KMUNet: A novel medical image segmentation model based on KAN and Mamba

Hongsheng Zhang[1[0009-0005-1039-6170]], Yuting Duan[1[0009-0001-4648-8270]], Ting Liu[1,2[0000-0002-8468-4926]], Weifeng Zhang [3(✉)[0000-0001-9273-7197]], and Hongzhong Tang [1(✉) [0000-0003-4703-4972]]

[1] Xiangtan University, 411105, China
[2] the College of Electronic Science, National University of Defense Technology, China
[3] The First Affiliated Hospital of Shantou University Medical College, 515041, China
18wfzhang@stu.edu.cn, diandiant@126.com

Abstract. Medical image segmentation is essential for identifying lesion regions and diagnosing disease. Convolutional neural networks (CNNs) and transformer-based models often struggle to effectively capture both local details and global contextual features in medical images, leading to a decline in segmentation performance. To address this problem, a novel medical image segmentation model, KMUNet, is proposed by integrating Kolmogorov-Arnold networks (KAN) and Mamba based on the traditional U-shape architecture. This model employs a CNN-based encoder to extract local features and integrates a State Space Model-based Mamba module in the decoder to capture long-range dependencies. Initially, a global downsampling module, called KAN-PatchEmbed is presented. This module differs from traditional convolutional operations in utilizing an interval sampling strategy to alleviate the loss of feature information and KAN to reduce computational complexity, respectively. Furthermore, the Kolmogorov-Arnold Spatial-Channel Attention module is designed for skip connections, where KAN is employed to allocate the weight of the current channel by aggregating features across all stages. Finally, the proposed model was evaluated on three publicly available datasets. Experimental results reveal that KMUNet outperforms other models in segmentation tasks and produces more visually appealing segmentation results. Our code is available at https://github.com/zhang-hongsheng/KMUNet.

Keywords: Medical Image Segmentation, State Space Model, KANs

1 Introduction

Medical image segmentation is a crucial aspect of medical image analysis, focused on automatically identifying lesion regions in medical images. Over the years, widespread deep-learning approaches have been proposed in medical image segmentation tasks [4][5]. UNet [1], a prominent architecture of convolutional neural network (CNN) based deep learning model, established a U-shaped encoder-decoder and proved to be highly suitable for medical image segmentation tasks. Numerous studies have explored

CNNs [2][3], successfully capturing local features in medical image classification and segmentation. However, a limitation of CNNs is the inability to effectively capture global features in images, leading to error predictions for small target regions. To address this problem, a context encoder network (CE-Net) [4] was presented by leveraging dilated convolutions and multi-scale pooling operations to extract contextual features in medical image. Moreover, a novel context pyramid fusion network (CPFNet) was presented by combining two pyramidal modules to fuse global/multi-scale context information. Furthermore, Xue et al. [6] proposed an adversarial network with multi-scale l_1 loss to capture global and local features. However, these approaches still struggle to address the need for global modeling in medical images.

Recently, state space models (SSMs) have achieved significant advancements and demonstrated superior performance in effectiveness and computational efficiency for long-sequence modeling [7][8]. Mamba [9], a network based on SSMs, has demonstrated competitive performance compared to traditional SSMs by utilizing an efficient selective scanning mechanism for global modeling in visual tasks. Vision Mamba [10] introduced a cross-scanning module to improve global modeling capabilities in visual applications. However, Mamba-based methods face challenges in capturing local features in medical image segmentation. Our preliminary experiments reveal that Mamba-based models often lead to under-segmentation due to insufficient attention to local details lesion boundaries and small regions.

Existing CNN-based and Mamba-based models have made significant progress in medical image segmentation. However, the lack of interpretability in these models limits their application in clinical decision-making [28]. Recently, the Kolmogorov–Arnold Network (KAN) [11], based on the Kolmogorov-Arnold theorem, replaced traditional Multilayer Perceptron (MLP) with learnable activation functions, aiming for an enhanced interpretable neural network. Consequently, we utilized KAN to enhance the interpretability of the model and segmentation performance, thereby improving its potential for application in clinical diagnosis.

