U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation

Chenxin Li*, Xinyu Liu*, Wuyang Li*, Cheng Wang*,
Hengyu Liu, Yifan Liu, Zhen Chen, Yixuan Yuan

The Chinese University of Hong Kong
yxyuan@ee.cuhk.edu.hk

Abstract

U-Net has become a cornerstone in various visual applications such as image segmentation and diffusion probability models. While numerous innovative designs and improvements have been introduced by incorporating transformers or MLPs, the networks are still limited to linearly modeling patterns as well as the deficient interpretability. To address these challenges, our intuition is inspired by the impressive results of the Kolmogorov-Arnold Networks (KANs) in terms of accuracy and interpretability, which reshape the neural network learning via the stack of non-linear learnable activation functions derived from the Kolmogorov-Anold representation theorem. Specifically, in this paper, we explore the untapped potential of KANs in improving backbones for vision tasks. We investigate, modify and re-design the established U-Net pipeline by integrating the dedicated KAN layers on the tokenized intermediate representation, termed U-KAN. Rigorous medical image segmentation benchmarks verify the superiority of U-KAN by higher accuracy even with less computation cost. We further delved into the potential of U-KAN as an alternative U-Net noise predictor in diffusion models, demonstrating its applicability in generating task-oriented model architectures.

Website — https://yes-u-kan.github.io/
Code — https://github.com/CUHK-AIM-Group/U-KAN
Extended version — https://arxiv.org/pdf/2406.02918

Introduction

In the past decade, many works have developed efficient segmentation methods for medical imaging due to the need for computer-aided diagnosis and image-guided surgical systems (Sun et al. 2022; Li et al. 2022b, 2021b; Liu et al. 2024a). U-Net (Ronneberger, Fischer, and Brox 2015) is a significant work, showing the effectiveness of encoder-decoder CNNs with skip connections for medical image segmentation (Wang et al. 2022; Li et al. 2021a; Ding et al. 2022), and also achieving good results in image translation tasks (Torbunov et al. 2023). Recently, diffusion models have used U-Net as a backbone and trained it to predict the noise to be removed in each denoising step (Ho, Jain, and Abbeel 2020).

Since U-Net's debut (Ronneberger, Fischer, and Brox 2015), med imaging advanced via U-Net++ (Zhou et al. 2018), 3D U-Net (Çiçek et al. 2016), V-Net (Milletari, Navab, and Ahmadi 2016), Y-Net (Mehta et al. 2018). U-NeXt (Valanarasu and Patel 2022) & Rolling U-Net (Liu et al. 2024c) build on, blend conv & MLP for res-limited seg. Simultaneously, Trans-UNet (Chen et al. 2021), MedT (Valanarasu et al. 2021), UNETR (Hatamizadeh et al. 2022) arose to enhance U-Net's global context. But they overfit on small data, needing alternatives. To address, SSMs (Gu and Dao 2023) showed potential in long-seq modeling. U-Mamba (Ma, Li, and Wang 2024) & SegMamba (Xing et al. 2024), based on nn-UNet (Isensee et al. 2021) & Swin UNETR (Hatamizadeh et al. 2021), got good results in vision tasks, maybe beating transformer limits.

Existing U-shape variations in medical image segmentation, despite being advanced, encounter fundamental challenges from sub-optimal kernel design and lack of explainability. Conventional kernels like convolution, Transformers, and MLPs are restricted to linearly modeling patterns across channels and struggle to capture the complex nonlinear relationships common in medical imaging. Such nonlinear patterns are vital as channels can signify different clinical, anatomical, or pathological aspects. Moreover, these models usually depend on empirical network search and heuristic design, overlooking the explainability of black-box U-shape models. This lack of explainability risks clinical decision-making and impedes the development of reliable diagnostic systems. Recent progress in Kolmogorov-Arnold Networks (KANs) has demonstrated potential for enhancing network interpretability (Yu et al. 2024). Utilizing KAN's architecture can effectively bridge the gap between a network's physical attributes and empirical performance, thus overcoming the limitations of current U-shape models.

In this paper, we present U-KAN, a universal framework integrating Kolmogorov-Arnold Networks (KAN) into the UNet backbone via a convolutional KAN mixed style. U-KAN keeps U-Net's benchmark setup, using a multilayered encoder-decoder with skip connections and adding a tokenized KAN block at higher-level representations near the bottleneck to project features into tokens and apply the KAN operator for pattern extraction. U-KAN utilizes KAN's nonlinear modeling and interpretability, standing out in the U-Net architecture. Empirical evaluations on medical segmentation

*These authors contributed equally.

benchmarks show its superior performance, surpassing existing U-Net backbones with higher accuracy and lower computation cost. We also explore its potential as a U-Net noise predictor in diffusion models. In essence, U-KAN is a significant step in integrating math theory-inspired operators into efficient visual pipelines, with broad application prospects in visual tasks. Our contributions are:

- We present the first effort to incoporate the advantage of emerging KAN, improving the established U-Net pipeline to be more accurate, efficient, and interpretable.
- We propose a tokenized KAN block to effectively steer the KAN operators to be compatible with the existing convolution-based designs.
- We empirically validate U-KAN on a wide range of medical segmentation benchmarks, achieving impressive accuracy and efficiency.
- The application of U-KAN to existing diffusion models as an improved noise predictor demonstrates its potential in backbone generative tasks and broader vision settings.

