MMU-KAN: Multi-path Mamba Fusion-based U-KAN For Medical Image Semantic Segmentation

Haifeng Li[1], Wei Tan[1*], Weichun Bu[1], Hengshuai Fan[1], Weiwei Sun[1], Yue Du[2*], and Yuxuan Guo[1]

[1] School of Mathematics and Statistics, Fuyang Normal University, Fuyang 236037, China
[2] School of Computer and Information Science, Qinghai Institute of Technology, Xining 810016, China
[*] Corresponding author. E-mail: tanweilina@163.com and duyue2007new@163.com

Received: Apr. 14, 2026; Accepted: May 05, 2026

Current U-Net based medical image segmentation methods often rely on stacked static nonlinear functions to model inter-pixel relationships, resulting in suboptimal segmentation performance. To address this, we propose a multi-path Mamba fusion-based U-KAN medical image segmentation method. Firstly, the U-Net architecture is reconstructed using Kolmogorov-Arnold Networks to enhance the nonlinear modeling capacity. Then, a dualpath encoder is designed to extract multi-granularity semantic representations, and the Mamba is introduced to achieve adaptive cross-path feature fusion. Finally, a multi-stage supervised decoding process generates the precise segmentation results. Experiments on three medical datasets demonstrate that the proposed method outperforms state-of-the-art approaches in both Dice and mIoU metrics, confirming its superiority.

Keywords: Semantic segmentation; U-KAN architecture; Mamba fusion; multi-stage supervised segmentation

http://dx.doi.org/10.6180/jase.202610_33.024

1. Introduction

Precise semantic segmentation of medical images stands as a cornerstone in the contemporary digital healthcare paradigm [1–3], providing critical morphological evidence for computer-aided diagnosis, treatment planning, and intraoperative navigation [4, 5]. By partitioning complex anatomical structures or pathological regions, such as tumors, vascular networks, and organ boundaries, into biologically meaningful subsets, segmentation enables clinicians to perform quantitative analysis of volumetric changes and spatial correlations [6–8]. However, the inherent characteristics of medical modalities, including low contrast-to-noise ratios in Ultrasound, intensity inhomogeneities in MRI, and the pervasive presence of speckle noise, pose significant challenges to robust boundary delineation [9–11]. Furthermore, the substantial inter-patient variability in anatomical topology and the subtle semantic differences between healthy and lesion tissues necessitate highly sophisticated feature representations [12–14]. While traditional manual contouring remains a gold standard, it is increasingly becoming a bottleneck in clinical workflows due to its labor-intensive nature and susceptibility to inter-observer variability [15–17]. Consequently, developing automated, high-precision segmentation frameworks that can capture both intricate local textures and long-range contextual dependencies has become an imperative frontier in medical image analysis.

In recent years, various medical image semantic segmentation models based on the U-Net architecture have evolved into three primary categories: CNN-based U-Net, Transformerbased U-Net, and Mamba-based U-Net models [18–20]. CNN-based U-Net models focus on constructing convolutional encoder-decoder architectures with diverse skip connections. These models leverage the powerful in-

ductive bias of convolutional kernels to capture detailed boundaries and textural information for segmentation. For instance, Shu et al. propose a dual-attention-guided convolutional U-Net, which employs joint channel and spatial attention within skip connections across different layers to adaptively fuse low-level structural details with high-level semantic information [21]. Rahman et al. embedded a multi-scale convolutional attention decoder into the U-Net framework, utilizing a grouped gating mechanism based on channel and spatial attention to enhance the capture of complex pixel interdependencies during decoding [22]. Similarly, Dai et al. introduce a dual-stream convolutional U-Net that facilitates the reuse and re-exploration of historical features through cross-path and cross-layer interactions, thereby extracting comprehensive information to improve segmentation performance [23]. Transformer-based U-Net models aim to leverage self-attention mechanisms to extract global dependencies between pixels. Yin et al. propose a U-Net model with a dual-encoder and single-decoder structure; the former captures longrange dependencies in semantic encoding at a low computational cost via lightweight multihead self-attention, while the latter utilizes graph feature pyramids and Kolmogorov-Arnold Networks to enhance image resolution recovery [1]. Lin et al. construct a U-Net driven by dual-stream encoders and a single-stream decoder, using Transformers to adaptively fuse contextual information of varying granularities [24]. Hatamizadeh et al. presented a pure Transformer U-Net, where window-based and shifted-window self-attention are employed in both the encoder and decoder to model global pixel dependencies [25]. Mamba-based U-Net models introduce State Space Models (SSMs) as the foundational building blocks for U-shaped architectures, capturing long-range dependencies across pixels with linear computational complexity. For example, Wu et al. proposed a selective scanning Mamba U-Net that enhances performance by leveraging high-order feature interactions between local and global scans [26]. Kui et al. designed a dual-encoder U-Net combining Mamba and CNN to capture both global and local information, utilizing multi-stage attention to fuse complementary cross-path information and strengthen feature representation [27].

