



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

MM-UNet: Meta Mamba UNet for Volumetric Medical Image Segmentation

Bin Xie¹, Yan Yan², and Gady Agam¹

¹ Department of Computer Science, Illinois Institute of Technology, Chicago, USA
bxie9@hawk.iit.edu

² Department of Computer Science, University of Illinois Chicago, Chicago, USA

Abstract. State Space Models (SSMs) have recently demonstrated outstanding performance in long-sequence modeling, particularly in natural language processing. However, their direct application to medical image segmentation poses several challenges. SSMs, originally designed for 1D sequences, struggle with 3D spatial structures in medical images due to discontinuities introduced by flattening. Additionally, SSMs have difficulty fitting high-variance data, which is common in medical imaging. In this paper, we analyze the intrinsic limitations of SSMs in medical image segmentation and propose a unified U-shaped encoder-decoder architecture, Meta Mamba UNet (MM-UNet), designed to leverage the advantages of SSMs while mitigating their drawbacks. MM-UNet incorporates hybrid modules that integrate SSMs within residual connections, reducing variance and improving performance. Furthermore, we introduce a novel bi-directional scan order strategy to alleviate discontinuities when processing medical images. Extensive experiments on AMOS22, ACDC, and BTCV datasets demonstrate the superiority of MM-UNet over state-of-the-art methods. MM-UNet achieves a Dice score of 91.5% and 89.0% on AMOS22 task 1 and task 2, surpassing nnUNet by 1.7% and 1.0%, 93.2% Dice on ACDC, and 87.1% Dice on BTCV. These results confirm the effectiveness of integrating SSMs in medical image segmentation through architectural design optimizations. We release our code at <https://github.com/bxie9/MM-UNet>.

Keywords: Medical Image Segmentation · Mamba

1 Introduction

Medical image segmentation plays a crucial role in biomedical image analysis, aiding in disease diagnosis, abnormality detection, and surgical planning. In recent years, significant progress has been achieved in medical image segmentation through deep learning-based methods [14, 7, 5, 17], with convolutional neural networks (CNNs) and vision transformers (ViTs) emerging as dominant architectures. However, CNNs are inherently limited by their local receptive fields, making them less effective at capturing long-range dependencies. ViTs suffer from high computational complexity due to their quadratic attention mechanism, which restricts their ability to efficiently model global dependencies.

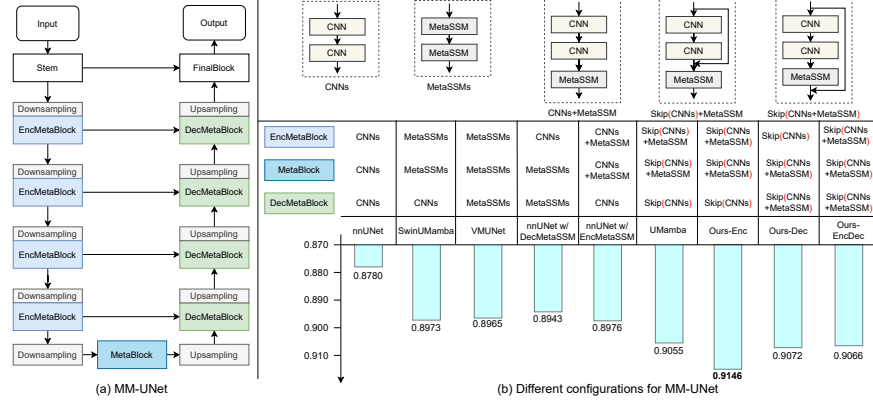


Fig. 1: (a) Overview of our proposed MM-UNet architecture. (b) Experiments replacing meta blocks in MM-UNet with different modules, including pure CNN-based, hybrid, and pure SSM-based modules. Skip connections represent residual connections.

Recently, State Space Models (SSMs) have gained significant attention in natural language processing (NLP) due to their ability to model long sequences with linear time complexity. Their success has also influenced the field of computer vision. This raises the challenge of determining whether SSMs, originally designed for 1D sequences, can effectively handle images with 2D or even 3D spatial structures, particularly in medical imaging?