Currently, numerous studies have focused on the role of multi-scale and multi-stage information in medical image segmentation [5][16]. Specifically, literature [16] combined four modules with their U-shape architecture for a light-weight medical image segmentation model, called MALUNet. This model introduced a Channel Attention Bridge (CAB) block to integrate cross-stage information from various stages along the channel dimension. While CAB achieved multi-stage feature fusion through MLP, it has two limitations: (1) weak interpretability of computational mechanisms, and (2) quadratic growth in computational complexity as feature dimensions increase.

Based on the above discussions, we propose a unified medical image segmentation model, called KMUNet based on the strengths of both KAN and Mamba. KMUNet utilizes CNN-based encoding to extract local features of images and employs Mamba decoding for global modeling. The main contributions of this paper are as follows:

• We propose a medical image segmentation model named KMUNet, which integrates KAN and Mamba based on the UNet architecture.

• We design KAN-PatchEmbed as a global downsampling operation, effectively reducing the loss of feature information through interval sampling. Moreover, it improves the feature extraction capabilities of KAN.

• We propose a Kolmogorov-Arnold Spatial-Channel Attention Block (KAN-SCA), which enhances multi-scale local features and integrates multi-stage global contextual information during decoding. Particularly, we employ KAN to capture global contextual relationships, thereby allocating weights to feature maps.

• The proposed model was compared with 15 models on three publicly available datasets, demonstrating its superior performance quantitatively and qualitatively. It has been proven to be a versatile model capable of segmenting various types of medical images.

2 Methodology

2.1 KMUNet Architecture

Our proposed KMUNet model is a four-layer U-shaped architecture with encoder, decoder, and skip-connect parts, as shown in Figure 1 (b). The proposed KMUNet employs a CNN-based encoder to extract local features. These convolutional operations hierarchically capture local features in images, with lower layers focusing on fine-grained local details and deeper layers encoding higher-level semantic information. In the decoder, we introduce Mamba modules for global modeling to effectively capture long-range dependencies, which differs from traditional U-Net. This design preserves sensitivity to local details while significantly modeling the semantic correlation between pathological regions and surrounding tissues. The architecture is appropriate for medical image segmentation tasks involving irregular shapes and poorly defined boundaries.

2.2 KAN-PatchEmbed

Vision Transformer (ViT) [12] introduced the PatchEmbed module, which utilizes a 4×4 convolution kernel with a stride of 4 to extract image information and reduce the image size. However, the large-stride convolution operations during downsampling may result in irreversible loss of spatial information, particularly in edge details that are crucial for medical image analysis. Therefore, we propose KAN-PatchEmbed for downsampling to resolve the loss of feature information and reduce the computational cost of the model. As displayed in Figure 1 (a), the input image $X \in \mathbb{R}^{H\times W\times C}$ is sampled at intervals of four pixels both in rows and columns. For instance, H=W=8, and the formula is defined as follows:

$$\{X_0, X_1, ..., X_{15}\} = \text{downsampling}(X) \tag{1}$$

All sampled images are concatenated as $X^* \in \mathbb{R}^{H/4\times W/4\times 16C}$,

$$X^* = \text{Concat}\{X_0, X_1, ..., X_{15}\} \tag{2}$$

where Concat is the concatenation operation.