Despite their success, these U-Net variants typically rely on stacking sequences of static nonlinear activation functions within deep neural networks to model the intricate interactions between image pixels. Such a paradigm often leads to suboptimal semantic segmentation performance due to the constrained flexibility of traditional MLP-based structures. Inspired by the robust and adaptive functional approximation capabilities of Kolmogorov-Arnold Networks (KAN), we propose a novel Multi-path Mamba-fused U-KAN framework for medical image semantic segmentation. In this approach, we re-architect the U-Net encoder-decoder backbone using KAN to significantly enhance the learning of complex nonlinear patterns among pixels. Furthermore, a multi-path encoding mechanism is designed to extract multi-scale semantic representations, which are subsequently aggregated via Mamba blocks. This fusion strategy strengthens the skip connections by effectively integrating cross-path semantic contexts, thereby advancing the state-of-the-art in high-precision medical image segmentation.

The contributions are as follows: (1) We propose a novel U-shaped segmentation framework based on KAN, which replaces traditional static linear layers with learnable activation functions to enhance non-linear modeling. (2) We design a dual-path encoding mechanism that processes multi-resolution inputs to simultaneously capture coarse-grained anatomical contours and fine-grained textural details. (3) We introduce a Mamba selective state spacebased fusion module to achieve dynamic, content-aware cross-path interaction, significantly strengthening feature representation during skip connections. (4) We demonstrate through extensive experiments on three heterogeneous datasets that our MMU-KAN outperforms state-of-the-art methods in both accuracy and robustness.

2. Method

As illustrated in Fig. 1, the proposed Multi-path Mamba-fused U-KAN (MMU-KAN) framework for medical image semantic segmentation primarily consists of three core modules: the multi-path semantic representation extraction module, the cross-path adaptive fusion module, and the multi-stage supervised decoding segmentation module.

2.1. Multi-path Semantic Representation Extraction

This module aims to capture multi-granularity semantic representations by constructing two parallel encoders to enhance the model's understanding of medical imagery. Specifically, the Kolmogorov-Arnold Network is introduced to architect the U-KAN encoders. Each medical image is resized into sub-images of varying resolutions and fed into these encoders to extract both coarse-grained and fine-grained semantic features. These features provide rich and robust semantic information for the subsequent decoding process via skip connections.

U-KAN Encoder: The encoder is composed of three convolutional stages followed by two KAN stages. Each convolutional stage includes a convolutional layer (*Conv*), a normalization layer (*Norm*), and a pooling layer (*Pool*),

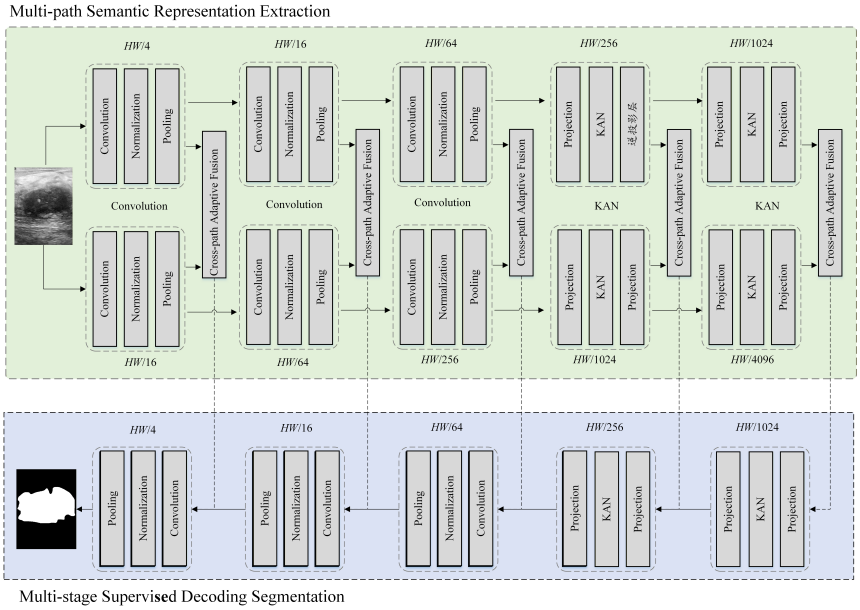

Fig. 1. The illustration of MMU-KAN. It consists of the multi-path semantic representation extraction module, the cross-path adaptive fusion module, and the multi-stage supervised decoding segmentation module.

where the convolutional kernel size is 3×3 with a stride of 1 , and the pooling layer size is 2×2. Given an input medical image $X \in R^{H \times W \times C}$, the computational flow of the i-th convolutional stage is defined as:

$$X_i = \text{Pool}\left(\text{Norm}\left(\text{Conv}\left(X_{i-1}\right)\right)\right) \tag{1}$$