To investigate this, we first explore the intrinsic limitations of SSMs in image-based tasks. Typically, SSMs process images by flattening them in a predefined order, such as width-first (H-W order). We conduct experiments using the Structured State Space Sequence Model (S4) [3] to fit 1D sequences derived from 2D medical images (see Fig. 2(a)). The results indicate that S4 struggles to fit points that deviate significantly from the mean, particularly those with high variance. Additionally, flattening images into a 1D sequence introduces discontinuities, making it difficult for SSMs to infer relationships across adjacent rows. However, these discontinuities can be mitigated using an inverse scan order, which enhances the ability of SSMs to model spatial structures more effectively.

Building on the advancements of Transformers in medical image segmentation (e.g., SwinUNETR [16], nnFormer [17]), several U-shaped architectures incorporating Mamba have been developed. These models take inspiration from Transformer-based segmentation approaches, including SegMamba, which employs a pure Mamba encoder with a CNN-based decoder, VM-UNet, a fully Mamba-based U-Net, and U-Mamba, a hybrid model combining Mamba and CNN. A key consideration is determining the most effective way to integrate Mamba into a U-shaped architecture to maximize performance. The comparison between pure Mamba-based models and hybrid approaches remains crucial in identifying the optimal design for medical segmentation.

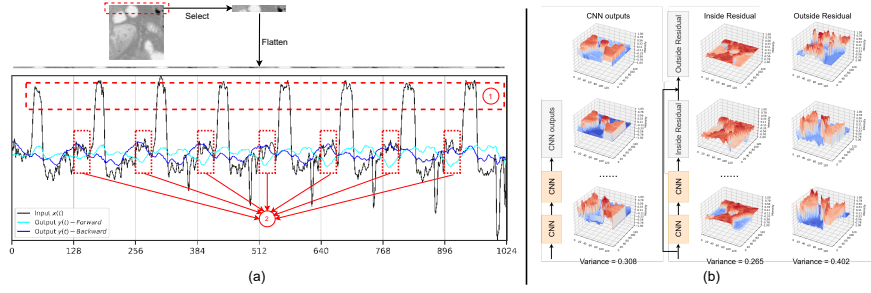


Fig. 2: (a) S4 fitting results on flattened 2D medical images. (b) Feature-map intensity distributions from pre-trained models. Averaged over 3.2M feature samples, variances are 0.209 for two sequential CNNs, 0.097 outside residual connections, and 0.042 inside residual connections, indicating substantially lower variance within residual paths.

To address this, we introduce MM-UNet (Meta Mamba UNet), a unified U-shaped encoder-decoder architecture with skip connections, as shown in Fig. 1. Each stage of the encoder, bottleneck, and decoder consists of a replaceable meta-block, allowing flexibility in module selection. Five different configurations are evaluated, including pure CNN-based, hybrid CNN-Mamba, and pure Mamba-based modules. Extensive experiments reveal that a hybrid module, where SSMs are placed after two sequential CNNs and within residual connections, delivers the best performance. Additionally, replacing meta-blocks in the encoder and bottleneck with this hybrid module further enhances results.

These improvements are attributed to the ability of pre-trained feature maps behind CNN layers and inside residual connections to exhibit lower variance compared to those outside residual connections (Fig. 2(b)), reinforcing the observation that SSMs struggle with high-variance inputs. With the macro-architecture defined, optimal scan orders for SSMs are explored. Fig. 5 illustrates various scan order strategies, and results indicate that a simple bi-directional scan order achieves the best performance. This strategy is adopted in the final model to maximize effectiveness.

In summary, we present the following key contributions:

- We propose MM-UNet, a unified U-shaped encoder-decoder architecture with skip connections, capable of representing existing Mamba-based models.
- We identify the challenges of SSMs in handling high-variance data and discontinuities caused by flattening spatial dimensions, and address these limitations through architectural refinements.
- We design a hybrid module that integrates SSMs within residual connections, reducing variance and improving segmentation accuracy.
- We introduce a bi-directional scan order strategy to mitigate discontinuities and enhance spatial coherence in medical image segmentation.
- MM-UNet achieves state-of-the-art performance with a 91.5% and 89.0% Dice score on AMOS2022 [8] task 1 and task 2, surpassing nnUNet by 1.7% and 1.0%, 93.2% Dice on ACDC [1], and 87.1% Dice on BTCV [9].

