2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

CMUNet: Enhancing Image Segmentation through Advanced Channel Dependency Modeling

Sizhe Yang, Yutao Qin, and Wei Ren(✉)

School of Computer and Information Science, Southwest University, Chongqing, China

Abstract. As a cornerstone of computer vision, image segmentation decomposes visuals into semantically coherent regions to simplify subsequent analysis. U-shaped models inspired by UNet have become the dominant architecture in this field, leveraging an encoder-decoder design with skip connections to retain spatial details. Within this framework, despite their impressive performance, CNNs are constrained by local receptive fields, while Transformers incur substantial computational overhead when capturing long-range dependencies. Recently, Mamba has emerged as a powerful approach for modeling long-range dependencies with linear complexity. However, existing efforts to integrate Mamba into U-Net primarily emphasize spatial feature extraction, largely overlooking the intricate inter-channel relationships encapsulating diverse semantic patterns. In this paper, we propose Channel Mamba UNet (CMUNet), which explicitly captures channel dependencies with two key components: Channel Mamba (CMamba), a module that adaptively recalibrates channel-wise features in the encoder, and Skip Connection Mamba (SkiM), a mechanism that facilitates multi-level channel fusion to align semantic information between encoder and decoder layers. Comprehensive evaluations on MoNuSeg, GlaS and ISIC-2018 confirm the effectiveness of CMUNet, achieving Dice scores of 81.28, 92.18, and 91.13, along with superior IoU and HD95 metrics.

Keywords: Image Segmentation, Channel-wise Modeling, UNet, Mamba.

1 Introduction

Image segmentation, a core challenge in computer vision, focuses on delineating coherent regions or object instances within an image to enable deeper structural interpretation. [8]. In this field, U-shaped architectures inspired by U-Net[22] have become the de facto standard, with an encoder path to extract semantic features, a decoder path to restore spatial details, and skip connections to preserve fine-grained information across corresponding levels.Both CNN-based and transformer-based models have achieved impressive results with this framework, but they face their own challenges --- CNNs struggle with their limited receptive field [15], while transformers suffer the costly long-term modeling due to the self-attention mechanism [1].

Recent advances in structured state space models (SSMs) [7], exemplified by Mamba [5], provide an efficient solution for long-range sequence modeling with input-length–linear complexity. Many studies attempt to improve accuracy by directly plugging Mamba into UNet encoders or decoders, leveraging its efficiency in modeling long-range dependencies [30]. However, these efforts primarily focus on spatial feature extraction but largely ignore the intricate relationships between channels. Different channels usually focus on different semantic patterns [10], failing to explicitly model them limits the network's ability to fully exploit the richness of multi-channel representations, potentially constraining segmentation performance.

In this paper, we focus on channel relationships and propose Channel Mamba UNet (**CMUNet**). Our approach introduces channel-wise modeling from two key perspectives. First, in the encoder, we design Channel Mamba (**CMamba**), a module that adaptively recalibrates channel-wise features after spatial modeling, selectively enhancing informative representations while suppressing less relevant ones. Second, for skip connections, we introduce Skip Connection Mamba (**SkiM**), a novel mechanism that effectively captures non-local semantic dependencies, enabling multi-level channel-wise information fusion to reduce the mismatch in abstraction levels between encoder outputs and decoder inputs. By jointly leveraging these two components, CMUNet fully exploits channel interactions, leading to enhanced segmentation performance.

Main contributions of this paper are as follows:

(1) Channel Mamba (**CMamba**): We propose CMamba, a novel module that dynamically adjusts channel-wise features in the encoder to refine semantic representations

(2) Skip Connection Mamba (**SkiM**): To mitigate the disparity in semantic representation between the encoder and decoder, we propose SkiM, facilitating multi-level channel-wise feature fusion.

(3) Channel Mamba UNet (**CMUNet**): We propose CMUNet, a U-Net architecture that utilizes CMamba and SkiM. Simulations on three benchmark datasets: GlaS[24], MoNuSeg [14], and ISIC-2018 [4] show that CMUNet achieves state-of-the-art performance in terms of Dice scores, IoU, and HD95, highlighting its effectiveness in capturing both spatial and channel-wise information for precise segmentation.