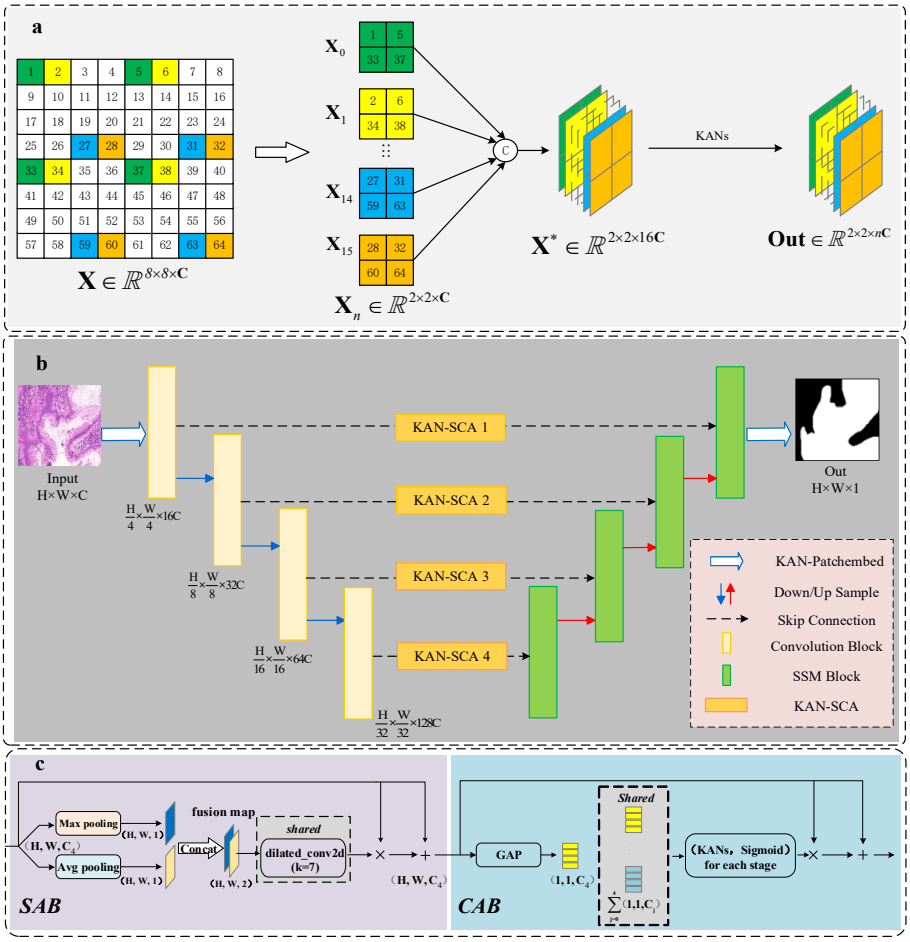

Fig. 1. (a) KAN-PatchEmbed module. (b) the proposed KMUNet model. (c) KAN-SCA, Kolmogorov-Arnold Spatial-Channel Attention module.

Finally, we utilize KANs to map X^* to a high-dimensional space by adjusting the numbers of channels. The KAN-PatchEmbed expression can be represented as follows:

$$Out = \mathrm{LN}[\mathrm{KANs}(X^*)] \tag{3}$$

where LN is the LayerNorm operation, and n is determined by the number of feature channels required by the first convolutional layer.

2.3 KAN-SCA

We design a KAN-SCA module to fuse multi-scale and multi-stage global contextual information, as illustrated in Figure 1 (c). Unlike MALUNet [16], we replace the MLP with KAN. For instance, in the fourth-stage KAN-SCA module, we first incorporate

the Spatial Attention Bridge (SAB). The SAB enables the model to focus on critical spatial features and suppress irrelevant background information. To achieve this, max pooling and average pooling are initially operated on the feature map, and the resulting feature maps are then concatenated. Subsequently, we utilize a shared dilated convolution operation to fuse feature maps and generate a spatial attention map by a sigmoid function. Finally, the element-wise multiplication operation is performed between the original image and the spatial attention map, adding this result to the residual information to produce the final spatial attention map.