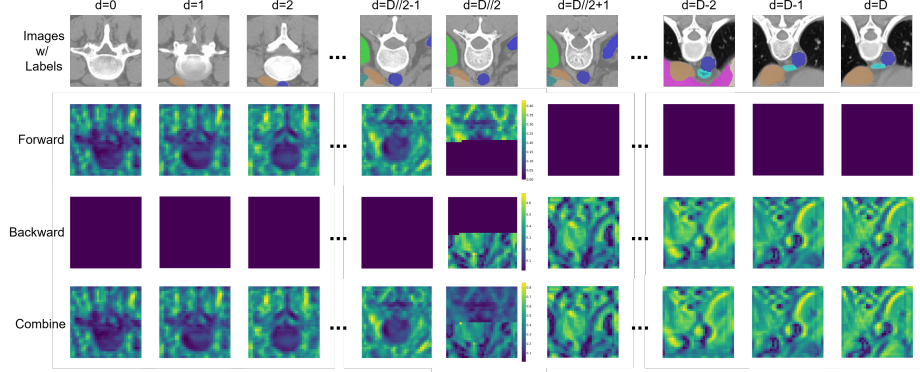


Fig. 3: Visualization of attention maps of QK^T for Mamba. Each attention map effectively captures image patterns across the temporal dimension.

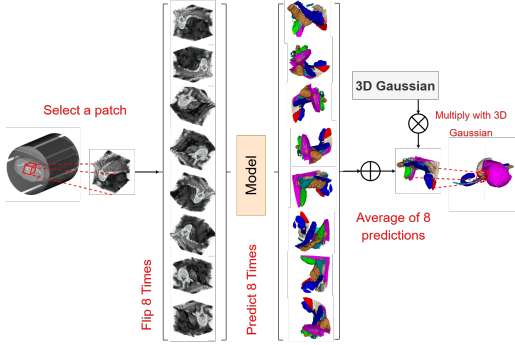


Fig. 4: Inference flipping along axial, coronal, and sagittal axes, akin to multiple scans in Mamba.

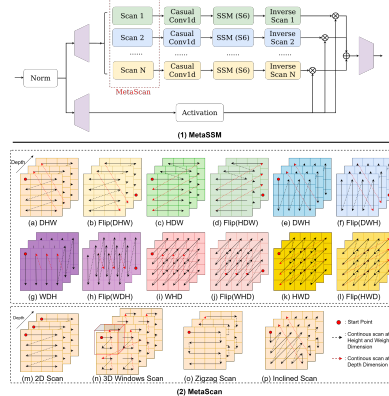


Fig. 5: (1) MetaSSM overview. (2) MetaScan scan order evaluation.

2 The Proposed Method

2.1 Applying Mamba to Medical Imaging

This section explores the use of State Space Models (SSMs) in medical imaging, highlighting their properties and limitations. SSMs are designed for one-dimensional sequences, while images, including medical images, have multiple spatial dimensions. To apply SSMs, images are typically flattened using a predefined scan order, e.g., row by row along width then height.

To assess the effect of scan order, a 128×128 medical image was flattened width-first and processed with the Structured State Space Sequence Model (S4) [3], with the first eight rows shown in Fig. 2(a). S4 struggles with points far from the mean due to its High-Order Polynomial Projection Operator (HiPPO) [2], which determines the coefficients of several polynomials that are combined to approximate the one-dimensional sequence, as shown in the case (1) of Fig. 2(a). This process inherently smooths the predictions, which is problematic in medical