2 Related Works

2.1 Mamba in Image Segmentation

In recent years, U-Net [22] variants integrating the Mamba architecture have achieved significant progress in image segmentation. For instance, VMUNet [23] pioneers the use of Mamba throughout both encoder and decoder stages by integrating VSS [17] modules, enabling enhanced extraction of spatially detailed features. SwinUMamba [16], on the other hand, utilizes a Mamba backbone pre-trained on ImageNet within its encoder, leveraging large-scale pretraining to enhance performance. Meanwhile, LKMUNet [28] leverages the efficiency of Mamba to construct large-kernel variants, aiming to expand the receptive field and capture wider spatial dependencies. Despite these advancements, these models predominantly focus on spatial information, often neglecting the intricate and dynamic interactions within the channels.

2.2 Channel Attention Modules

Channel-focused attention strategies have become a common tool in computer vision to strengthen representational power and boost segmentation accuracy. Classic approaches, such as Squeeze-and-Excitation (SE [10]), first transform an image into channel descriptors, which are then mapped to a set of channel weights using two fully connected (FC) layers followed by a sigmoid activation function. Similarly, CBAM [29] combines spatial and channel attention, where channel-wise relations are modeled using an MLP.

In these methods, channel sequences are typically processed using simple MLP [21] or FC layers, limiting their ability to capture complex dependencies. In contrast, Mamba excels in sequence modeling by efficiently capturing long-range dependencies [5], allowing for a more nuanced understanding of inter-channel relationships. This advantage makes Mamba particularly well-suited for effectively handling channel sequences.

3 Method

3.1 Overview of Channel Mamba UNet (CMUNet)

An overview of our Channel Mamba UNet (CMUNet) is shown in Fig 1. Following the classic U-shape structure of UNet. Let F_l represent the feature map at the l-th layer. Given an input image $F_{in} \in R^{C\times W\times H}$, the convolutional STEM layer initial encodes it into the feature map $F_0 \in R^{64\times\frac{H}{2}\times\frac{W}{2}}$. At each successive level, the spatial dimensions of the feature map are halved, while the number of channels is progressively doubled. $F_l \in R^{(64\times 2^l)\times\frac{H}{2^{l+1}}\times\frac{W}{2^{l+1}}}$.

To enhance feature representation at both the spatial-wise and channel-wise, we introduce Spatial-and-Channel Mamba (SCMamba) in the encoder, which simultaneously captures fine-grained spatial dependencies and global channel interactions. SCMamba consists of two components: (1) VSS Block from VMamba [17], serving as the token-mixer to model spatial dependencies. (2) Channel Mamba (CMamba) serves as the Channel-MLP, fine-tuning the channel-specific features by uncovering intricate inter-channel correlations. The process inside each encoder can be formulated as:

