR-Net [8R-Rot]	<u>93.31</u> ±1.43	**86.69** ±0.08	<u>86.34</u> ±0.22	<u>70.51</u> ±0.46	**95.94** ±0.20	**82.58** ±0.80	**91.87** ±0.03	**74.76** ±1.00	**85.25** ±0.20	**12.20** ±3.34

bottom–up, allowing semantic features to guide boundary localization while boundary cues reinforce semantic consistency.

At level ℓ, BPFB receives the encoder feature $E_\ell \in \mathbb{R}^{C_\ell \times D_\ell \times H_\ell \times W_\ell}$ and an adjacent feature E_j from a neighboring scale $j \in \{\ell - 1, \ell + 1\}$. The adjacent feature is resampled to match the spatial resolution of level ℓ:

$$N_{\ell \leftarrow j} = \mathrm{Interp}_{tri}\big(E_j \to (D_\ell, H_\ell, W_\ell)\big), \tag{5}$$

where $N_{\ell \leftarrow j}$ denotes the aligned cross-scale feature. The aligned feature is concatenated with the current feature, i.e., $F_\ell^{cat} = [E_\ell, N_{\ell \leftarrow j}]$, and then processed by Channel Shuffle Zhang et al. (2018) followed by a pointwise $1 \times 1 \times 1$ convolution:

$$F_\ell^{out} = \sigma\Big(\mathrm{IN}\big(\mathrm{Conv}_{1\times1\times1}(\mathrm{Shuffle}(F_\ell^{cat}))\big)\Big), \tag{6}$$

where σ denotes the LeakyReLU activation. Channel Shuffle promotes cross-branch mixing, while the pointwise projection reduces redundancy introduced by concatenation.

By progressively applying bidirectional fusion, BPFB transmits semantic information downward while propagating boundary precision upward. This two-way refinement harmonizes global and local representations, improving spatially consistent feature fusion for both small boundary-sensitive organs and large volumetric structures.

4 Experimental Results

We evaluate SPAR-Net from three complementary perspectives: segmentation accuracy under a unified protocol, the contribution of each architectural choice, and the resulting computational trade-off. The analysis focuses on whether scale-preserving token alignment and radial propagation improve both overlap and boundary consistency.

Table 3: MSD Pancreas results using the unified 225/56 split. Pancr. denotes the pancreas class. Bold rows indicate our methods.

Method	Pancr. Dice	Cancer Dice	Mean Dice↑	HD95 ↓
3D U-Net (Çiçek et al., 2016)	73.78	43.06	58.42	51.62
SegResNet (Myronenko, 2018)	72.34	42.35	57.35	53.92
3D UX-Net (Lee et al., 2022)	75.87	39.44	57.66	46.30
MedNeXt (Roy et al., 2023)	76.64	41.52	59.08	54.66
UNETR (Hatamizadeh et al., 2022)	67.19	27.80	47.50	64.48
Swin UNETR (Hatamizadeh et al., 2021)	74.83	41.31	58.07	61.09
SegMamba (Xing et al., 2024)	74.17	42.07	58.12	67.04
U-Mamba (Ma et al., 2024)	74.91	43.43	59.17	41.15
SegMamba-V2 (Xing et al., 2025)	75.64	45.05	60.35	49.98
SPAR-Net	76.44	44.18	60.31	36.86
SPAR-Net [8R-Rot]	**77.76**	**50.27**	**64.02**	**17.15**

Table 4: Single-patch efficiency on a 128^3 input using one RTX 3090. Latency excludes sliding-window aggregation.

Model	FLOPs (G)↓	Latency (ms)↓	Patch/s ↑	Peak (GB)↓
3D U-Net	**36.4**	**11.1**	**89.90**	**0.34**
3D UX-Net	2743.4	209.2	4.78	3.16
MedNeXt	336.8	115.1	8.69	2.90
UNETR	353.7	54.2	18.46	1.47
Swin UNETR	1467.6	192.4	5.20	3.92
U-Mamba	5617.2	126.7	7.89	2.41
SegMamba	2878.1	210.5	4.75	3.24
SegMamba-V2	4180.8	263.4	3.80	3.66
SPAR-Net (ours)	1140.2	153.5	6.51	1.88

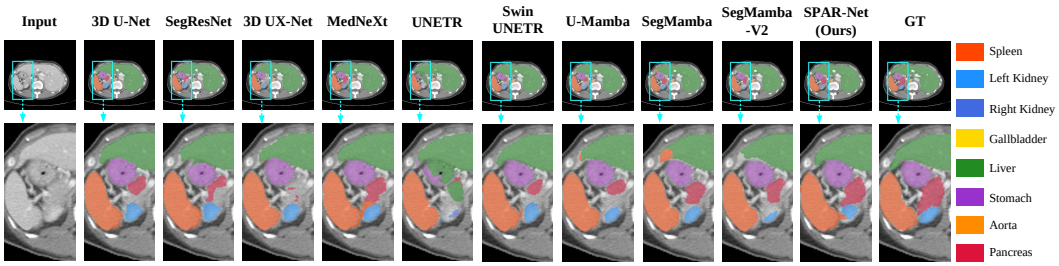


Figure 3: Qualitative comparison on Synapse. SPAR-Net produces complete masks and preserves thin boundaries in challenging low-contrast regions.

Datasets: Synapse contains 30 contrast-enhanced abdominal CT volumes with reference masks for eight organs. The labels were produced by trained interpreters under radiologist supervision. Following the common split (Hatamizadeh et al., 2021), we use 18 volumes for training and 12 for validation, with 128^3 training crops. MSD Pancreas (Task07) contains 281 contrast-enhanced CT scans from patients with pancreatic masses and provides pancreas and cancer masks (Antonelli et al., 2022). We use a fixed 225/56 train/validation split and 96^3 crops. The public releases do not provide per-case inter-rater agreement; we therefore use the released labels as the reference standard.

Unified baseline protocol: We implement all methods in PyTorch and MONAI and evaluate them with the same data split, intensity preprocessing, patch sampling, augmentation, loss, optimization schedule, checkpoint rule, and 50% overlapping sliding-window inference. All models are trained for 500 epochs on one NVIDIA RTX 3090. We use SGD with momentum 0.9, initial learning rate 10^{-3}, weight decay 3×10^{-5}, and a warmup–cosine schedule. The objective is the sum of soft Dice and cross-entropy losses. The best checkpoint is selected by validation Dice. This unified protocol is intended to compare architectures rather than numbers copied from heterogeneous publications; full settings are given in the supplementary material.