Here, X_i and X_{i-1} represent the output and input of the i-th convolutional stage, respectively, where $i \in \{1, 2, 3\}$. The final output of the convolutional stages in the U-KAN encoder is X_3. Each KAN stage consists of a vector projection layer $VP(\cdot)$, a KAN layer $\Phi(\cdot)$, and a vector inverse projection layer VIP$(\cdot)$. The computational flow of the j-th KAN stage is as follows:

$$Z_j = VIP\left(\Phi\left(VP\left(Z_{j-1}\right)\right)\right) \tag{2}$$

where Z_j and Z_{j-1} denote the input and output of the j-th KAN stage, respectively. $j \in \{1, 2\}$ and $Z_0 = X_3$. The specific mathematical formulation of the KAN layer $\Phi(\cdot)$ is given by:

$$\Phi\left(Z_{j-1}\right) = \begin{Bmatrix} \phi_{j,1,1}(\cdot) & \cdots & \phi_{j,1,n_{j-1}}(\cdot) \\ \vdots & \vdots & \vdots \\ \phi_{j,n_j,1}(\cdot) & \cdots & \phi_{j,n_j,n_{j-1}}(\cdot) \end{Bmatrix} Z_{j-1} \tag{3}$$

where n_j and n_{j-1} represent the number of neurons in the j-th and $(j-1)$-th KAN layers.

By adopting learnable and parameterized activation functions as edges and weights instead of traditional linear weight matrices, KAN significantly enhances the model's capacity to capture complex nonlinear relationships between pixels in medical images while improving interpretability and flexibility.

Multi-granularity representation learning: Medical images encompass rich semantics ranging from large-scale tissue contours to cell-level details. Feature extraction at a single granularity often struggles to balance global and local information, leading to imprecise segmentation results. To address this, multi-granularity representation learning is constructed by designing sub-image inputs with different resolutions to extract both coarse-grained and fine-grained semantic representations. This approach provides rich and robust semantic information for subsequent segmentation stages.

First, a global U-KAN encoder is defined for the high-resolution image branch to extract coarse-grained representations. The inputs for the five representation extraction stages are defined as $H \times W$, $(H/2) \times (W/2)$, $(H/4) \times (W/4)$, $(H/8) \times (W/8)$, and $(H/16) \times (W/16)$. Next, a local U-KAN encoder is defined for the low-resolution image branch to extract finegrained representations. The inputs for its five stages are defined as $(H/2) \times (W/2)$, $(H/4) \times (W/4)$, $(H/8) \times (W/8)$, $(H/16) \times (W/16)$, and $(H/32) \times (W/32)$. Through this method, two different resolutions of semantic representations can be obtained at each extraction stage for skip connections, greatly enriching the semantic information required for the decoding process. Finally, the

semantic representations of the five stages for the global and local U-KAN encoders are defined as follows:

$$\begin{aligned} F^H &= \left\{X_1^H, X_2^H, X_3^H, Z_1^H, Z_2^H\right\} \\ F^L &= \left\{X_1^L, X_2^L, X_3^L, Z_1^L, Z_2^L\right\} \end{aligned} \tag{4}$$

where F^H and F^L represent the sets of semantic representations captured by the global and local U-KAN encoders across the five stages, respectively.

2.2. Cross-path Adaptive Fusion

Cross-path adaptive fusion aims to effectively and efficiently aggregate multi-granularity semantic representations extracted from the dual-path encoders. Therefore, the Mamba selective state space model is introduced to construct a stage-wise, refined interaction mechanism for dynamic cross-path semantic aggregation. Cross-path adaptive fusion consists of a Mamba interaction phase and a Mamba aggregation phase. The Mamba interaction phase simulates a cross-attention mechanism, using the semantic representation of one path to enhance the expression of the other. The Mamba aggregation phase performs concatenation on the enhanced semantic representations and applies a selective scanning mechanism to obtain the final skip connection representation. Given the coarse-grained representation F_k^H and fine-grained representation F_k^L extracted from the k-th stage of the dual-path encoders, the calculation process for the k-th stage skip connection is as follows:

$$\begin{aligned} \bar{F}_k^H, \bar{F}_k^L &= \mathrm{CroMB}\left(F_k^H, F_k^L\right) \\ H_k &= \mathrm{FusMB}\left(\left[\bar{F}_k^H, \bar{F}_k^L\right]\right) \end{aligned} \tag{5}$$

where $\bar{F}_k^H$ and $\bar{F}_k^L$ denote the two granularities of semantic representations enhanced by the Mamba interaction stage $\mathrm{CroMB}(\cdot).[\cdot]$ represents the concatenation operation, and H_k represents the k-th stage skip connection representation generated by the Mamba aggregation stage $FusMB(\cdot)$.