$$F_l' = LN\big(VSS(F_l)\big) \tag{1}$$

$$F_l'' = LN\big(CMamba(F_l')\big) \tag{2}$$

$$F_{l+1} = downsampling(F_l'') \tag{3}$$

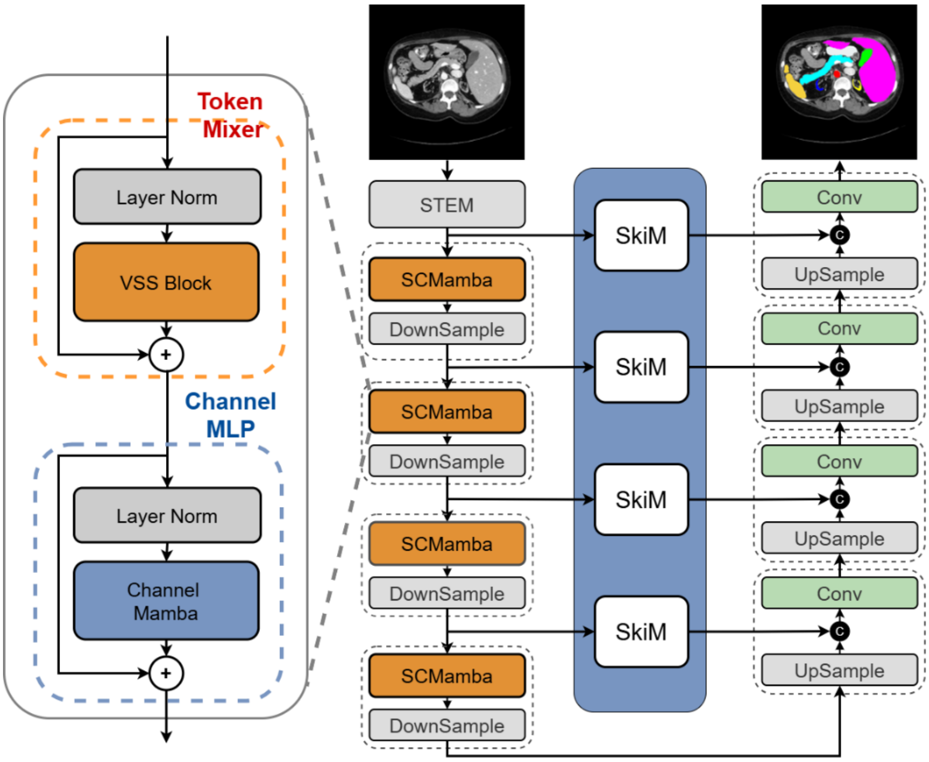


Fig.1. The Overall architecture of CMUNet.

Beyond the encoder, we further refine feature transmission using Skip Mamba (SkiM) in the skip connections, which consolidates multi-level channel information, ensuring smoother semantic transitions from the encoder to the decoder. At the l-th level, the process can be briefly formulated in:

$$F_l = CMamba\left(LN\left(VSS\left(LN\left(F_{l-1}^{enc}\right)\right)\right)\right) \tag{4}$$

$$F_l^{skip} = SKiM_l\left(F_1^{enc}, F_2^{enc}, F_3^{enc}, F_4^{enc}\right) \tag{5}$$

$$F_l^{dec} = Conv\left(Concate\left(F_l^{skip}, F_{l-1}^{dec}\right)\right) \tag{6}$$

where F_l^{enc},F_l^{skip} ,F_l^{dec} represents the feature map in the encoder, skip connection and decoder of the l-th level

3.2 Channel Mamba (CMamba)

In this section, we introduce Channel Mamba (CMamba), which extracts a global channel vector using Global Average Pooling (GAP), processes it with Non-Causal Mamba, and subsequently recalibrates the feature map based on the refined channel vector.

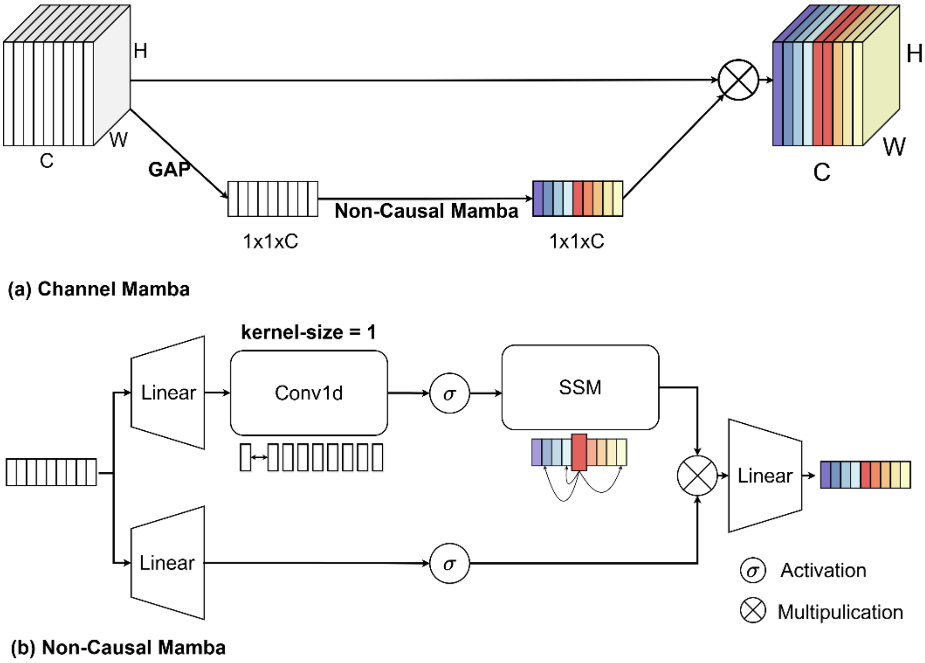


Fig.2. (a) the process of Channel Mamba (CMamba), (b) the process of Non-Causual Mamba with a kernel size of 1 for the 1d conv step

Channel Vector. Mamba models on sequence, so CMamba first extracts a global channel descriptor by applying Global Average Pooling (GAP) to the feature map. Given a feature map $F \in R^{C\times H\times W}$

$$V_c = GAP(F) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F(i,j) \tag{7}$$

the Channel Vector $V_c \in R^{1\times 1\times C}$, which represents a compressed representation of global context inside each channel.