Metrics: We report mean foreground Dice and HD95. Dice measures volumetric overlap, while HD95 captures large surface deviations and is especially informative for thin or fragmented structures. Synapse values are reported as mean $\pm$ standard deviation across repeated runs. MSD Pancreas values are reported on the fixed validation split.

4.1 Comparison with Prior Architectures

Synapse: Table 2 shows a consistent trend. SPAR-Net [8R-Rot] achieves 85.25 Dice and 12.20 HD95, improving the strongest baseline by 1.83 Dice points and reducing mean HD95 by 16.14. Dice gains are the largest for the gallbladder and pancreas, while HD95 improves broadly across

Table 5: Scan-order comparisons.

No.	Scanning Strategy	Dice↑	HD95↓
1	Linear	83.19	31.51
2	Spiral-In	83.94	24.99
3	Stochastic-In	83.97	21.90
4	Spiral-Out	84.33	21.20
5	Stochastic-Out	84.45	19.82
6	Stochastic-Out + Spiral-Out	85.09	19.35
7	**Stochastic-In + Spiral-Out**	**85.25**	**12.20**

Table 6: SAFE scale ablation on Synapse.

No.	Patch streams	Dice↑	HD95↓
1	P_l only	80.02	13.51
2	$P_l + P_m$	83.36	19.04
3	P_s only	84.17	22.59
4	$P_s + P_m + P_l$	**85.25**	**12.20**

organs. This supports the role of SLS and BPFB in reducing discontinuous and displaced boundaries, not just recovering more foreground voxels.

MSD Pancreas: Table 3 tests the same design in a lesion setting. SPAR-Net [8R-Rot] reaches 77.76 Dice on the pancreas and 50.27 on cancer, for a 64.02 mean Dice. Relative to SegMamba-V2, the cancer Dice rises by 5.22 points and HD95 falls from 49.98 to 17.15. The larger improvement on cancer than on the pancreas is consistent with the scale-preserving multi-scale design: the fine stream protects small targets, while the coarser streams and radial paths provide the surrounding anatomical context needed to disambiguate weak tumor contrast.

Qualitative behavior: Figure 3 shows that the quantitative surface gains correspond to fewer fragmented masks and cleaner interfaces between adjacent organs. The improvement is most evident where the target is small, curved, or weakly contrasted. We interpret these visual results as evidence of better geometric consistency, not as a substitute for clinical validation.

Efficiency: Table 4 places the accuracy gains in context. SPAR-Net is not the fastest model, but it uses less peak memory than the evaluated high-capacity CNN and Mamba baselines and substantially fewer FLOPs than 3D UX-Net, Swin UNETR, U-Mamba, and both SegMamba variants. The result is a middle operating point: moderate latency and memory with the strongest segmentation accuracy under the unified protocol.

4.2 Ablation Study

Scan order: Table 5 changes only the traversal rule while holding the backbone and capacity fixed. Every radial alternative improves HD95 over linear scanning. A single stochastic outward path is the strongest one-path variant, while the inward stochastic and outward spiral pair gives the best overall result. The complementary directions matter: one path aggregates peripheral evidence toward the center, and the other redistributes central context toward the boundary.

Token scale: Table 6 shows that no single scale is uniformly sufficient. The fine stream is important for the pancreas and aorta, but the complete three-scale representation gives the best mean Dice and HD95. This result explains why SAFE precedes SLS: radial propagation is most effective when the sequence already contains both detailed and contextual evidence.

5 Discussion

Preserving evidence before propagation: The experiments support the ordering of operations that motivates SPAR-Net. SAFE first determines what evidence is available to the sequence model, and SLS then determines how that evidence travels. The scale ablation shows that a fine stream alone lacks context for larger structures, while coarse streams reduce overlap on boundary-sensitive organs. The scan-order comparison holds the token representation and backbone fixed and shows that radial paths consistently improve surface accuracy over linear serialization. Together, these results indicate that stronger state-space modeling cannot recover detail that was removed during tokenization, and richer tokenization is not sufficient when context is propagated through a poorly aligned sequence.

What radial scanning contributes: SLS should be interpreted as a soft geometric prior, not as a claim that points on the same shell are immediate neighbors. Its radial layers preserve distance to a stable reference and remove a single privileged Cartesian direction. The paired paths then expose the same token grid to opposite information flows: the inward path aggregates peripheral evidence toward the center, whereas the outward path redistributes central context toward object boundaries. Two observations isolate this effect. First, the paired inward–outward configuration improves HD95 substantially over every one-path strategy in the same backbone. Second, replacing the native scan in U-Mamba, SegMamba, and SegMamba-V2 reduces HD95 without changing their parameter counts. The recurring surface improvement across different encoders suggests that scan construction contributes independently of SAFE and BPFB.

Why paired traversal matters: A radial order remains directional when used alone. Stochastic-In gathers peripheral evidence toward the crop center, whereas Spiral-Out redistributes central context toward boundary layers. Each path is inverse-permuted before fusion, so complementary states are combined at matched spatial coordinates rather than matched sequence positions. The scan-order ablation supports this design: every radial one-path strategy improves on linear serialization, while the inward–outward pair gives the strongest joint Dice–HD95 result without changing the linear sequence complexity of the state-space operator.

Stability and failure profile: The repeated Synapse runs show that the final ranking is not produced by one favorable optimization seed: SPAR-Net [8R-Rot] reports 85.25 ± 0.20 Dice and 12.20 ± 3.34 HD95. The surface variance is markedly lower than the evaluated Mamba baselines, whose HD95 deviations range from 8.11 to 9.56 for SegMamba and SegMamba-V2, while the Dice variance remains small. The organ-wise results also reveal where the prior is less helpful. SPAR-Net does not improve the aorta HD95 over the best baseline, and the rotational variant sacrifices part of the spleen advantage obtained by the base scan. These exceptions are informative: a center-referenced traversal is not uniformly optimal for every geometry, and broader angular coverage can trade organ-specific performance for a stronger mean result. The base and rotational variants should therefore be viewed as two operating points rather than one universally dominating the other. This profile motivates future adaptive routing that selects the reference, shell width, or angular schedule from the feature geometry while retaining deterministic coordinate restoration.