Mamba interaction: This phase aims to enhance two different granularities of semantic representations through a cross-selective scanning mechanism, thereby achieving mutual reinforcement. Specifically, for the coarse-grained representation F_k^H and fine-grained representation F_k^L at the same encoding stage, the $Flatten(\cdot)$ operation is first used to convert 2D spatial feature maps into 1D sequences to adapt to the sequential processing characteristics of the State Space Model:

$$S_k^H = \mathrm{Flatten}\left(F_k^H\right), \quad S_k^L = \mathrm{Flatten}\left(F_k^L\right) \tag{6}$$

Then, based on the 1D sequences, linear projection layers are used to dynamically generate State Space Model (SSM) parameters. Specifically, the SSM parameter set for S_k^H is $\left\{B_k^H, C_k^H, \Delta_k^H\right\}$, and for S_k^L it is $\left\{B_k^L, C_k^L, \Delta_k^L\right\}$. Discretization is performed to convert the continuous-time system into a discrete-time system:

$$\begin{aligned} \bar{A}_k^H &= \exp\left(\Delta_k^H A_k^H\right), \quad \bar{A}_k^L = \exp\left(\Delta_k^L A_k^L\right) \\ \bar{B}_k^H &= \exp\left(\Delta_k^H A_k^H\right), \quad \bar{B}_k^L = \exp\left(\Delta_k^L A_k^L\right) \end{aligned} \tag{7}$$

where A_k^H and A_k^L are predefined learnable state transition matrices. Next, for each time step t, the following calculation process is executed sequentially, achieving cross-modal information exchange by swapping the output matrix C :

$$\begin{aligned} T_k^H(t) &= \bar{A}_k^H T_k^H(t-1) + \bar{B}_k^H S_k^H(t), O_k^H(t) = C_k^L T_k^H(t) \\ T_k^L(t) &= \bar{A}_k^L T_k^L(t-1) + \bar{B}_k^L S_k^L(t), O_k^L(t) = C_k^H T_k^L(t) \end{aligned} \tag{8}$$

During this process, $T_k^H(t)$ and $T_k^L(t)$ represent the hidden state vectors at time step t, while $S_k^H(t)$ and $S_k^L(t)$ represent the feature vectors of the input sequence at time step t. After T steps, the $\mathrm{Reshape}(\cdot)$ operation is utilized to map the 1D sequences back to 2D spatial feature maps $\bar{F}_k^H$ and $\bar{F}_k^L$.

Mamba aggregation: This phase first fuses the cross-enhanced semantic representations of different granularities into a unified representation and employs a bidirectional scanning mechanism to fully capture contextual information within the sequence. Specifically, the $Flatten(\cdot)$ operation is used to convert the 2D spatial feature maps into 1D sequences, and a concatenation operation is executed to obtain an initial fused representation:

$$Q_k = \left[\mathrm{Flatten}\left(\bar{F}_k^H\right), \mathrm{Flatten}\left(\bar{F}_k^L\right)\right] \tag{9}$$

To capture bidirectional context, the $Reverse(\cdot)$ operation is used to generate the reverse sequence Q_k^R, i.e., $Q_k^R = \mathrm{Reverse}\left(Q_k\right)$. The standard State Space Model is applied to the forward and reverse sequences respectively for feature extraction:

$$\hat{Q}_k = \mathrm{SSM}\left(Q_k\right), \quad \hat{Q}_k^R = \mathrm{SSM}\left(Q_k^R\right) \tag{10}$$

where $\mathrm{SSM}(\cdot)$ represents the complete forward propagation process of the State Space Model, including parameter generation, discretization, and sequence scanning. The processed forward and reverse sequences are integrated via average fusion, and the $Reshape(\cdot)$ operation is used to map the 1 D sequence back to the 2 D spatial feature map H_k:

$$H_k = \mathrm{Reshape}\left(\hat{Q}_k + \hat{Q}_k^R\right) \tag{11}$$

2.3. Multi-stage Supervised Decoding Segmentation

To improve segmentation accuracy, a deep hierarchical loss is proposed to force the model to maintain prediction consistency across different resolutions. This enhances the collaborative optimization of high-level semantics and low-level structures during image segmentation. Specifically, supervisory signals are added at three stages to train the overall segmentation network: the final output of the U-KAN encoder, the output of the first stage of the U-KAN decoder, and the final output of the U-KAN decoder. The deep hierarchical loss is defined as follows:

$$L = \lambda_1 L_1\left(Y, \bar{Y}_1\right) + \lambda_2 L_2\left(Y, \bar{Y}_2\right) + \lambda_3 L_3\left(Y, \bar{Y}_3\right) \tag{12}$$

where $\lambda_1 = 0.2, \lambda_2 = 0.2$, and $\lambda_3 = 0.6$ are the hyper-parameters for the supervision loss at the three stages. Y represents the ground truth segmentation mask. $\bar{Y}_1$ denotes the predicted segmentation mask generated from the final output of the U-KAN encoder $H_5 \cdot \bar{Y}_2$ denotes the predicted segmentation mask generated from the first stage output of the U-KAN decoder $\bar{H}_1$. $\bar{Y}_3$ denotes the predicted segmentation mask generated from the final output of the U-KAN decoder $\bar{H}_5$. Each loss $L_i\left(Y, \bar{Y}_i\right)$ consists of two sub-losses:

$$L_i\left(Y, \bar{Y}_i\right) = L_i^{IoU}\left(Y, \bar{Y}_i\right) + L_i^{bce}\left(Y, \bar{Y}_i\right) \tag{13}$$

Among them, $L^{IoU}(\cdot)$ denotes the weighted Intersection over Union (IoU) loss, used to measure the overlap between the predicted and ground truth boundaries. $L^{bce}(\cdot)$ denotes the weighted binary cross-entropy loss.