Non-Causality. In many sequences -- such as temporal or spatial ones - the order of elements inherently implies causality: neighboring elements are typically more strongly correlated because their positions encode valuable information. In contrast, the channel vector V_c lacks this intrinsic ordering, meaning that its channels operate independently and do not reflect any sequential causality.

Inside Mamba [5], the standard 1D convolution tends to amplify local relationships, inadvertently introducing sequential causality into the channels. To mitigate this effect, as shown in Fig.2 (b), we configure the 1D convolution with a kernel size of 1, thereby eliminating the locality bias typically introduced by convolution. This modification ensures that CMamba focuses solely on capturing the global dependencies among channels. The process can be formulated as:

$$V_c' = non - caual - Mamba(V_c) \tag{8}$$

where V_c' represents the processed Channel Vector.

Feature Recalibration. After non-causal Mamba, the processed Channel Vector $V_c{}'$. is applied to the original feature map F to recalibrate its channel responses. This is achieved by performing element-wise multiplication (Hadamard product), enabling the dynamic adjustment of channel activations by enhancing or suppressing them selectively. The process can be formulated as:

$$F_{out} = F \odot V_c' \tag{9}$$

where $\odot$ represents Hadamard product, and $F_{out} \in R^{C\times H\times W}$is the output feature map with channel-wise attention.

3.3 Skip connection Mamba (SkiM)

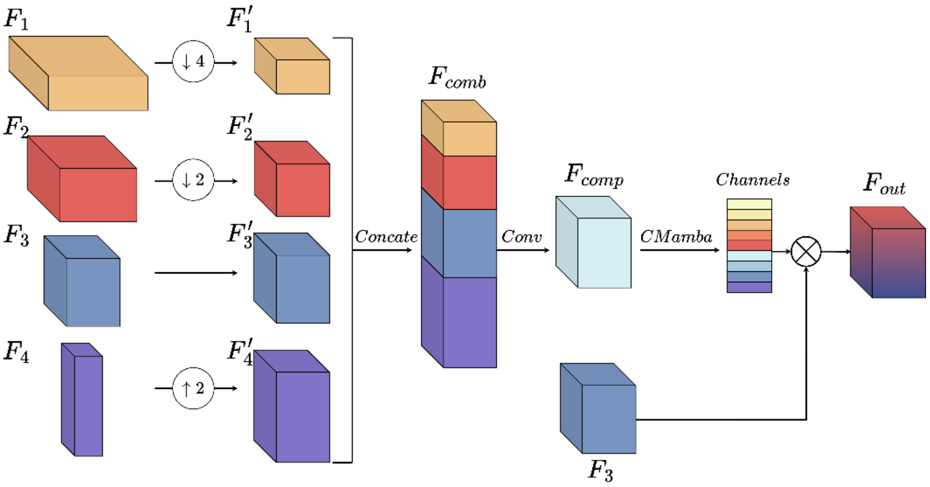


Fig.3. Illustation of Skip connection Mamba (SkiM) , taking $SKiM_3$ as example

Relying solely on channel operations in the encoder is insufficient, as semantic gaps still exist between the encoder and decoder. To bridge this, we introduce Skip Connection Mamba (SkiM), which enhances skip connections with multi-level channel-wise information.

At the l-th level, we first resize feature maps from different encoder levels to a uniform resolution and concatenate them:

$$F_i' = resize_l(F_i^{enc}) \in R^{C_i\times\frac{H}{2^{l+1}}\times\frac{W}{2^{l+1}}} \tag{10}$$

$$F_l^{comb} = Concatenate(F_1', F_2', F_3', F_4') \tag{11}$$

where F_l^{comb} represents the combined feature map that aggregates channel information across multiple semantic levels. A convolution then reduces it to align with the channel dimensionality of the l-th encoder feature map:

$$F_l^{fused} = Conv\left(F_l^{comb}\right) \tag{12}$$

where $F_l^{fused}$is the compressed feature map with 64×2^l channels.