Accuracy-efficiency trade-off: SPAR-Net prioritizes boundary-robust segmentation over minimum latency. Its multi-scale streams and dual radial paths increase computation relative to a minimal Mamba encoder, but FLOPs and peak memory remain below several high-capacity volumetric baselines. This trade-off preserves multi-scale evidence and coordinate-consistent propagation, yielding the clearest return in HD95 and small-target Dice. Table 4 shows these gains at moderate computational cost rather than claiming minimum latency.

Determinism and limitations: The stochastic component is used only for scan construction. A seeded permutation is generated once and reused across training, validation, and inference, ensuring deterministic behavior and coordinate consistency. The geometric crop center avoids organ-specific priors but may be suboptimal for eccentric or peripheral targets. The 92.36M-parameter model is also relatively heavy. Future work will explore adaptive reference estimation, learned traversal, and lighter fusion. Since Dice and HD95 measure geometric agreement, our conclusions are limited to segmentation accuracy, boundary quality, and computational efficiency.

6 Conclusion

SPAR-Net follows a simple principle: preserve spatial evidence before deciding how it should propagate. SAFE aligns fine-to-coarse receptive fields on a shared high-resolution token grid, SLS replaces fixed Cartesian serialization with complementary radial paths, and FR with BPFB restores voxel structure while reconciling semantic and boundary information. On Synapse and MSD Pancreas, this design improves both mean Dice and HD95, with the clearest gains on small, irregular, and low-contrast targets. Across component, scan-order, transfer, scale, and efficiency analyses, scale-preserving tokenization strengthens the state-space representation, while coordinate-aligned radial propagation maintains more continuous 3D boundaries.

AI USE STATEMENT

Generative AI tools were used to support manuscript editing, literature and bibliographic verification, and consistency checks among the described methodology, reported results, and released code. All research decisions, experimental design, implementation, results, and conclusions were independently reviewed, verified, and finalized by the authors.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Ao Chang, Jiajun Zeng, Ruobing Huang, and Dong Ni. Em-net: Efficient channel and frequency learning with mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 266–275. Springer, 2024.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. 3d transunet: Advancing medical image segmentation through vision transformers. *arXiv preprint arXiv:2310.07781*, 2023.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

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

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022.

Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6202–6212, 2023.

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.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, pp. 12. Munich, Germany, 2015.

Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076*, 2022.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pp. 615–625. Springer, 2024a.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024b.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, pp. 311–320. Springer, 2018.

Thang Quoc Pham, Thai Hoang Le, Khai Dinh Lai, Dat Quoc Ngo, Tan Van Pham, Quang Hong Hua, Khang Quang Le, Huyen Duy Mai Le, and Tuyen Ngoc Lam Nguyen. Caaf-resunet: Adaptive attention fusion with boundary-aware loss for lung nodule segmentation. *Medicina*, 61(7): 1126, 2025.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 405–415. Springer, 2023.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Hongyi Wang, Yingying Xu, Qingqing Chen, Ruofeng Tong, Yen-Wei Chen, Hongjie Hu, and Lanfen Lin. Adaptive decomposition and shared weight volumetric transformer blocks for efficient patch-free 3d medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4854–4865, 2023.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 578–588. Springer, 2024.

Zhaohu Xing, Tian Ye, Yijun Yang, Du Cai, Baowen Gai, Xiao-Jian Wu, Feng Gao, and Lei Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Jin Yang, Peijie Qiu, Yichi Zhang, Daniel S Marcus, and Aristeidis Sotiras. D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation. *Biomedical Signal Processing and Control*, 113:108837, 2026.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Appendix

In this supplementary material, we provide additional architectural details, algorithmic formulations, and experiments to complement the main paper. The contents are organized as follows:

- § A presents the architecture and algorithmic formulation of the SLS Block, including parallel scanning, dual-branch Mamba modeling, and hierarchical fusion.
- § B describes the datasets Antonelli et al. (2022), Hatamizadeh et al. (2022), preprocessing, training hyperparameters, implementation details, evaluation metrics, and quantitative/qualitative analyses on Synapse and MSD Pancreas.
- § C reports additional ablation studies, including plug-and-play integration of SLS into Mamba-based backbones and the impact of different layer-width configurations.

Together, these sections summarize the design of SPAR-Net and assess the robustness of the proposed layer-aware scanning framework across datasets and architectures.

A SLS Block: Detailed Architecture

This section provides additional architectural details of the SLS Block, which serves as the core sequence modeling module in SPAR-Net. Here, layer-aware sequence modeling refers to serializing a 3D feature volume into a token sequence whose ordering follows the underlying anatomical layout, rather than the arbitrary Cartesian axes typically used in prior state-space models. Purely axis-aligned flattening (e.g, D/H/W scanning) will break local spatial continuity, place anatomically distant voxels adjacently in the sequence, and introduce strong directional bias, all of which hinder effective state propagation in recurrent state-space layers.

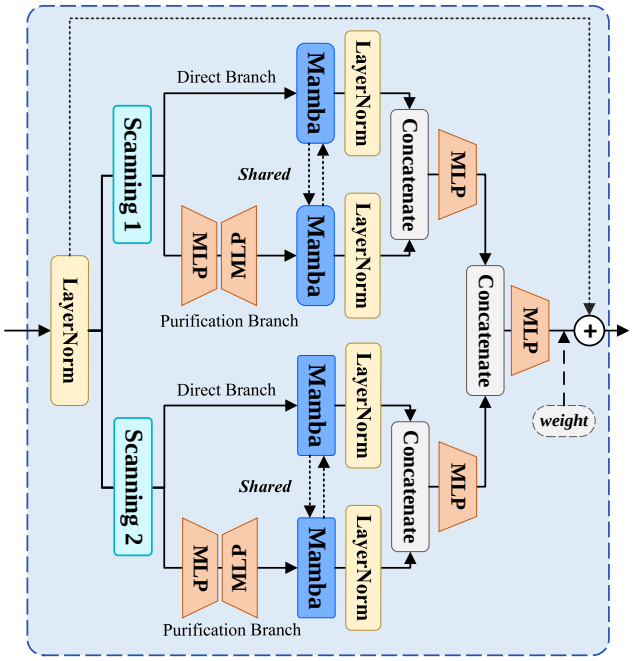


Figure 4: Architecture of the SLS Block. Two layer-aligned scanning paths (*e.g.*, Stochastic-In and Spiral-Out) generate complementary token sequences, each processed by a shared-weight dual-branch Mamba module. The outputs are fused hierarchically—first within each path, then across paths—and combined with the input through a weighted residual connection.