3. Results and discussion

3.1. Set up

Dataset: Three heterogeneous medical datasets were selected as benchmark datasets to verify the superiority and effectiveness of the proposed method in medical image semantic segmentation. Among them, the BUSI dataset contains 647 breast tumor images and corresponding semantic segmentation masks. The CVC-ClinicDB dataset consists of 612 colonoscopy sequence images and their corresponding masks. The Kvasir-seg dataset includes 1,000 polyp images and their corresponding masks. In the experiments, all three datasets were divided into training and testing sets with a ratio of 8:2, and all images were resized to 256×256 pixels.

Implementation detail: The entire network architecture of the proposed method is implemented using PyTorch and runs on a Linux system equipped with an NVIDIA GTX A100 GPU. The encoder consists of three convolutional stages and two KAN stages, and the decoder is constructed with a symmetrical architecture to the encoder. During the training process, the Adam optimizer is employed to optimize the entire network with a learning rate of 0.0001 , a training cycle of 350 epochs, and a batch size of 8 . Dice and mIoU were selected as evaluation metrics for semantic segmentation performance. Both Dice and mIoU range from $[0, 1]$, where a higher value indicates better segmentation performance.

3.2. Performance Comparison

Comparison methods:

We selected nine state-of-the-art medical image semantic segmentation models as baseline methods for comparison. These include three CNN-based U-Net models (CU-Net [21], EU-Net [22], and IU-Net [23]), three Transformer-based U-Net models (LKAFormer [1], TransU-Net [24], and MFormer [28]), and three Mamba-based U-Net models (HMAFNet [26], MambaNet [29], and VMamba-Net [30]).

Comparison analysis: As shown in Table 1, the proposed method demonstrates a significant performance advantage over all comparative baseline models across three distinct medical imaging datasets. In comparison to traditional CNN-based architectures, our approach provides more coherent segmentation results, particularly in delineating complex anatomical boundaries. Furthermore, while Transformer-based and recent Mamba-based models show competitive performance on specific datasets, the proposed framework maintains superior consistency and robustness across different imaging modalities. These results indicate that our method effectively handles the inherent variability in medical data, ranging from subtle breast tumor lesions to large-scale polyp structures, ultimately achieving the highest segmentation precision among all evaluated methods.

The performance gain can be attributed to the synergistic integration of KAN and the Mamba-based multi-path fusion mechanism. First, unlike conventional U-Net variants that rely on static activation functions, the U-KAN architecture utilizes learnable activation functions as edges, which significantly enhances the model's ability to approximate complex, non-linear pixel relationships inherent in medical images. Second, the Multi-path Mamba-fused mechanism effectively captures both coarse-grained and fine-grained semantic representations by processing multi-resolution inputs. The selective scanning capability of Mamba allows for dynamic cross-path information aggregation, ensuring that long-range contextual dependencies are preserved while maintaining linear computational complexity. Finally, the multi-stage supervised decoding forces prediction consistency across resolutions, leading to more refined boundary

Table 1. Comparison results of three heterogeneous medical datasets. Bold represents the best results.

Method	BUSI		CVC-ClinicDB		Kvasir-seg	
	mIoU	Dice	mIoU	Dice	mIoU	Dice
CU-Net	0.5785	0.7157	0.8029	0.8695	0.7075	0.8171
EU-Net	0.5692	0.6911	0.7756	0.8224	0.6889	0.8014
IU-Net	0.6073	0.7267	0.8118	0.9007	0.7628	0.8310
TransU-Net	0.5989	0.7172	0.8117	0.9092	0.7701	0.8456
MFormer	0.5528	0.6664	0.7526	0.8043	0.6770	0.7852
LKAFormer	0.5924	0.7373	0.8200	0.8923	0.7557	0.8277
HMAFNet	0.6073	0.7267	0.8118	0.8997	0.7528	0.8210
MambaNet	0.6025	0.7364	0.8194	0.9001	0.7474	0.8247
VMamba-Net	0.5895	0.7257	0.8171	0.8929	0.7457	0.8233
Ours	**0.6284**	**0.7654**	**0.8447**	**0.9124**	**0.7792**	**0.8684**

delineation.