The novelty of KAN-SCA is its initial integration of KAN into the Channel Attention Bridge (CAB) to provide the interpretability. This model utilizes KANs to generate channel attention maps, which are subsequently utilized to guide residual feature fusion in subsequent layers via adaptive weight allocation. As a result, CAB enriches the multi-stage features and allows the model to focus on essential channels by suppressing channels with low context correlation. This process can be expressed using the following formulas:

$$y = \text{Concat}(\text{AvgPool}(x_i)), i \in \{1,2,3,4\} \tag{4}$$

$$att_4 = \sigma(\text{KANs}(\text{Conv1D}(y))), i \in \{1,2,3,4\} \tag{5}$$

$$\text{Out}_4 = x_4 + att_4 \cdot x_4 \tag{6}$$

where AvgPool refers to global average pooling, and Conv1D represents 1D convolution operation.

By utilizing the powerful fitting ability of KAN, our proposed KAN-SCA module effectively fuses multi-scale and multi-stage information during decoding, thereby producing superior segmentation performance.

3 Experiment

3.1 Dataset

To validate the effectiveness and generalization capability of KMUNet, we evaluated our network using three public datasets. CVC-ClinicDB [25] comprises 612 images from 31 colonoscopy sequences. The Glas [26] contains 165 images of colorectal adenocarcinoma. Additionally, the BUSI [27] includes ultrasound images of breast cancer pathology along with their corresponding segmentation maps. BUSI contains 210 malignant breast ultrasound images, and we split all datasets into 70% training and 30% test. Images were resized to 256×256.

3.2 Competition methods and implementation details

To verify the effectiveness of our KMUNet, we compared it with the following models: UNet [1], UNet++ [13], SCR-Net [14], U-Next [15], MALUNet [16], Meta-UNet [2], ACC-UNet [17], Rolling-Unet (Roll-UNet) [18], MHorUNet [3], ATTENTION SWIN

U-NET (Att-UNet) [24], VM-UNet [19], H-vmunet [20], SkinMamba(skinmamba) [21], U-Mamba [22], and U-KAN [23]. Our experiment was conducted on an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The proposed model utilized the AdamW optimizer with an initial learning rate of 0.001 and a batch size of 16 on the CVC-ClinicDB and Glas datasets. On the BUSI dataset, the model used the Adam optimizer with a cosine annealing learning rate scheduler, an initial learning rate of 1e-4, and a batch size of 8， other hyperparameters follow PyTorch defaults, all trained for 400 epochs. The loss function integrates cross-entropy and Dice losses with 1:1. All models were evaluated using three metrics: Dice, Sensitivity (Sen), and Precision (Pre).

3.3 Comparison with Several State-of-the-art Segmentation Models

Table 1 reports a quantitative results comparison with state-of-the-art methods on three datasets. The best results are indicated in bold, and the suboptimal results have been underlined. Our KMUNet outperforms other models on the CVC-ClinicDB dataset and achieves the highest Dice score on three datasets. Additionally, it maintains the best Sensitivity on the CVC-ClinicDB and Glas datasets, and the highest Precision on the CVC-ClinicDB and BUSI datasets.

Table 1. Quantitative comparison with state-of-the-art methods on the three datasets.