The SLS Block addresses these issues by reorganizing volumetric features into concentric radial layers centered around an anatomically meaningful reference point. This layer construction explicitly encodes near–mid–far relationships and groups tokens by geometric proximity instead of axis

order. On top of this layered representation, the SLS Block applies structured traversal strategies (*Stochastic-In* and *Spiral-Out*) that follow anatomically coherent paths: inward traversal aggregates boundary-informed cues toward the center, while outward traversal propagates high-level context back to peripheral regions. Together, these traversal schemes establish spatially consistent token neighborhoods that support long-range dependency modeling along anatomically meaningful directions.

To further enhance spatial coherence, we decompose the proposed SLS mechanism into two complementary scanning variants.

Global Stochastic Layer Scanning partitions the volume into a sequence of concentric radial layers whose thickness is determined by the layer-width configuration, rather than a fixed number of regions. Random permutations are applied independently within each layer to introduce spatial diversity while preserving local adjacency. Across layers, the traversal direction can be set to center-out or outer-in, enabling controllable long-range paths that better reflect anatomical structure than conventional linear sweeps.

Regional Rotational Stochastic Layer Scanning further subdivides each layer into multiple angular sectors to better handle small organs and regions with high local variability. Random permutations are applied within each sector, and rotational offsets are introduced across layers to prevent fixed sector boundaries. This rotation mitigates boundary fragmentation and allows the scanning path to capture organ shape from multiple viewing angles.

Both variants contribute complementary forms of spatial coherence, and their quantitative benefits are demonstrated in the main paper's *Quantitative Comparison on Synapse* and in the supplementary results reported in Table 9.

In the following, we describe how parallel scanning, dual-branch Mamba modeling, and hierarchical fusion are combined within the SLS Block to form a unified, spatially consistent representation.

Parallel Scanning Strategy. As indicated by the ablation study in the main paper, the best performance is obtained by combining two spatially coherent traversal strategies: *St-In* (Stochastic-In) and *Sp-Out* (Spiral-Out). Given a layer-aligned token map $X \in \mathbb{R}^{N\times C}$, we construct two ordered sequences by applying separate scan operators:

$$X^{(1)} = \mathrm{Scan}_{St-In}(X), \quad X^{(2)} = \mathrm{Scan}_{Sp-Out}(X), \tag{7}$$

where $\mathrm{Scan}_{St-In}(\cdot)$ performs inward traversal with per-layer stochastic permutation, and $\mathrm{Scan}_{Sp-Out}(\cdot)$ applies outward traversal along a deterministic spiral pattern. The former emphasizes boundary-aware aggregation toward the center, while the latter propagates high-level context toward the periphery and improves coverage of the full organ extent.

Dual-branch Mamba Modeling. Each scanned sequence $X^{(k)}$ is processed by the dual-branch Mamba module defined in the main paper. The direct branch operates on the original sequence to preserve fine structural detail, whereas the purification branch passes the features through a reduce–expand MLP before Mamba, yielding a more compact semantic representation. Both branches share the same Mamba parameters, ensuring consistent state dynamics while separating structural detail from high-level abstraction.

Hierarchical Fusion Across Branches and Paths. The SLS Block employs a two-stage fusion scheme designed to (i) merge local and global cues within each traversal path and (ii) consolidate the complementary information from the two paths. For each path $k \in \{1, 2\}$, the direct and purification outputs $Z^{(k)}_{dir}$ and $Z^{(k)}_{pur}$ are first concatenated and fused by a lightweight MLP:

$$U^{(k)} = \mathrm{MLP}\Big(\mathrm{LN}\big([Z^{(k)}_{dir}, Z^{(k)}_{pur}]\big)\Big), \tag{8}$$

producing a path-specific representation that combines high-frequency boundary detail with compact semantics. The two path-level features $\{U^{(1)}, U^{(2)}\}$ are then concatenated and fused:

$$Z_{fused} = \mathrm{MLP}\Big(\mathrm{LN}\big([U^{(1)}, U^{(2)}]\big)\Big), \tag{9}$$

which integrates inward-aggregating and outward-propagating dependencies into a single spatially consistent sequence and empirically reduces directional bias, especially for small or asymmetric organs.

Algorithm 1 Stochastic Layer Scanning Block (SLS Block, depth T)

Require: Consistent token sequence $F_{tok} \in \mathbb{R}^{N\times C}$; scan indices $\mathbf{i}^{SI}, \mathbf{i}^{SO} \in \{1,\dots,N\}$ for *Stochastic-In* and *Spiral-Out*; block depth T; learnable weight α;
Ensure: Refined token sequence $Z_{out} \in \mathbb{R}^{N\times C}$;

$X \leftarrow F_{tok}$; ▷ canonical adjacency-preserving order
$X^{(2)} \leftarrow \text{Permute}(X, \mathbf{i}^{SO})$; ▷ Spiral-Out path
for $t \leftarrow 1$ to T **do** ▷ depth (*e.g.*, T=3)
  for $k \leftarrow 1$ to 2 **do** ▷ two parallel scanning paths
    $Z_{dir}^{(k)} \leftarrow \mathrm{LN}\big(\mathrm{Mamba}(\mathrm{LN}(X^{(k)}))\big)$;
    $Y^{(k)} \leftarrow \mathrm{MLP}_{\downarrow}(X^{(k)})$;
    $Y^{(k)} \leftarrow \mathrm{MLP}_{\uparrow}(Y^{(k)})$;
    $Z_{pur}^{(k)} \leftarrow \mathrm{LN}\big(\mathrm{Mamba}(\mathrm{LN}(Y^{(k)}))\big)$;
    $U^{(k)} \leftarrow \mathrm{MLP}\big(\mathrm{LN}([\,Z_{dir}^{(k)},\, Z_{pur}^{(k)}\,])\big)$;
    $X^{(k)} \leftarrow U^{(k)}$; ▷ update representation of path k
  end for
end for
$Z_{fused} \leftarrow \mathrm{MLP}\big(\mathrm{LN}([\,U^{(1)},\, U^{(2)}\,])\big)$; ▷ fusion in scanning space
$\tilde{Z} \leftarrow X + \alpha\,\text{InversePermute}(Z_{fused}, \mathbf{i}^{SI})$; ▷ map back to canonical order and apply weighted residual
$Z_{out} \leftarrow \tilde{Z}$;

Table 7: Summary of datasets and preprocessing configurations used in our experiments.