Next, CMamba extracts a channel vector to encode multi-level channel-wise dependencies:

$$V_c = CMamba\left(F_l^{fused}\right) \tag{13}$$

Finally, this vector recalibrates the original encoder feature map:

$$F_l^{skip} = F_l^{enc} \odot V_c \tag{14}$$

where F_l^{skip} represents the output of $SKiM$ in the l-th level, which enhances feature transmission with richer channel interactions, and mitigating the misalignment of semantic content between the encoding and decoding phases.

4 Experiment

4.1 Datasets and Implementation Details

We benchmark CMUNet on three established segmentation benchmarks. The GlaS [24] collection offers 85 gland images for training and 80 for evaluation. MoNuSeg [14] provides 30 annotated slides for training and 14 held out for testing, targeting multi-organ nuclei delineation. For skin lesion segmentation, we utilize the ISIC-2018[4] challenge set, which comprises 2,594 labeled images for model development.

All training and inference routines were executed on an NVIDIA RTX 4090 GPU. We assessed segmentation quality using the Dice score, Intersection over Union (IoU), and the 95th-percentile Hausdorff distance (HD95). To improve generalization, input images were subject to random flips (horizontal/vertical) and rotations [25]. Optimization was handled by the Adam optimizer [13] with a Cosine Annealing learning-rate schedule [20], and we minimized a hybrid loss combining cross-entropy and Dice terms. Every image was uniformly resized to 224 × 224 pixels before feeding into the network [27].

4.2 Comparisons of Model Performances

Table 1 shows the performance of CMUNet across three datasets, demonstrating its superior segmentation capabilities. CMUNet outperforms state-of-the-art methods in Dice, IoU, and HD95 metrics for all datasets.

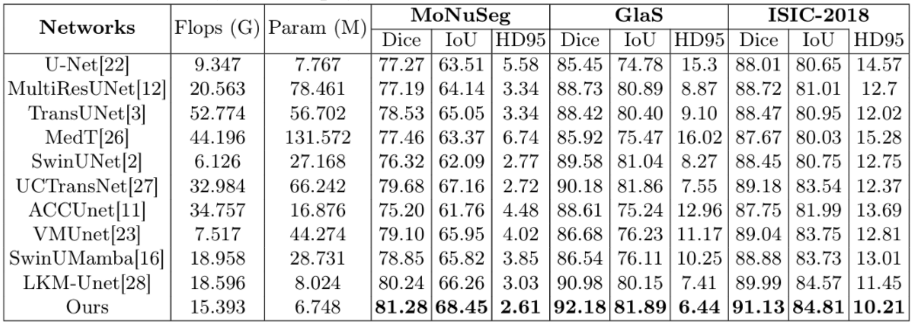
We organize the benchmarked approaches into two distinct paradigms: established architectures rooted in conventional designs, and recent Mamba-integrated frameworks tailored for long-sequence modeling. Traditional models encompass a variety of CNN-based architectures (such as U-Net [22], MultiResUNet [12], and ACCUNet [11]), Transformer-based models (including SwinUNet [2], and MedT [26]), and hybrid models (e.g., UCTransNet [27], TransUNet [3]). In contrast, Mamba-based models, which emerged in 2024 and have demonstrated advanced capabilities in handling complex segmentation tasks, include VM-Unet [23], SwinUMamba [16], and LKM-Unet [28].

Specifically, CMUNet surpasses the second-place CNN-based models by 3.45 in **Dice** score and the Transformer-based models by 2.00 on the GlaS dataset. When compared to the latest Mamba networks, CMUNet achieves an additional 1.20 in Dice score. On the large-scale ISIC-2018 dataset, CMUNet outperforms CNN models by 2.41 and Transformer-based models by 1.95 in Dice score. Additionally, CMUNet exceeds the Mamba networks by 1.14 in Dice, demonstrating its high consistency with the ground truth.