Dataset	Method	Spleen	R.Kd	L.Kd	GB	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg.
AMOS-CT (task 1)	UNETR [5]	0.928	0.913	0.903	0.719	0.763	0.955	0.849	0.922	0.838	0.766	0.663	0.663	0.662	0.815	0.744	0.807
	nnFormer [17]	0.950	0.948	0.944	0.789	0.784	0.967	0.914	0.931	0.868	0.828	0.654	0.695	0.759	0.865	0.773	0.845
	SwinUNETR [4]	0.954	0.954	0.950	0.819	0.852	0.972	0.919	0.955	0.911	0.875	0.775	0.801	0.816	0.895	0.812	0.884
	SwinUNETRv2 [6]	0.959	0.962	0.958	0.842	0.867	0.976	0.933	0.957	0.920	0.889	0.783	0.812	0.843	0.913	0.836	0.897
	3D UX-Net [10]	0.955	0.956	0.953	0.826	0.858	0.972	0.922	0.955	0.915	0.881	0.781	0.809	0.820	0.902	0.823	0.889
	nn-UNet [7]	0.951	0.961	0.956	0.826	0.869	0.973	0.931	0.957	0.923	0.880	0.784	0.809	0.846	0.898	0.827	0.893
	VMUNet [15]	0.958	0.967	0.964	0.847	0.866	0.976	0.938	0.957	0.923	0.882	0.785	0.805	0.852	0.904	0.819	0.896
	SwinUMamba [11]	0.958	0.967	0.965	0.846	0.870	0.976	0.940	0.958	0.922	0.883	0.783	0.811	0.856	0.902	0.818	0.897
	UMamba [13]	0.971	0.969	0.967	0.863	0.879	0.979	0.944	0.960	0.929	0.894	0.795	0.819	0.864	0.914	0.830	0.905
	MM-UNet (Ours)	0.977	0.973	0.970	0.883	0.891	0.981	0.949	0.963	0.933	0.906	0.806	0.832	0.875	0.932	0.852	0.915
AMOS-CTMRI (task 2)	UNETR [5]	0.925	0.938	0.923	0.737	0.776	0.948	0.854	0.918	0.836	0.803	0.672	0.684	0.723	0.814	0.757	0.821
	nnFormer [17]	0.945	0.946	0.942	0.773	0.779	0.971	0.903	0.931	0.856	0.826	0.696	0.726	0.740	0.837	0.741	0.840
	SwinUNETR [4]	0.953	0.963	0.955	0.795	0.828	0.976	0.919	0.949	0.891	0.858	0.733	0.764	0.793	0.888	0.785	0.869
	SwinUNETRv2 [6]	0.951	0.960	0.949	0.815	0.838	0.962	0.911	0.940	0.887	0.862	0.740	0.777	0.809	0.879	0.801	0.872
	3D UX-Net [10]	0.948	0.950	0.945	0.814	0.832	0.972	0.915	0.945	0.887	0.869	0.738	0.772	0.802	0.873	0.806	0.870
	nn-UNet [7]	0.955	0.961	0.957	0.823	0.841	0.976	0.919	0.952	0.894	0.877	0.744	0.787	0.819	0.880	0.816	0.880
	VMUNet [15]	0.956	0.965	0.957	0.842	0.840	0.972	0.928	0.948	0.897	0.878	0.744	0.776	0.822	0.878	0.818	0.881
	SwinUMamba [11]	0.951	0.964	0.958	0.838	0.833	0.973	0.934	0.942	0.898	0.872	0.737	0.767	0.821	0.879	0.820	0.879
	UMamba [13]	0.954	0.964	0.956	0.841	0.845	0.975	0.930	0.951	0.897	0.876	0.752	0.781	0.821	0.881	0.817	0.882
	MM-UNet (Ours)	0.965	0.967	0.960	0.847	0.852	0.978	0.937	0.953	0.902	0.886	0.758	0.792	0.836	0.902	0.827	0.890

Table 1: AMOS-CT and AMOS-CTMRI test results on official leaderboards.

imaging since different organs and tissues often exhibit high-intensity variations that are critical for accurate segmentation. Reducing input variance is therefore necessary when using SSMs.

Another challenge occurs at row transitions, marked by the cyan line in case (2) of Fig. 2(a). In the red boxes, predictions are nearly the inverse of the ground truth (black line) because the start of a new row is not necessarily correlated with the end of the previous one, making cross-row dependencies difficult to model. Thus, discontinuities introduced by flattening should be minimized.

Errors caused by discontinuities are specific to a given scan order. A simple and effective solution is to apply SSMs in the reverse order, where row beginnings and endings are swapped. The blue line in Fig. 2(a) shows this inverse ordering, which significantly improves predictions in the red boxes. This demonstrates that using a pair of opposite scan orders mitigates discontinuity-induced errors.