Dataset	Anatomical Region	Modality	#Samples	Data Splits	Pre-processing
Synapse Hatamizadeh et al. (2022)	Abdomen	CT	30	Train: 18 Validation: 12	1. Interpolate scans to isotropic voxel spacing of [1.5×1.5×2.0] mm. 2. Clip intensities using soft tissue window [−175, 250], normalize to [0, 1]. 3. Random sampling of 128 ×128×128 voxels (1:1 positive-to-negative ratio). 4. Augmentations: random flip (0.2), rotation (0.2), intensity shifting (0.1), and intensity scaling (0.1). 5. Labels: Spleen, Left kidney, Right kidney, Gallbladder, Liver, Stomach, Aorta, and Pancreas.
MSD **Pancreas Antonelli et al. (2022)**	Pancreas	CT	281	Train: 225 Validation: 56	1. Clip intensities to [−87, 199] and normalize to [0, 1]. 2. Sample patches of 96×96×96 voxels with a 1:1 positive-to-negative ratio. 3. Augmentations: random flip (0.5), rotation (0.25), intensity scaling (0.5), and shifting (0.5). 4. Labels: Pancreas, Cancer.

Weighted Residual Integration. To stabilize optimization and avoid over-amplification from dual-path fusion, the fused representation is mapped back to the canonical token order and combined with the input via a learnable residual:

$$Z_{out} = X + \alpha\, Z_{fused}, \tag{10}$$

where α is a trainable scalar that controls the update magnitude. This residual design preserves a strong identity path while allowing the fused sequence to refine spatially consistent features in a controlled manner, reducing convergence variance at deeper encoder stages.

Algorithmic Formulation. Algorithm 1 summarizes the full computation flow of the SLS Block, including parallel layer-wise traversal, dual-branch Mamba modeling, and hierarchical fusion. The algorithm starts from the layer-aligned token sequence F_{tok}, applies *Stochastic-In* and *Spiral-Out* permutations to construct two traversal paths, and iteratively updates the path-specific features through shared-weight direct and purification branches. The final fusion step aggregates the two paths in the scanning space, followed by an inverse permutation and a weighted residual update in the canonical token order. All symbols and module definitions (*e.g.*, Mamba, reduce–expand MLP, scan indices) are consistent with the notation introduced in Section 3 of the main paper.

Table 8: Hyperparameters used for the Synapse and MSD Pancreas datasets.

Hyperparameters	**Synapse**	**MSD Pancreas**
Encoder Stage	4	4
Original Volume Size	$128 \times 512 \times 512$	$96 \times 512 \times 512$
Layer-wise Channel Size	[48, 96, 192, 384]	[48, 96, 192, 384]
Patch Size (Training)	$128 \times 128 \times 128$	$96 \times 96 \times 96$
Layer Depth of SLS Block	[3, 3, 3, 3]	[3, 3, 3, 3]
Per-layer Patch Size	{[2,4,8], [2,4,8], [1,2,4], [1,2]}	{[2,4,8], [1,2,4], [1,2,4], [1,2]}
Number of Sub-volumes Cropped	4	4
Max Epochs	500	500
Optimizer	SGD	SGD
Learning Rate	0.001	0.001
Weight Decay	3×10^{-5}	3×10^{-5}
LR Scheduler	Warmup–Cosine	Warmup–Cosine
Loss Function	Dice + CrossEntropy	Dice + CrossEntropy

B Experimental Details

B.1 Datasets

Synapse Hatamizadeh et al. (2022). The Synapse multi-organ CT dataset contains 30 contrast-enhanced abdominal volumes with 80–225 slices per scan, each annotated for 13 organs by trained interpreters under radiologist supervision from Vanderbilt University Medical Center. Following the common protocol in prior work, we use 18 scans for training and 12 for validation, and focus on voxel-wise segmentation of eight major abdominal organs: spleen, left kidney, right kidney, gallbladder, liver, stomach, aorta, and pancreas. Pre-processing and augmentation configurations are listed in Table 7.

MSD Pancreas Antonelli et al. (2022). The MSD Pancreas Tumor dataset (Task07 of the Medical Segmentation Decathlon) comprises 281 contrast-enhanced abdominal CT scans from the Memorial Sloan Kettering Cancer Center, annotated with two classes: pancreas and cancer. The task is challenging due to severe class imbalance between large background regions, medium-sized pancreas structures, and small tumor lesions. All volumes are resampled to an isotropic spacing and cropped into fixed-size 3D patches for training; the key pre-processing and sampling settings are listed in Table 7.

B.2 Implementation Details

We implemented all models using PyTorch and MONAI[1], and train them on a single NVIDIA RTX 3090 GPU with 24 GB memory. For all datasets, we use an initial learning rate of 1×10^{-3} and train each model for 500 epochs. The best checkpoint is selected based on the highest validation Dice score.

We employed a hybrid segmentation loss that balances volumetric overlap and voxel-wise classification accuracy. The final loss is defined as:

$$\mathcal{L} = \mathcal{L}_{Dice} + \mathcal{L}_{CE}, \tag{11}$$

where $\mathcal{L}_{Dice}$ denotes the soft Dice loss that maximizes region-overlap consistency, and $\mathcal{L}_{CE}$ denotes the voxel-wise Cross-Entropy loss that stabilizes per-pixel classification.