For **IoU**, which reflects object detection ability, CMUNet outperforms the second-place model by 0.83 on MoNuSeg, 2.19 on GlaS, 0.26 on ISIC, indicating its superior object detection accuracy across all three datasets. And for edge detection, CMUNet attains the lowest **HD95** values on all three datasets—2.61 on MoNuSeg, 6.44 on GlaS, and 10.21 on ISIC-2018—indicating its excellent capability in clearly delineating object boundaries. Benefiting from its capacity to highlight salient representations while downplaying irrelevant signals, CMUNet consistently delivers strong results across datasets of varying scales. This reflects not only its adaptability to differing data complexities, but also its robustness and precision in diverse segmentation environments.

With 15.393G FLOPs and 6.748M parameters, CMUNet strikes an effective balance between computational overhead and segmentation quality, offering a resource-efficient solution suitable for deployment in real-world scenarios."

Table 1. Comparison of Dice, IoU, and HD95 scores across MoNuSeg, GlaS, and ISIC-2018 datasets, and Flops, Param for different networks



Networks	Flops (G)	Param (M)	MoNuSeg			GlaS			ISIC-2018		
			Dice	IoU	HD95	Dice	IoU	HD95	Dice	IoU	HD95
U-Net[22]	9.347	7.767	77.27	63.51	5.58	85.45	74.78	15.3	88.01	80.65	14.57
MultiResUNet[12]	20.563	78.461	77.19	64.14	3.34	88.73	80.89	8.87	88.72	81.01	12.7
TransUNet[3]	52.774	56.702	78.53	65.05	3.34	88.42	80.40	9.10	88.47	80.95	12.02
MedT[26]	44.196	131.572	77.46	63.37	6.74	85.92	75.47	16.02	87.67	80.03	15.28
SwinUNet[2]	6.126	27.168	76.32	62.09	2.77	89.58	81.04	8.27	88.45	80.75	12.75
UCTransNet[27]	32.984	66.242	79.68	67.16	2.72	90.18	81.86	7.55	89.18	83.54	12.37
ACCUnet[11]	34.757	16.876	75.20	61.76	4.48	88.61	75.24	12.96	87.75	81.99	13.69
VMUnet[23]	7.517	44.274	79.10	65.95	4.02	86.68	76.23	11.17	89.04	83.75	12.81
SwinUMamba[16]	18.958	28.731	78.85	65.82	3.85	86.54	76.11	10.25	88.88	83.73	13.01
LKM-Unet[28]	18.596	8.024	80.24	66.26	3.03	90.98	80.15	7.41	89.99	84.57	11.45
Ours	15.393	6.748	**81.28**	**68.45**	**2.61**	**92.18**	**81.89**	**6.44**	**91.13**	**84.81**	**10.21**

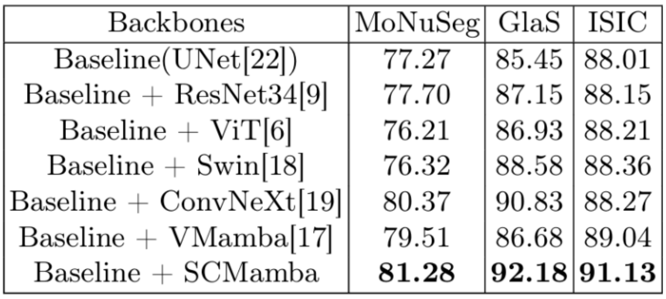
4.3 Ablation Study

Effectiveness of CMUNet's Encoder. To promote richer feature abstraction across spatial layouts and cross-channel interactions, CMUNet integrates SCMamba in the encoder. Its performance is benchmarked against prevalent alternatives in Table 2.

The compared encoders include popular and classic backbones. The UNet [22] encoder, ResNet34[9] and ConvNeXt [19] are convolutional backbones; ViT [6], Swin [18] are transformer-based backbones; VMamba [17] is the newly proposed Mamba backbone. The baseline model for comparison is UNet, a widely adopted architecture in image segmentation.

As evidenced by Table 3, SCMamba yields the most favorable Dice scores—81.28, 92.18, and 91.13 on MoNuSeg, GlaS, and ISIC respectively. These results demonstrate the superior performance of SCMamba compared to other backbones, highlighting its ability to model accurate spatial structure as well as the interdependencies between channels.