Beyond the quantitative superiority, several noteworthy observations can be made regarding the comparative baselines. While Transformer-based models such as TransU-Net and LKAFormer show strong performance on the Kvasir-seg dataset due to their global receptive fields, they are often hindered by high computational overhead and suboptimal performance on smaller datasets like BUSI. Conversely, while recent Mamba-based models like MambaNet and VMamba-Net improve efficiency, they may struggle with fine-grained local textures. Our method effectively bridges this gap by combining the high-capacity non-linear modeling of KAN with the efficient global context capture of Mamba. The results suggest that the "learnable-activation-on-edge" paradigm of KAN, coupled with multi-fidelity feature fusion, provides a more versatile solution for medical image segmentation than traditional MLP-based or pure attention-based architectures.

3.3. Ablation Analysis

To thoroughly evaluate the individual contributions and effectiveness of each component in the proposed framework, we constructed three ablation variants for comparative analysis. Ours w/o MPE: Replacing the multi-path encoder with a conventional single-path encoding structure. Ours w/o MMF: Substituting the Mamba-based multi-path fusion with standard concatenation-based skip connections. Ours w/o KAN: Reverting the KAN-based encoder-decoder architecture to a standard convolutional backbone. As reported in Table 2, the ablation results across all three heterogeneous datasets consistently indicate that the removal of any single module leads to a discernible degradation in segmentation performance. Specifically, the following observations and conclusions can be drawn: The variant w/o KAN exhibits a substantial drop in both mIoU and Dice metrics. This phenomenon suggests that traditional convolutional layers, which rely on static activation functions, are less efficient at modeling the complex non-linear inter-pixel dependencies found in medical images. The superior performance of our full model underscores the effectiveness of KAN in enhancing the model's non-linear approximation capacity through learnable activation functions. In the w/o MPE variant, the loss of multi-granularity semantic features results in suboptimal segmentation, particularly for lesions with varying scales. This demonstrates that the multi-path encoding mechanism is crucial for capturing both global anatomical contours and local textural nuances simultaneously. The performance decline in the w/o MMF variant validates that traditional skip connections struggle to effectively integrate cross-path information. In contrast, the Mamba-based fusion strategy leverages selective scanning to adaptively aggregate complementary features, thereby significantly enriching the robustness of the feature representations. In conclusion, these ablation findings provide strong empirical evidence for the rationality and necessity of each design choice, confirming that the synergy between U-KAN, MPE, and MMF is the key to achieving state-of-the-art medical image segmentation.

3.4. Visualization Analysis

To further qualitatively evaluate the segmentation performance, we visualize the results of four representative methods across three heterogeneous datasets, as illustrated in Fig. 2. A comparative observation reveals several distinct advantages of our proposed framework: First, in the BUSI dataset, which is characterized by blurred boundaries and low contrast, traditional models often produce over-segmented or fragmented masks. In contrast, our method effectively delineates the tumor contours with high smoothness, demonstrating that the learnable activation functions in the U-KAN blocks possess superior non-linear fitting capabilities for capturing subtle edge gradients. Second, for the CVC-ClinicDB and Kvasir-seg datasets, where polyp

Table 2. Ablation results of three heterogeneous medical datasets. Bold represents the best results.

Variant	BUSI		CVC-ClinicDB		Kvasir-seg	
	mIoU	Dice	mIoU	Dice	mIoU	Dice
Ours w/o MPE	0.6098	0.7386	0.8256	0.9095	0.7503	0.8422
Ours w/o MMF	0.6133	0.7429	0.8337	0.9087	0.7474	0.8398
Ours w/o KAN	0.6002	0.7284	0.8224	0.9014	0.7444	0.8427
Ours	**0.6284**	**0.7654**	**0.8447**	**0.9124**	**0.7792**	**0.8684**

targets exhibit significant scale variations and complex backgrounds, our model shows remarkable structural consistency. While Mamba-based baselines occasionally miss small-scale lesions or generate artifacts in shadowed areas, our Multi-path Mamba-fused architecture successfully preserves the global topological structure while capturing fine-grained textures. This suggests that the dual-path encoding mechanism, combined with Mamba's selective scanning, provides a more robust contextual understanding of diverse pathological shapes. Finally, a closer inspection of the prediction masks indicates that our method achieves the highest degree of overlap with the Ground Truth. The reduction in false positives and the precision in localized details confirm that the deep hierarchical loss effectively guides the model to maintain multi-resolution semantic consistency. Overall, the visualization results align with the quantitative metrics, providing strong evidence that our framework excels in generating high-fidelity segmentation masks for clinical diagnosis support.

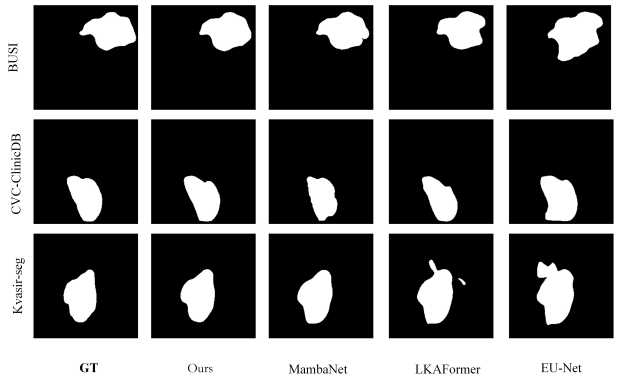

Fig. 2. Visualization results of semantic segmentation of four methods on three heterogeneous datasets.