All training hyperparameters, including optimizer configuration, learning-rate scheduling, layer depth of the SLS Block, patch sizes for each dataset, and the per-layer multi-scale patch settings, are summarized in Table 8. We adopted the SGD optimizer with momentum and a warmup–cosine

[1] https://monai.io/

Table 9: Organ-wise HD95 results on the Synapse multi-organ CT dataset; lower is better. Values are reported as mean ± standard deviation across repeated runs. Bold and underlined values indicate the best and second-best results, respectively. SPAR-Net [8RR] uses eight-region rotational stochastic layer scanning. Spl., L.Kid., R.Kid., Gall., Liv., Sto., Aor., and Panc. denote the spleen, left kidney, right kidney, gallbladder, liver, stomach, aorta, and pancreas, respectively.

Method	**Spl.**	**L.Kid.**	**R.Kid.**	**Gall.**	**Liv.**	**Sto.**	**Aor.**	**Panc.**	**HD95↓**
3D U-Net Çiçek et al. (2016)	51.97 ±5.02	19.11 ±1.63	7.60 ±0.23	24.94 ±17.97	20.79 ±2.00	31.39 ±1.07	<u>4.27</u> ±0.77	15.53 ±3.51	21.95 ±2.83
SegResNet Myronenko (2018)	58.81 ±21.60	36.17 ±10.76	13.67 ±5.61	48.35 ±4.43	40.75 ±2.54	41.91 ±7.06	10.73 ±10.44	20.99 ±7.21	33.92 ±3.23
3D UX-Net Lee et al. (2022)	52.92 ±27.06	50.10 ±20.69	15.26 ±6.96	27.83 ±13.02	40.37 ±2.08	58.34 ±11.39	9.65 ±1.07	20.84 ±5.73	34.41 ±2.08
MedNeXt Roy et al. (2023)	83.80 ±28.07	28.98 ±6.97	12.83 ±8.10	26.71 ±19.17	43.70 ±5.28	39.92 ±2.62	8.75 ±1.67	16.40 ±5.96	32.64 ±6.05
UNETR Hatamizadeh et al. (2022)	43.88 ±20.62	36.57 ±23.88	42.39 ±5.88	51.51 ±16.32	45.09 ±11.29	29.88 ±7.48	26.00 ±11.93	29.06 ±7.01	38.05 ±4.63
Swin UNETR Hatamizadeh et al. (2021)	64.19 ±25.16	43.93 ±4.42	28.10 ±5.58	35.69 ±35.97	30.11 ±3.14	43.82 ±13.05	12.09 ±3.83	34.98 ±10.09	36.61 ±7.09
SegMamba Xing et al. (2024)	50.08 ±6.24	42.04 ±40.80	15.49 ±3.18	45.70 ±16.80	33.96 ±3.01	36.62 ±9.43	7.85 ±3.04	13.78 ±0.75	30.69 ±8.11
U-Mamba Ma et al. (2024)	57.87 ±15.75	33.16 ±24.61	14.36 ±7.27	58.41 ±6.94	27.39 ±6.96	40.75 ±7.40	5.60 ±0.99	14.47 ±0.81	31.50 ±2.32
SegMamba-V2 Xing et al. (2025)	43.83 ±33.09	34.94 ±17.40	16.57 ±12.88	41.60 ±38.52	32.75 ±9.02	39.17 ±2.43	**4.15** ±0.53	<u>13.72</u> ±3.94	28.34 ±9.56
SPAR-Net	**13.27** ±2.84	<u>13.64</u> ±9.66	<u>4.51</u> ±1.66	<u>13.52</u> ±10.71	<u>15.72</u> ±1.01	<u>24.72</u> ±8.66	4.64 ±1.55	13.91 ±1.42	<u>12.99</u> ±3.29
SPAR-Net [8RR]	<u>23.02</u> ±19.40	**13.48** ±9.51	**4.49** ±3.23	**7.66** ±0.33	**13.17** ±1.06	**19.14** ±0.10	4.66 ±1.73	**11.98** ±1.46	**12.20** ±3.34

learning rate schedule. For Synapse, we cropped input volumes into 128^3 patches, while for the MSD Pancreas dataset, we used 96^3 patches. Both tasks use four encoder stages with channel dimensions $[48, 96, 192, 384]$, SLS Block depths of $[3, 3, 3, 3]$, and four cropped sub-volumes per iteration.

B.3 Evaluation Metrics

We evaluated segmentation performance using two standard metrics: the Dice similarity coefficient (Dice) and the 95th percentile Hausdorff distance (HD95). Dice measures volumetric overlap between the predicted mask $\hat{Y}$ and the ground truth Y:

$$\mathrm{Dice}(Y, \hat{Y}) = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \times 100. \tag{12}$$

To assess boundary accuracy, we computed the Hausdorff distance between the two surfaces and report its 95th percentile (HD95):

$$\mathrm{HD}(Y, \hat{Y}) = \max\left\{ \max_{y \in Y} \min_{\hat{y} \in \hat{Y}} d(y, \hat{y}),\ \max_{\hat{y} \in \hat{Y}} \min_{y \in Y} d(y, \hat{y}) \right\}, \tag{13}$$

where $d(y, \hat{y})$ denotes the Euclidean distance between boundary points y and $\hat{y}$. HD95 is obtained by taking the 95th percentile of the bidirectional surface distances. Dice primarily reflects region-level agreement, while HD95 is more sensitive to boundary deviations; together, they provide a balanced view of volumetric and contour quality.