Although images have multiple spatial dimensions, our experiments show that a single pair of opposite scan orders is sufficient to represent the input. Fig. 3 presents Mamba attention maps, indicating effective pattern modeling.

During inference (Fig. 4), each patch is flipped along axial, coronal, and sagittal axes, generating eight variants whose predictions are averaged. This is analogous to applying multiple scan orders in Mamba. A 3D Gaussian weighting is then applied to emphasize central regions and suppress peripheral areas, further reducing boundary-related errors. This strategy is the default in the nnU-Net codebase, and all reported results in this paper, including baselines, are obtained using this setting.

2.2 Architectural Design of MM-UNet

Recent advances in medical image segmentation, driven by Transformer-based models such as SwinUNETR and nnFormer, have led to U-shaped architectures incorporating Mamba modules, including SegMamba (Mamba encoder + CNN decoder), VM-UNet (fully Mamba-based), and U-Mamba (CNN-Mamba hybrid).

Method	Spl.	R.Kd	L.Kd	GB	Eso.	Liv.	Stom.	Aorta	IVC	Veins	Panc.	AG	DSC
3D UX-Net [10]	0.946	0.942	0.943	0.593	0.722	0.964	0.734	0.872	0.849	0.722	0.809	0.671	0.814
UNETR [5]	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
Swin-UNETR [4]	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	0.812	0.794	0.765	0.869
nnUNet [7]	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
nnFormer [17]	0.935	0.949	0.950	0.641	0.795	0.968	0.901	0.897	0.859	0.778	0.856	0.739	0.856
VMUNet [15]	0.962	0.948	0.951	0.581	0.784	0.968	0.866	0.900	0.861	0.754	0.815	0.722	0.843
SwinUMamba [11]	0.961	0.950	0.951	0.614	0.777	0.970	0.876	0.904	0.866	0.755	0.822	0.712	0.847
UMamba [13]	0.962	0.950	0.950	0.627	0.789	0.970	0.888	0.909	0.871	0.769	0.835	0.714	0.853
MM-UNet (Ours)	0.967	0.952	0.953	0.660	0.806	0.973	0.932	0.916	0.886	0.803	0.862	0.740	0.871

Table 2: Comparison of MM-UNet with state-of-the-art methods on BTCV (DSC in %).

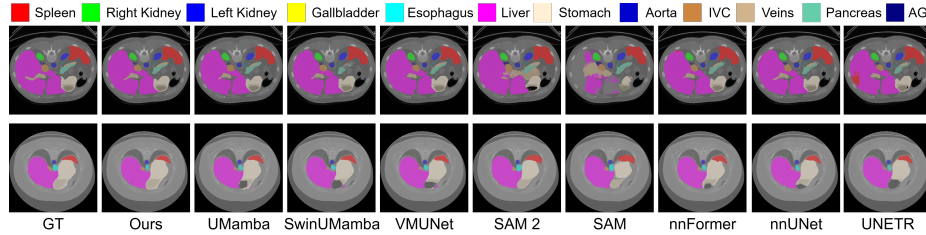


Fig. 6: Qualitative comparison on BTCV. MM-UNet is the most precise for each class.

However, the optimal way to integrate Mamba into U-shaped architectures, and whether hybrid designs outperform pure Mamba models, remains unclear.

To address this, we propose MM-UNet, a unified meta U-Net framework (Fig. 1(a)) with a symmetric encoder-decoder and skip connections for fusing low-level and high-level features. It consists of a stem for channel expansion, multi-stage encoders (downsampling + EncMetaBlock), a bottleneck (MetaBlock), mirrored decoders (upsampling + DecMetaBlock), and a FinalBlock for prediction. The EncMetaBlock, DecMetaBlock, and MetaBlock are interchangeable, supporting CNN-based, pure Mamba-based, or hybrid CNN-Mamba designs (Fig. 1(b)). Replacing all blocks with CNN layers yields an nnUNet-like model, while using only Mamba blocks forms a VM-UNet-like architecture.

We systematically evaluate configurations using a bi-directional Mamba block (MetaSSM). Replacing EncMetaBlock and MetaBlock with Mamba modules (SwinUMamba-style) or replacing all three meta blocks (VM-UNet-style) improves performance by 1.9% and 1.8%, respectively. In contrast, replacing only DecMetaBlock and MetaBlock reduces performance to 89.4%, indicating that Mamba is most effective in the encoder and bottleneck.