4. Conclusion

This study presents MMU-KAN, a novel medical image semantic segmentation framework that redefines the traditional U-Net paradigm by synergizing KAN with a multi-path Mamba fusion strategy. By replacing static activation functions with learnable, edge-based nonlinearities, the architecture significantly enhances the model's capacity to approximate complex inter-pixel relationships. The integration of a dual-path encoding mechanism allows for the concurrent extraction of multi-granularity semantic features, while the Mamba-based selective scanning facilitates adaptive cross-path interaction to resolve the semantic gap between global context and local textures.

Acknowledgments

This work was supported in part by the Doctoral Foundation of Fuyang Normal University under Grant 2025KYQD0031, 2024KYQD0119 and 2021KYQD0037, the Teacher Education Collaborative Quality Improvement Plan of Fuyang Normal University under Grant 2025XTKYTD01 and 2025JYXM0002, the Undergraduate Teaching Quality Enhancement Project of Fuyang Normal University under Grant 2025JYXM0011 and the University Natural Science Research Key Project in Anhui Province of China under Grant 2024AH051469 and 2025AHGXZK31290, the Fuyang Normal University - Bengbu Municipal Education Bureau Collaborative Project for Research and Cultivation of Basic Education Results(2024BBXTJY32).

References

[1] S. Yin, L. Wang, T. Chen, H. Huang, J. Gao, J. Zhang, M. Liu, P. Li, and C. Xu, (2026) *"LKAFormer: A lightweight kolmogorov-arnold transformer model for image semantic segmentation"* **ACM Transactions on Intelligent Systems and Technology *17***(3): 1–24. DOI: 10.1145/3759254.

[2] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2024) *"Deep Incomplete Multiview Clustering via Information Bottleneck for Pattern Mining of Data in Extreme-Environment IoT"* **IEEE Internet of Things Journal *11***(16): 26700–26712. DOI: 10.1109/JIOT.2023.3325272.

[3] W. Liu, (2024) *"Channel Reorganization for Few-Shot Segmentation"* **Journal of Artificial Intelligence Research *1***(1): 36–40. DOI: 10.70891/JAIR.2024.100025.

[4] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems** ***35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[5] W. Zhang and J. Wang, (2024) *"English text sentiment analysis network based on CNN and U-Net"* **IFS/ACM Transactions on Machine Learning** ***1***(1): 13–18. DOI: 10.70891/JSE.2024.100009.

[6] G. Gao, C. Chen, K. Xu, K. Liu, and A. Mashhadi, (2024) *"Automatic face detection based on bidirectional recurrent neural network optimized by improved Ebola optimization search algorithm"* **Scientific Reports** ***14***(1): 27798. DOI: 10.1038/s41598-024-79067-x.

[7] X. Xue and H. Zhu, (2022) *"Matching knowledge graphs with compact niching evolutionary algorithm"* **Expert Systems with Applications** ***203***: 117371. DOI: 10.1016/j.eswa.2022.117371.

[8] D. Qian, L. Yu, H. Tang, and J. Zhao, (2020) *"Multiview feature fusion optimization method for image retrieval based on matrix correlation"* **Journal of Electronic Imaging** ***29***(5): 053007–053007. DOI: 10.1117/1.JEI.29.5.053007.

[9] W. Wang, Q. Dong, and Z. Hu, (2023) *"Interactive piecewise planar building reconstruction from a single image based on geometric priors"* **Expert Systems with Applications** ***230***: 120572. DOI: 10.1016/j.eswa.2023.120572.

[10] L. Yu, P. Liu, L. Jiang, and Z. Zhao, (2021) *"Tensor dispersion-based multi-view feature embedding for dimension reduction"* **Journal of Electronic Imaging** ***30***(3): 033019–033019. DOI: 10.1117/1.JEI.30.3.033019.

[11] A. A. Bin-Salem, D. Z. Haider, M. Alzubaidi, Z. U. A. Tariq, and H. Naeem, (2022) *"A scoping review on COVID-19's early detection using deep learning model and computed tomography and ultrasound"* **Traitement du Signal** ***39***(1): 205. DOI: 10.18280/ts.390121.

[12] H. Naeem, A. Alsirhani, M. M. Alshahrani, and A. Alomari, (2022) *"Android Device Malware Classification Framework Using Multistep Image Feature Extraction and Multihead Deep Neural Ensemble."* **Traitement du Signal** ***39***(3): DOI: 10.18280/ts.390326.

[13] D. liang Zhang, Z. Jiang, F. Mohammadzadeh, S. M. H. Azhdari, L. Abualigah, and T. M. Ghazal, (2024) *"FUZ-SMO: A fuzzy slime mould optimizer for mitigating false alarm rates in the classification of underwater datasets using deep convolutional neural networks"* **Heliyon** ***10***(7): DOI: 10.1016/j.heliyon.2024.e28681.