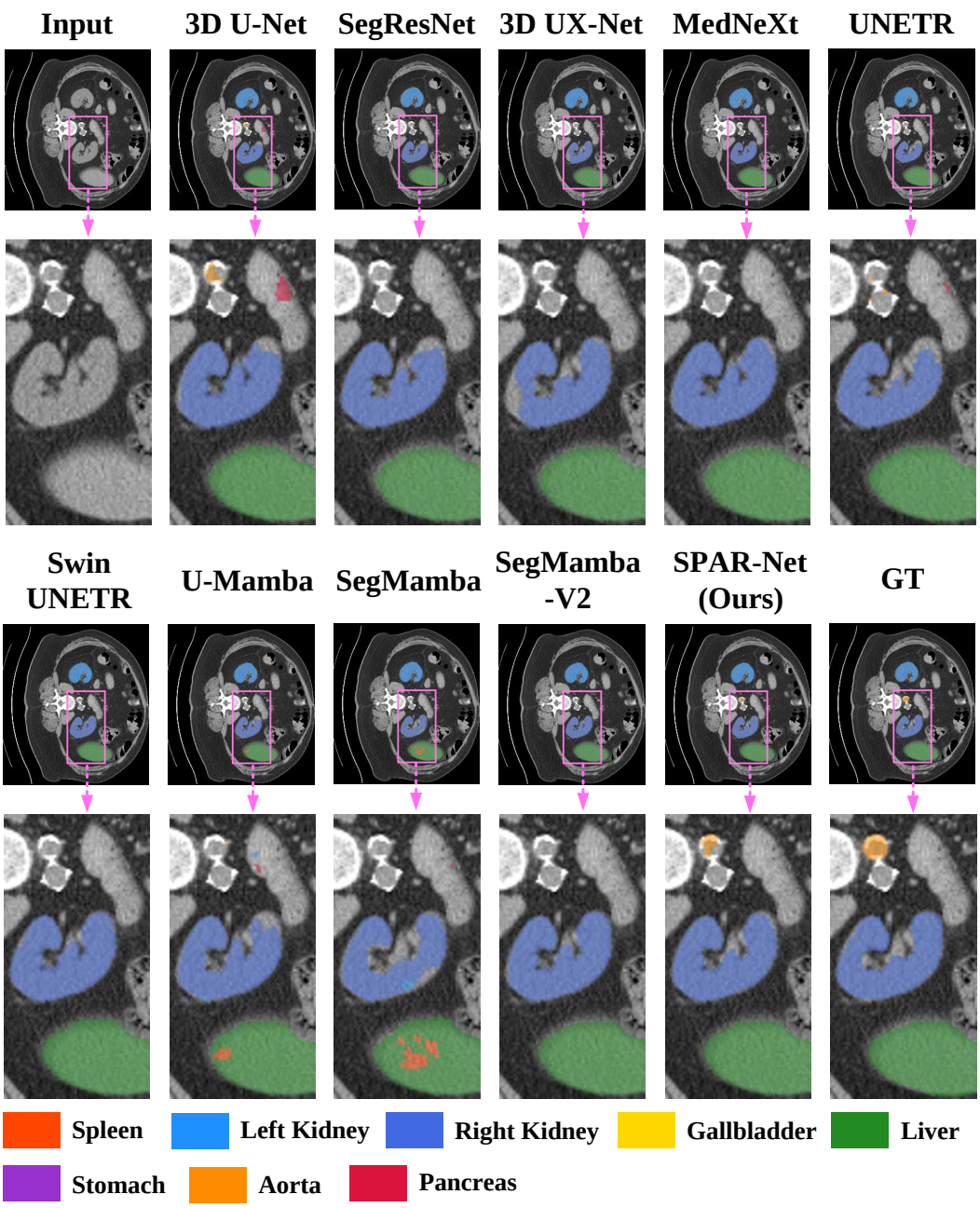


Figure 5: Qualitative comparison on the Synapse dataset Hatamizadeh et al. (2022). SPAR-Net produces sharper, more continuous organ boundaries and better preserves small, low-contrast structures than prior methods.

B.4 HD95 Analysis on Synapse

Table 9 reports the organ-wise HD95 scores on the Synapse dataset, complementing the Dice results in the main paper. HD95 is particularly sensitive to boundary errors and thus highlights structure consistency, which is a known weakness of axis-aligned scanning strategies.

CNN- and Transformer-based baselines exhibit large boundary deviations for several organs, especially the gallbladder and pancreas, where thin or low-contrast structures are common. Mamba-

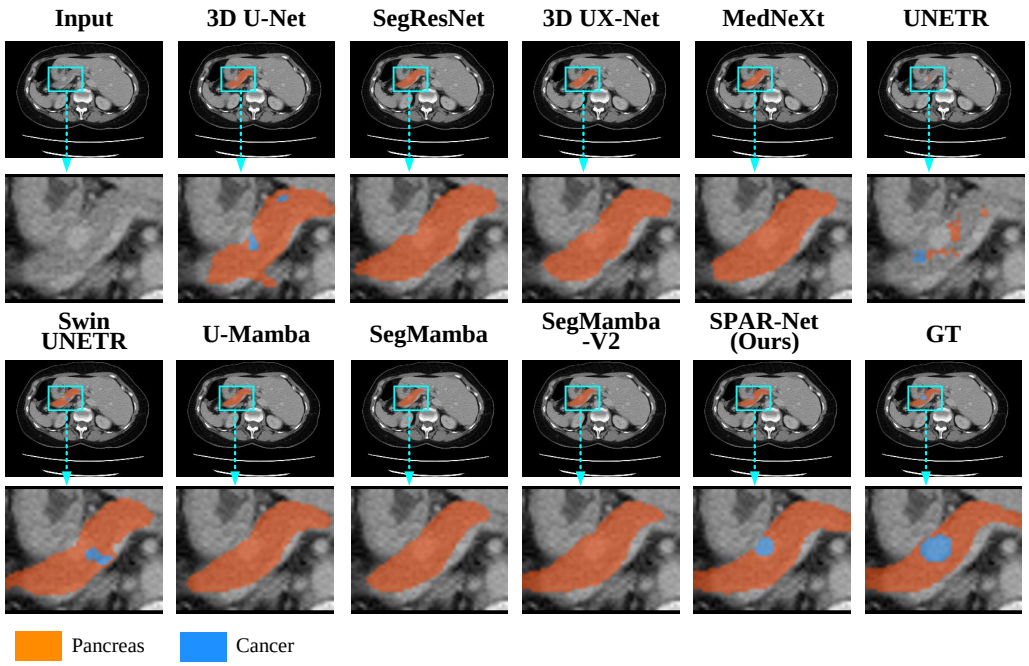


Figure 6: Qualitative comparison on the MSD Pancreas dataset Antonelli et al. (2022). SPAR-Net better recovers the full pancreas anatomy and detects low-contrast tumors more reliably, while prior methods often yield fragmented boundaries or miss the lesion.

based variants generally improve spatial boundary consistency but still show inconsistent performance across organs, reflecting the limitations of strictly axis-wise token serialization.