Methods	CVC-ClinicDB			Glas			BUSI		
	Dice	Sen	Pre	Dice	Sen	Pre	Dice	Sen	Pre
UNet	0.9102	0.9245	0.8975	0.9237	0.9271	0.9203	0.6923	0.6601	0.7521
UNet++	0.9033	0.9084	0.9112	0.9347	0.9255	0.9441	0.7220	0.7003	0.7688
SCR-Net	0.9229	0.9353	0.9109	0.9452	0.9400	<u>0.9506</u>	0.7168	0.7130	0.7453
Att-UNet	0.7212	0.6851	0.7660	0.8915	0.9074	0.8969	0.4777	0.4904	0.5015
U-Next	0.9171	0.9176	0.9181	0.9437	0.9401	0.9475	0.7395	0.7324	0.7636
MALUNet	0.8753	0.9124	0.842	0.9140	0.9189	0.9096	0.6866	0.6815	0.6991
Meta-UNet	0.9301	0.9210	0.9394	0.9190	0.9170	0.9211	0.7575	**0.7474**	0.7816
ACC-UNet	0.9290	**0.9361**	0.9232	0.9442	<u>0.9441</u>	0.9444	0.7363	0.7124	0.7917
Roll-Unet	0.9298	<u>0.9359</u>	0.9238	<u>0.9465</u>	0.9422	**0.9510**	0.7327	0.7355	0.7425
VM-UNet	0.8845	0.8741	0.8961	0.9393	0.9430	0.9358	0.7455	0.6923	<u>0.8377</u>
H-vmunet	0.8977	0.8924	0.9056	0.9332	0.9233	0.9434	0.7382	0.7387	0.7559
MHorUNet	0.9112	0.9325	0.8925	0.9207	0.9051	0.9372	0.7318	0.7109	0.7596
SkiMamba	0.9194	0.9064	0.9332	0.9394	0.9354	0.9436	0.7563	0.7085	0.8142
U-Mamba	0.9161	0.9162	0.9170	0.9419	0.9389	0.9449	<u>0.7677</u>	<u>0.7424</u>	0.8000
U-KAN	<u>0.9371</u>	0.9307	<u>0.9439</u>	0.9411	0.9358	0.9464	0.7638	0.7371	0.8101
Our Model	**0.9384**	0.9325	**0.9451**	**0.9485**	**0.9504**	0.9466	**0.7731**	0.7221	**0.8463**

Figure 2, Figure 3, and Figure 4 illustrate the visual comparison results on the CVC-ClinicDB, Glas, and BUSI datasets, respectively, where false positives are denoted in

red and false negatives in green. As illustrated in these figures, our proposed KMUNet demonstrates excellent performance on visual results, preserving fine-grained structure and capturing subtle details, particularly in complex regions and boundaries.

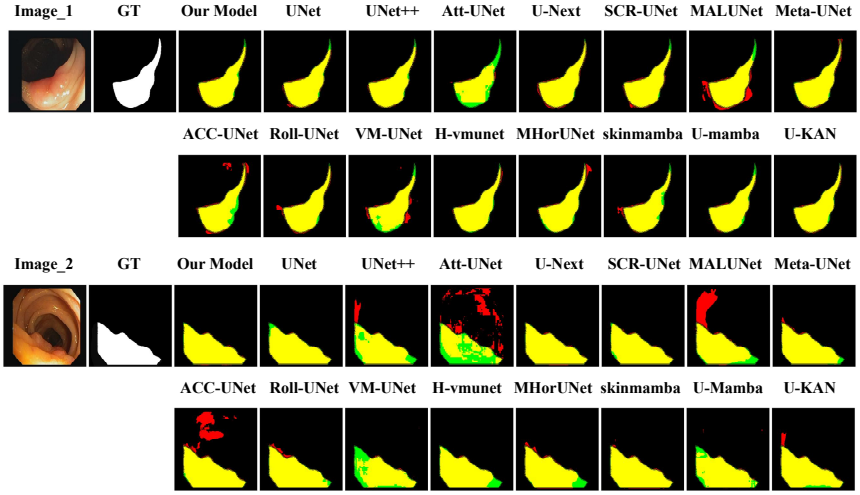

Fig. 2. Segmentation results of different methods applied to the CVC-ClinicDB dataset. Red indicates false positives, while green indicates false negatives.