Table 3. SCMamba encoder vs. other backbones



Backbones	MoNuSeg	GlaS	ISIC
Baseline(UNet[22])	77.27	85.45	88.01
Baseline + ResNet34[9]	77.70	87.15	88.15
Baseline + ViT[6]	76.21	86.93	88.21
Baseline + Swin[18]	76.32	88.58	88.36
Baseline + ConvNeXt[19]	80.37	90.83	88.27
Baseline + VMamba[17]	79.51	86.68	89.04
Baseline + SCMamba	**81.28**	**92.18**	**91.13**

Effectiveness of Channel-wise Components: CMamba and SkiM. CMUNet handles inter-channel relationships from two perspectives. On one hand, CMamba captures global inter-channel dependencies with negligible extra cost (+0.03G in Flops and +0.01M in param). On the other hand, SkiM mitigates the representational disparity between encoding and decoding stages through multi-level channel aggregation, with only a small additional cost.

To evaluate their effectiveness, we compare different configurations where the baseline is equipped only with the spatial module (the VSS Block) in the SCMamba encoder. As presented in Table 4, the baseline yields Dice scores of 79.51, 91.28, and 89.04 on MoNuSeg, GlaS, and ISIC-2018, respectively. Introducing CMamba alone elevates the performance to 81.15, 91.77, and 90.33, demonstrating its ability to effectively model global inter-channel dependencies. Similarly, enabling SkiM alone improves the Dice scores to 80.54, 92.09, and 90.75, confirming that multi-level skip channel aggregation enhances segmentation accuracy. When both CMamba and SkiM are employed, the model attains the best performance with Dice scores of 81.28 on MoNuSeg, 92.18 on GlaS, and 91.13 on ISIC-2018.

These results demonstrate that CMamba and SkiM contribute complementarily to refining channel representations and reducing the semantic gap, all with only marginal computational overhead.

Table 4. Channel-wise Components

Components	Flops (G)	Param (M)	MoNuSeg	GlaS	ISIC-2018
VSS[17]	13.872	6.129	79.51	91.28	89.04
VSS + CMamba	13.875	6.130	80.54	91.77	90.33
VSS + SkiM	15.390	6.649	81.15	92.09	90.75
VSS + CMamba + SkiM	15.393	6.650	**81.28**	**92.18**	**91.13**

Impact of Kernel Size on Non-Causality. As mentioned in the previous sections, the Channel Vector is inherently non-causal -- there is no distinction between adjacent channels and those far apart. To preserve this non-causal behavior, the 1D convolution in Mamba is ideally set with a kernel size of 1, thereby completely avoiding any locality bias.

Reported in Tab 5, the result demonstrates how varying the kernel size affects performance:

Table 5. Impact of Kernel Size on Channel Mamba

Kernel Size	MoNuSeg	GlaS	ISIC
1	81.28	92.18	91.13
4	76.05	87.93	88.51
16	79.57	88.99	89.64
64	80.77	91.33	90.25

Kernel size = 1: Each channel is modeled individually in the conv 1d steps of Mamba, maintaining non-causality and yielding the best performance.

Kernel size = 4: Introducing a small kernel imposes strong locality. This results in a dramatic drop in performance (-5.23 on MoNuSeg, -4.25 on GlaS, and -2.62 on ISIC-2018), indicating that the enforced locality disrupts the ideal channel modeling.

Kernel sizes = 16 and 64: As the kernel size increases further, the harmful effects of locality diminish. With a kernel size of 16, performance partially recovers (-1.71 on MoNuSeg, -3.19 on GlaS, -1.49 on ISIC-2018), and with a kernel size of 64, the results improve further (-0.51, -0.85, and -0.88, respectively).

In conclusion, small kernels (except for size 1) introduce strong locality into channel modeling, inadvertently imposing causality on channels, which significantly degrades performance. As the kernel size increases, it gradually mitigates these negative effects by approximating global context at very large scales. However, overall, using a kernel size of 1 remains the optimal choice for preserving effective channel modeling.