Further analysis shows that a hybrid design, two convolutional layers followed by a Mamba, outperforms both pure CNN and pure Mamba variants. As illustrated in Fig. 2(b), feature maps within residual connections exhibit lower variance, which benefits SSMS that are sensitive to high-variance inputs. Embedding Mamba inside residual connections yields an additional 0.9% improvement.

Accordingly, MM-UNet adopts hybrid CNN-Mamba modules with Mamba embedded within residual connections in the encoder and bottleneck, achieving the best segmentation performance.

2.3 Scan Design

We investigate different scan order strategies for the Mamba block (MetaSSM), as shown in Fig. 5. Since medical images are typically 3D, flattening them into 1D sequences can follow multiple orders. For example, the Depth–Height–Width (DHW) order scans along width, then height, and finally depth. Other permutations (DWH, WDH, WHD, HDW, HWD), together with their inverse orders, are illustrated in Fig. 5(2a)–(2l). Additional strategies, including 2D scan, 3D window-based scan, zigzag, and inclined scan, are shown in Fig. 5(2m)–(2p).

Experimental results are summarized in Table 4. The standard 1D DHW scan (B1) achieves a Dice score of 0.902. Introducing its reverse order to form a bi-directional scan (1D BiScan, B2) improves performance to 0.915 (+0.013), demonstrating that opposite-direction scanning effectively mitigates prediction errors caused by discontinuities during flattening.

We further examine adding more scan directions. However, introducing an additional unpaired direction (HWD) does not improve performance, suggesting that unpaired scans introduce harmful discontinuities. Extending to three pairs (B4) or six pairs (B5) yields Dice scores of 0.908 and 0.907, respectively, indicating that multiple pairs provide no additional benefit and may introduce redundancy.

Other scan strategies are also evaluated. The 2D scan (B6) underperforms due to insufficient modeling of sequential continuity. The 3D window-based scan (B7) achieves comparable performance (0.902) to the 1D scan but increases discontinuities due to frequent boundary transitions. The zigzag scan (B8) performs similarly to 1D BiScan (0.909), while the inclined scan (B9) performs slightly worse (0.901), likely due to additional discontinuities.

Overall, a simple bi-directional strategy with one pair of opposite-direction orders provides the best trade-off between spatial continuity and discontinuity mitigation. Therefore, we adopt the 1D BiScan, consisting of DHW and its inverse flip(DHW), as the default scan strategy in MM-UNet.

3 Experiments

Datasets and Evaluation Metrics. We evaluate on three public datasets: AMOS22 [8], BTCV [9], and ACDC [1]. AMOS22 has two tasks: CT (Task 1) with 300 training and 200 test CT scans, and CT+MRI (Task 2) adding 60 and 40 MRI scans for training and test, both covering 16 abdominal organs. BTCV contains 30 abdominal CT scans (24 train / 6 val) with 13 organs. ACDC includes 200 training and 100 test cardiac MRI scans with RV, MYO, and LV annotations. AMOS22 and ACDC both use 5-fold cross-validation. Performance is measured by Dice Similarity Coefficient (DSC) on all datasets.

Comparison with State-of-the-Art Methods. We compare MM-UNet with representative convolution-based methods (nnUNet [7, 10]), transformer-based methods (UNETR [5], SwinUNETR [4], and nnFormer [17]), and Mamba-based methods (VMUNet [15], SwinUMamba [11], and UMamba [13]) on ACDC, BTCV, and AMOS22 task1 and task 2 benchmarks. As shown in Tabs. 3–1, MM-UNet

Method	RV $\uparrow$	Myo $\uparrow$	LV $\uparrow$	DSC $\uparrow$
UNETR [5]	88.51	88.35	92.78	89.88
SwinUNETR [4]	89.97	89.07	93.12	90.72
3D UX-Net [10]	89.13	88.53	92.94	90.20
nnUNet [7]	90.59	89.52	93.86	91.32
nnFormer [17]	92.09	90.74	94.40	92.41
VMUNet [15]	91.71	90.57	94.50	92.26
SwinUMamba [11]	90.68	89.60	93.98	91.42
UMamba [13]	91.98	90.63	94.82	92.48
MM-UNet (Ours)	92.61	91.65	95.30	93.19

Table 3: Comparison on ACDC with 5-fold ensembles.