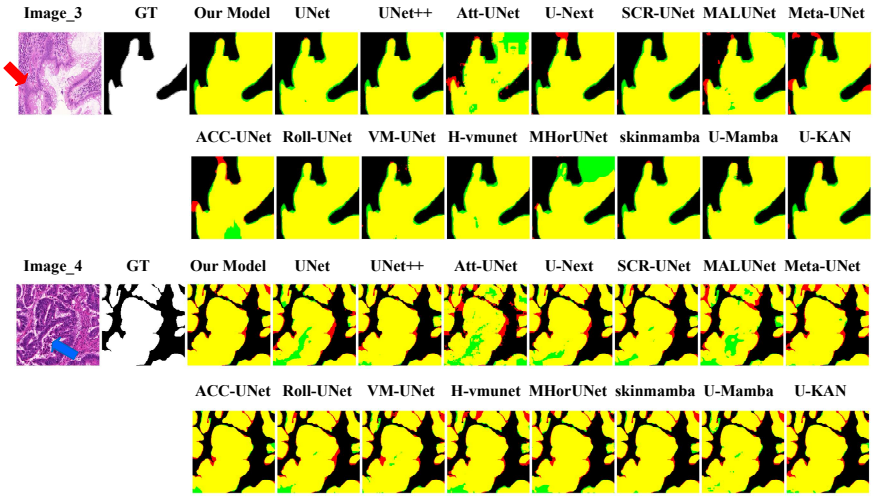

Fig. 3. Segmentation results of different methods applied to the Glas dataset.

As shown in Figure 3, in the area pointed by the red arrow in Image_3, CNN-based models (i.e. Att-UNet, U-Next, MALUNet, Meta-UNet, and ACC-UNet) produce more false positive regions. In the area indicated by the blue arrow in Image_4, Mamba-based and CNN-based models both generate false negatives, incorrectly classifying white pixels within the lesion area as normal regions. Unlike CNN-based or Mamba-based models, our KMUNet achieves the lowest false predictions and demonstrates high similarity

with the ground truth. These results suggest the abilities of our model to effectively capture local details and global contextual features in medical images, supporting the benefits of integrating CNN-based encoding and Mamba-based decoding.
To further investigate the effectiveness and generalization ability of our KMUNet, we conducted experiments on the BUSI dataset, which presents greater challenges in medical image segmentation tasks compared to the former two datasets. From the results of Image_5, we observe that CNN-based and Mamba-based models struggle with under-segmentation and over-segmentation. Our KMUNet outperforms the other models, with fewer false positives and clearer boundaries.

In addition, our KMUNet, with approximately 10M parameters and 3 GFLOPs, achieves an average reduction of 67% in parameter count and 94% in computational cost compared to UNet. Furthermore, KMUNet demonstrates superior computational efficiency over other Mamba- or Transformer-based models

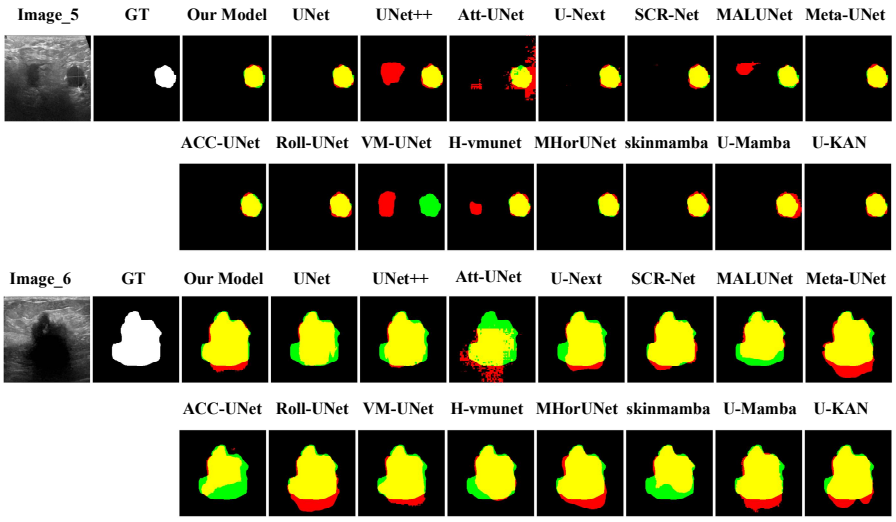

Fig. 4. Segmentation results of different methods applied to the BUSI dataset.