Both SPAR-Net variants substantially reduce HD95 compared with prior methods. The base SPAR-Net already achieves sizeable improvements for the spleen, kidneys, and aorta, indicating more stable boundary reconstruction. The enhanced SPAR-Net [8RR] further lowers the mean HD95 to 12.20 and attains the best or second-best score in seven out of eight organs. These results suggest that the spatially coherent traversal not only improves average boundary accuracy but also reduces outlier errors, particularly in small or highly curved structures.

B.5 Qualitative Comparison on Synapse

Figure 5 presents qualitative results on representative axial slices from the Synapse dataset Hatamizadeh et al. (2022). Classical CNNs (3D U-Net, SegResNet, 3D UX-Net) tend to blur organ interfaces and frequently miss fine or low-contrast regions, particularly around the pancreas and gallbladder. Transformer-based models (UNETR, Swin UNETR) yielthe edges of organs predictions but often exhibit discontinuities along organ edges.

Recent Mamba-based architectures (U-Mamba, SegMamba, SegMamba-V2) improve global consistency, yet they still undersegment small organs or produce fragmented pancreas masks in challenging cases. By comparison, SPAR-Net generates more anatomically coherent segmentations, with sharper boundaries, clearer separation between adjacent organs, and more complete reconstructions of small structures. These observations are consistent with the quantitative HD95 improvements reported earlier.

B.6 Qualitative Comparison on MSD Pancreas

Figure 6 shows qualitative results on the MSD Pancreas dataset Antonelli et al. (2022), which is challenging due to the small tumor size, low contrast against surrounding tissue, and irregular pancreas shape. CNN-based models (3D U-Net, SegResNet) often produce overly smooth or incomplete pan-

creas masks and frequently miss the tumor. Transformer-based approaches (UNETR, Swin UNETR) capture broader context but exhibit unstable delineation along thin or curved regions.

Mamba-based models (U-Mamba, SegMamba, SegMamba-V2) offer improved global modeling, yet still suffer from fragmented tumor predictions and boundary discontinuities. In contrast, SPAR-Net maintains coherent pancreas contours and preserves fine-scale tumor geometry more consistently. The combination of spatially consistent token serialization and dual-branch state-space modeling appears particularly beneficial for these low-contrast, small-lesion scenarios, aligning with the quantitative gains reported for MSD Pancreas.

C Additional Ablation Study

Table 10: Component ablations on Synapse.

No.	SPAR module	BPFB	Dice↑	HD95↓	Params (M)
1	✗	✗	82.72	24.27	41.22
2	✗	✓	83.12	19.40	52.24
3	✓	✗	84.25	23.39	81.35
4	✓	✓	**85.25**	**12.20**	92.36

C.1 Component ablations on Synapse

Table 10 separates the encoder module from cross-scale fusion. The SPAR module contributes the larger Dice increase, whereas BPFB has a stronger effect on HD95. Their combination improves both metrics, indicating distinct roles: scale-aligned state-space modeling strengthens the representation, and bidirectional fusion stabilizes the reconstructed boundaries. A step-wise decomposition in the supplement reaches the same conclusion.

C.2 Plug-and-Play Generalization of SLS

Table 11 evaluates the generality of the SLS mechanism by integrating it into representative Mamba-based segmentation backbones, including U-Mamba, SegMamba, and SegMamba-V2. Across architectures, SLS acts as a lightweight, plug-and-play enhancement that consistently improves geometric coherence.

U-Mamba shows the largest overall gain, with Dice increasing by 1.50 and HD95 decreasing by 0.78, indicating that layer-aware stochastic traversal can strengthen long-range spatial reasoning without increasing the parameter count. For SegMamba, the Dice change is negligible (-0.08), but HD95 drops by 10.46, suggesting markedly better boundary fidelity despite similar region-level overlap. SegMamba-V2 also benefits from SLS, with an 11.74 reduction in HD95 and a slight Dice improvement.

Overall, these results support that SLS can be seamlessly incorporated into existing Mamba-based 3D segmentation networks to obtain more structure-consistent and spatially coherent representations across organ scales.

Table 11: Effect of integrating the proposed SLS into existing Mamba-based backbones. We report mean Dice (%) and HD95 (↓). Each method is evaluated after incorporating SLS, with values in parentheses indicating performance changes relative to the original baseline (positive ↑ for Dice, negative ↓ for HD95).

Method	Dice ↑	HD95 ↓
U-Mamba **w/ SLS**	84.18 (+1.50)	30.16 (-0.78)
SegMamba **w/ SLS**	82.64 (-0.08)	19.82 (-10.46)
SegMamba-V2 **w/ SLS**	83.52 (+0.10)	16.23 (-11.74)
SPAR-Net (ours)	**85.25**	**12.20**

Table 12: Ablation of different layer-width configurations in the SLS Block. Wider layers capture broader semantic context, whereas narrower layers emphasize boundary-sensitive details.

No.	Layer Width per Stage	Dice ↑	HD95 ↓
1	[2.0, 2.0, 2.0, 2.0]	84.61	20.81
2	[3.0, 2.0, 1.5, 1.5]	85.14	13.57
3	[3.0, 2.0, 2.0, 2.0]	85.22	8.87
4	[4.0, 3.0, 3.0, 3.0]	85.24	14.13
5	[1.0, 1.0, 1.0, 1.0]	85.34	**8.62**
6	**[3.0, 3.0, 3.0, 3.0]**	**85.25**	12.20

C.3 Layer-Width Configurations of the SLS Block

Table 12 reports an ablation study on the layer-width configuration in the SLS Block. Each value controls the radial thickness of concentric layers at a given encoder stage. Narrow layers (*e.g.*, No. 5: [1.0, 1.0, 1.0, 1.0]) emphasize fine anatomical boundaries and achieve the lowest HD95, but their limited receptive field slightly restricts Dice performance. In contrast, wider layers (No. 1 and No. 4) aggregate broader context but dilute boundary cues, leading to larger HD95 errors. Among all tested settings, the uniform mid-range configuration [3.0, 3.0, 3.0, 3.0] (No. 6) offers the best Dice with competitive HD95, indicating a stable compromise between contextual aggregation and boundary fidelity.

These results highlight that balanced layer widths are important for layer-aware sequence modeling: excessively narrow or wide layers both introduce suboptimal spatial trade-offs, whereas a moderate, stage-consistent width yields more reliable segmentation results.