Method	DSC $\uparrow$
B1 (a) DHW	0.902
B2 (a) DHW + (b) flip(DHW)	0.915
B3 (a) DHW + (b) flip(DHW) + (k) HWD	0.898
B4 (a) + (b) + (c) + (d) + (e) + (f)	0.908
B5 (a) + (b) + (c) + (d) + (e) + (f) + (g) + (h) + (i) + (j) + (k) + (l)	0.907
B6 (m) 2D Scan w/ DHW + flip(DHW)	0.900
B7 (n) 3D Window Scan w/ DHW + flip(DHW)	0.902
B8 (o) Zigzag Scan w/ DHW + flip(DHW)	0.909
B9 (p) Inclined Scan w/ DHW + flip(DHW)	0.901

Table 4: Performance comparison of different MetaScan configurations with various scan orders.

consistently achieves state-of-the-art performance across all datasets. On ACDC, our method attains the highest average DSC of 93.19%, outperforming strong baselines such as nnFormer and UMamba. On BTCV, MM-UNet achieves a new SOTA DSC of 87.1%, surpassing both transformer-based and Mamba-based competitors, with notable improvements on challenging organs such as the pancreas and stomach. On the large-scale AMOS22 benchmark (CT and CT-MRI), MM-UNet achieves the highest accuracy (91.5% DSC and 89.0% DSC on task 1 and task 2) on the official leaderboards, delivering the best average DSC in both tasks and consistently improving segmentation accuracy across most organs. These results demonstrate MM-UNet’s robustness and generalization across diverse datasets, modalities, and anatomical structures.

Attention Maps for Mamba. Inspired by VMamba [12], we visualize the attention values of QK^T in MetaSSM to gain a deeper understanding of how our model processes 3D medical images (Fig. 3). Despite flattening the volumetric inputs into a 1D sequence, MetaSSM produces attention maps that clearly encode spatial-temporal anatomical patterns across distant slices. The forward and backward state transitions exhibit complementary receptive fields, allowing MetaSSM to model long-range dependencies along the depth dimension without relying on heavy multi-head attention.

Compared to transformer attention, which relies on large parameter budgets and costly token-token interactions, MetaSSM produces sharper and more structured attention maps with substantially fewer parameters and lower computational cost. The visualizations show that MetaSSM preserves depth continuity, captures inter-slice organ shape evolution, and maintains strong representations even under extreme token compression, highlighting its suitability for efficient and accurate 3D medical image segmentation.

4 Conclusion

We propose MM-UNet, a unified U-shaped encoder-decoder framework for medical image segmentation. It is the first architecture capable of representing existing Mamba-based models while effectively integrating CNNs and state space models (SSMs). We address key challenges of applying SSMs to medical images, including

sensitivity to high-variance inputs and discontinuities caused by flattening multi-dimensional data into 1D sequences. Through systematic analysis, we identify an optimal macro design for SSM placement and introduce a bi-directional scan strategy for 3D images to mitigate discontinuity-induced errors.

Extensive experiments demonstrate state-of-the-art performance: 91.5% and 89.0% Dice on AMOS2022 Task 1 and Task 2 (surpassing nnUNet by 1.7% and 1.0%), 93.2% on ACDC, and 87.1% on BTCV, confirming its robustness and generalization ability. Beyond strong empirical results, our work provides insights into CNN-SSM integration and establishes an effective hybrid design bridging convolutional and sequence-based modeling for medical vision.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging* **37**(11), 2514–2525 (2018)
2. Gu, A., Dao, T., Ermon, S., Rudra, A., Ré, C.: Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems* **33**, 1474–1487 (2020)
3. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021)
4. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI Brainlesion Workshop*. pp. 272–284. Springer (2021)
5. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 574–584 (2022)
6. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 416–426. Springer (2023)
7. Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128* (2019)
8. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems* **35**, 36722–36732 (2022)
9. Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In: *Proc. MICCAI: Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge* (2015)

10. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022)
11. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., 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 (2024)
12. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
13. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
15. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
16. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
17. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)