4 Conclusion

This article proposes a medical image segmentation model, KMUNet, which effectively combines the strengths of CNNs and Mamba to capture local and global features simultaneously. We first design a global downsampling (KAN-PatchEmbed) module to reduce the loss of feature information and computational complexity. Additionally, a KAN-SCA module is developed to effectively enhance multi-scale local features and capture correlation information from multi-stage information during decoding. Experiments on three public datasets demonstrate that KMUNet shows significantly improved segmentation results and generalization capabilities compared to other models.

Acknowledgments. We are especially grateful to the related experts from Shantou University Medical College and the First Affiliated Hospital of Shantou University Medical College, Yiqun Geng, and Liangli, Hong.

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

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing (2015)
2. Wu, H., Zhao, Z., Wang, Z.: META-Unet: Multi-scale efficient transformer attention Unet for fast and high-accuracy polyp segmentation. IEEE Transactions on Automation Science and Engineering (2023)
3. Wu, R., Liang, P., Huang, X., Shi, L., Gu, Y., Zhu, H., Chang, Q.: MHorUNet: High-order spatial interaction UNet for skin lesion segmentation. Biomedical Signal Processing and Control, 88, 10551 (2024)
4. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging, 38(10), 2281-2292 (2019)
5. Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., Chen, X.: CPFNet: Context pyramid fusion network for medical image segmentation. IEEE transactions on medical imaging, 39(10), 3008-3018 (2020)
6. Xue, Y., Xu, T., Zhang, H., Long, L. R., Huang, X.: Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. Neuroinformatics, 16, 383-392 (2018)
7. Dao, T., Gu, A.: Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060 (2024)
8. Gupta, A., Gu, A., Berant, J.: Diagonal state spaces are as effective as structured state spaces. Advances in Neural Information Processing Systems 35, 22982–22994 (2022)
9. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
10. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
11. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024)
12. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging, 39(6), 1856-1867 (2019)
14. Wu, H., Zhong, J., Wang, W., Wen, Z., Qin, J.: Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 4, pp. 2916-2924) (2021)

15. Song, T., Meng, F., Rodriguez-Paton, A., Li, P., Zheng, P., Wang, X.: U-next: A novel convolution neural network with an aggregation u-net architecture for gallstone segmentation in ct images. IEEE Access, 7, 166823-166832 (2019)
16. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation (2022)
17. Ibtehaz, N., Kihara, D.: Acc-unet: A completely convolutional unet model for the 2020s. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 692-702). Cham: Springer Nature Switzerland (2023)
18. Liu, Y., Zhu, H., Liu, M., Yu, H., Chen, Z., Gao, J.: Rolling-Unet: Revitalizing MLP's Ability to Efficiently Extract Long-Distance Dependencies for Medical Image Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 4, pp. 3819-3827) (2024)
19. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
20. Wu, R., Liu, Y., Liang, P., Chang, Q.: H-vmunet: High-order vision mamba unet for medical image segmentation. Neurocomputing, 129447 (2025)
21. Zou, S., Zhang, M., Fan, B., Zhou, Z., Zou, X.: SkinMamba: A Precision Skin Lesion Segmentation Architecture with Cross-Scale Global State Modeling and Frequency Boundary Guidance. arXiv preprint arXiv:2409.10890 (2024)
22. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
23. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., Chen, Z., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. arXiv preprint arXiv:2406.02918 (2024)
24. Aghdam, E. K., Azad, R., Zarvani, M., Merhof, D.: Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE (2023)
25. Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007 (2015)
26. K. Sirinukunwattana, J. P. W. Pluim, H. Chen, X Qi, P. Heng, Y. Guo, L. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. Ben Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. Snead, N. M. Rajpoot, "Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest" http://arxiv.org/abs/1603.00275
27. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief, 28, 104863 (2020)
28. Vellido, A.: The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications, 32(24), 18069-18083 (2020)