[14] L. Yu, D. Zhang, N. Liu, and W. Zhou, (2021) *"A multi-view fusion method via tensor learning and gradient descent for image features"* **IEEE Access** ***9***: 79389–79399. DOI: 10.1109/ACCESS.2021.3079499.

[15] H. Wang, (2020) *"A Synchronous Transmission Method for Array Signals of Sensor Network under Resonance Technology."* **Traitement du Signal** ***37***(4): DOI: 10.18280/ts.370405.

[16] F. Ullah, M. R. Naeem, H. Naeem, X. Cheng, and M. Alazab, (2022) *"CroLSSim: Cross-language software similarity detector using hybrid approach of LSA-based AST-MDrep features and CNN-LSTM model"* **International Journal of Intelligent Systems** ***37***(9): 5768–5795. DOI: 10.1002/int.22813.

[17] S. Dong, X.-g. Zhang, and W.-g. Zhou, (2020) *"A security localization algorithm based on DV-hop against Sybil attack in wireless sensor networks"* **Journal of Electrical Engineering & Technology** ***15***(2): 919–926. DOI: 10.1007/s42835-020-00361-5.

[18] H. Naeem, X. Cheng, F. Ullah, S. Jabbar, and S. Dong, (2022) *"A deep convolutional neural network stacked ensemble for malware threat classification in internet of things"* **Journal of Circuits, Systems and Computers** ***31***(17): 2250302. DOI: 10.1142/S0218126622503029.

[19] K. Abbas, M. K. Hasan, A. Abbasi, U. A. Mokhtar, A. Khan, S. N. H. S. Abdullah, S. Dong, S. Islam, D. Alboaneen, and F. R. A. Ahmed, (2023) *"Predicting the future popularity of academic publications using deep learning by considering it as temporal citation networks"* **IEEE Access** ***11***: 83052–83068. DOI: 10.1109/ACCESS.2023.3290906.

[20] Z. Wang, T. Tao, Y. Ge, Z. Chen, T. Chen, Z. Ye, and Y. Lei, (2026) *"Weak-mamba-unet: Visual mamba makes cnn and vit work better for scribble-based medical image segmentation"* **IEEE Transactions on Biomedical Engineering**: DOI: 10.1109/TBME.2026.3668882.

[21] X. Shu, J. Wang, A. Zhang, J. Shi, and X.-J. Wu, (2024) *"CSCA U-Net: A channel and space compound attention CNN for medical image segmentation"* **Artificial Intelligence in Medicine** ***150***: 102800. DOI: 10.1016/j.artmed.2024.102800.

[22] M. M. Rahman, M. Munir, and R. Marculescu. "Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, 11769–11779. DOI: 10.1109/CVPR52733.2024.01118.

[23] D. Dai, C. Dong, Q. Yan, Y. Sun, C. Zhang, Z. Li, and S. Xu, (2024) *"I2u-net: A dual-path u-net with rich information interaction for medical image segmentation"* **Medical Image Analysis *97***: 103241. DOI: 10.1016/j.media.2024.103241.

[24] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, (2022) *"Ds-transunet: Dual swin transformer u-net for medical image segmentation"* **IEEE Transactions on Instrumentation and Measurement *71***: 1–15. DOI: 10.1109/TIM.2022.3178991.

[25] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images". In: *International MICCAI brainlesion workshop*. 2021, 272–284. URL: https://arxiv.org/abs/2201.01266.

[26] R. Wu, Y. Liu, P. Liang, and Q. Chang, (2025) *"H-vmunet: High-order vision mamba unet for medical image segmentation"* **Neurocomputing *624***: 129447. DOI: 10.1016/j.neucom.2025.129447.

[27] P. Liang, L. Shi, B. Pu, R. Wu, J. Chen, L. Zhou, L. Xu, Z. Chen, Q. Chang, and Y. Li, (2025) *"Mambasam: A visual mamba-adapted sam framework for medical image segmentation"* **IEEE journal of biomedical and health informatics**: DOI: 10.1109/JBHI.2025.3544548.

[28] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, (2022) *"Missformer: An effective transformer for 2d medical image segmentation"* **IEEE transactions on medical imaging *42***(5): 1484–1494. DOI: 10.1109/TMI.2022.3230943.

[29] X. Kui, S. Jiang, Q. Li, Y. Peng, Z. Hu, and B. Zou, (2025) *"Gl-MambaNet: A global-local hybrid Mamba network for medical image segmentation"* **Neurocomputing *626***: 129580. DOI: 10.1016/j.neucom.2025.129580.

[30] T. D. Q. Dang, H. H. Nguyen, and A. Tiulpin. "LoG-VMamba: local-global vision mamba for medical image segmentation". In: *Proceedings of the Asian Conference on Computer Vision*. 2024, 548–565. DOI: 10.1007/978-981-96-0901-7_14.