5 Conclusion

In this paper, we have presented Channel Mamba UNet (CMUNet), a U-shaped framework that brings channel-centric modeling to the forefront of image segmentation. By embedding Channel Mamba (CMamba) units in the encoder and augmenting skip links with Skip Connection Mamba (SkiM), our design captures both broad inter-channel relationships and nuanced multi-level feature interactions. Comprehensive testing on MoNuSeg, GlaS, and ISIC-2018 confirms that CMUNet consistently delivers top-tier segmentation quality—achieving Dice scores of 81.28, 92.18, and 91.13, respectively—while also leading in IoU and HD95 benchmarks. These findings underscore the value of explicit channel modeling via Mamba architectures for pushing the boundaries of segmentation accuracy.

Acknowledgments. This work was supported by Chongqing Normal University (grant number 24XWB030).

References

1. Brauwers, G., Frăsincar, F.: A general survey on attention mechanisms in deep learning. IEEE Transactions on Knowledge and Data Engineer ing 35(4), 1234–1247 (2023). https://doi.org/10.1109/TKDE.2021.3126456, https://doi.org/10.1109/TKDE.2021.3126456
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European Conference on Computer Vision Workshops(ECCVW) (2022)
3. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Im age Anal. 97, 103280 (2024). https://doi.org/10.1016/J.MEDIA.2024.103280, https://doi.org/10.1016/j.media.2024.103280
4. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the interna tional skin imaging collaboration (isic) (2019), https://arxiv.org/abs/1902.03368
5. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient al gorithms through structured state space duality. In: International Conference on Machine Learning (ICML) (2024)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
7. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: The International Conference on Learning Representations (ICLR) (2022)
8. Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval 7, 87–93 (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
11. Ibtehaz, N., Kihara, D.: Acc-unet: A completely convolutional unet model for the 2020s. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention–MICCAI 2023. pp. 692–702. Springer Nature Switzerland (2023)

12. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Networks 121, 74–87 (2020)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Interna tional Conference on Learning Representations (ICLR) (2015)
14. Kumar, N., et al.: A multi-organ nucleus segmentation challenge. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2017)
15. Li, F., Zhou, L., Wang, Y., Chen, C., Yang, S., Shan, F., Liu, L.: Modeling long-range dependencies for weakly supervised disease classifica tion and localization on chest x-ray. Quantitative Imaging in Medicine and Surgery 12(6), 3364–3378 (2022). https://doi.org/10.21037/qims-21-1117, https://pmc.ncbi.nlm.nih.gov/articles/PMC9131331/
16. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Yu, Y., Liang, Y., Shi, G., Zhang, S., Zheng, H., Wang, S.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. arXiv preprint arXiv:2402.03302 (2024)
17. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems 37, 103031–103063 (2025)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B.: Swin trans former: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
20. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR) (2017)
21. Pinkus, A.: Approximation theory of the mlp model in neural networks. Acta numerica 8, 143–195 (1999)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomed ical image segmentation. Medical Image Computing and Computer-Assisted Inter vention (MICCAI) pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574 4_28
23. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image seg mentation (2024), https://arxiv.org/abs/2402.02491
24. Sirinukunwattana, K., et al.: Gland segmentation in colon histology images: The glas challenge contest. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2017)
25. Takahashi, R., Matsubara, T., Uehara, K.: Ricap: Random image cropping and patching data augmentation for deep cnns. In: Zhu, J., Takeuchi, I. (eds.) Pro ceedings of The 10th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 95, pp. 786–798. PMLR (14–16 Nov 2018), https://proceedings.mlr.press/v95/takahashi18a.html
26. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Com puting and Computer Assisted Intervention– MICCAI 2021. pp. 36–46. Springer International Publishing, Cham (2021)
27. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with trans former. Proceedings of the AAAI Conference on Artificial Intelligence 36(3), 2441–2449 (Jun 2022). https://doi.org/10.1609/aaai.v36i3.20144, https://ojs.aaai.org/index.php/AAAI/article/view/20144

28. Wang, J., Chen, J., Chen, D.Z., Wu, J.: LKM-UNet: Large Kernel Vision Mamba UNet for Medical Image Segmentation . In: proceedings of Medical Image Com puting and Computer Assisted Intervention– MICCAI 2024. vol. LNCS 15008. Springer Nature Switzerland (October 2024)
29. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
30. Xu, R., Yang, S., Wang, Y., Cai, Y., Du, B., Chen, H.: Visual mamba: A survey and new outlooks (2024)