Related Work

U-Net for Medical Image Segmentation. Medical image segmentation has seen significant advancements through deep learning methods (Ronneberger, Fischer, and Brox 2015; Li et al. 2024a). U-Net (Ronneberger, Fischer, and Brox 2015) popularized the encoder-decoder architecture, while subsequent models like CE-Net (Gu et al. 2019) and Unet++ (Zhou et al. 2018) enhanced contextual information and multi-scale feature fusion. Transformer-based models, including Vision Transformer (Dosovitskiy et al. 2021), Medical Transformer (Valanarasu et al. 2021), and TransUNet (Chen et al. 2021), have also shown promise in this field. Attention mechanisms (Schlemper et al. 2019) and multi-scale feature fusion (Huang et al. 2020) are widely employed, while 3D models (Andermatt, Pezold, and Cattin 2016; Kamnitsas et al. 2017) have yielded commendable results. Recently, Mamba (Gu and Dao 2023) achieved a breakthrough with linear-time inference and efficient training. Its visual applications, such as Vision Mamba (Liu et al. 2024b) and VMamba (Zhu et al. 2024), have shown superior performance in capturing global visual context. U-Mamba (Ma, Li, and Wang 2024) and related works (Xing et al. 2024; Ruan and Xiang 2024) have demonstrated excellence in medical image segmentation. The emergence of Kolmogorov–Arnold Network (KAN) (Liu et al. 2024d) as a promising alternative to MLP, with its precision, efficiency, and interpretability, opens new avenues for exploration in vision backbones.

U-Net Diffusion for Image Generation. Diffusion Probability Models have emerged as a focal point in computer vision research (Ho, Jain, and Abbeel 2020; Rombach et al. 2022), introducing a novel generative paradigm distinct from VAEs (Kingma and Welling 2013), GANs (Goodfellow et al. 2014; Brock, Donahue, and Simonyan 2018; Zhang et al. 2021), and vector quantization methods (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021). These models use a fixed Markov chain to map latent space, capturing intricate dataset structures. Diffusion models have shown impressive generative prowess in various applications, including image synthesis (Ho, Jain, and Abbeel 2020), editing (Avrahami, Lischinski, and Fried 2022; Choi et al. 2021), image-to-image translation (Saharia et al. 2022; Li et al. 2024c), and video generation (Li et al. 2024b). The process involves a diffusion phase, where Gaussian noise is gradually added to input data, and a denoising phase, where the original data is recovered through learned inverse diffusion operations. Convolutional U-Nets (Ronneberger, Fischer, and Brox 2015) are typically used as the backbone architecture for predicting noise removal at each denoising step. Recent research has shifted from utilizing pre-trained diffusion U-Nets to exploring their intrinsic features and structural properties. Free-U reassesses U-Net's skip connections and backbone feature maps, RINs (Jabri, Fleet, and Chen 2022) introduced an efficient attention-based architecture, and DiT (Peebles and Xie 2023) proposed a scalable pure transformer-diffusion combination. This paper exhibits the potential of combining U-Net and KAN in generation, extending the limits of generation backbone architecture.

Kolmogorov–Arnold Networks (KANs). The Kolmogorov-Arnold theorem (Kolmogorov 1957) posits that any continuous function can be expressed as a composition of continuous unary functions of finite variables, providing a theoretical basis for universal neural network models. Hornik et al. (Hornik, Stinchcombe, and White 1989) further demonstrated the universal approximation capabilities of feed-forward neural networks, laying the groundwork for deep learning. Inspired by this theorem, Kolmogorov-Arnold Networks (KANs) (Huang, Zhao, and Song 2014) were proposed. KANs comprise concatenated Kolmogorov-Arnold layers with learnable one-dimensional activation functions, effectively approximating high-dimensional complex functions across various applications. KANs are notable for their strong theoretical interpretability and explainability. Huang et al.(Huang, Zhao, and Xing 2017) analyzed KANs' optimization characteristics and convergence, validating their approximation capacity and generalization performance. Liang et al.(Liang, Zhao, and Huang 2018) introduced a deep KAN model for tasks like image classification, while Xing et al. (Xing, Zhao, and Huang 2018) applied KANs to time series prediction and control problems. Despite recent advancements, the integration of KANs into general-purpose vision networks lacks practical implementations. This paper presents a universal network architecture with KAN and validates it across diverse segmentation and generative tasks.

Method

Overview Fig. 1 illustrates the proposed U-KAN, featuring a two-phase encoder-decoder architecture: a Convolution Phase and a Tokenized Kolmogorov–Arnold Network (Tok-KAN) Phase. The encoder consists of three convolutional blocks followed by two tokenized MLP blocks, while the decoder comprises two tokenized KAN blocks and three convolutional blocks. Each encoder block halves the feature resolution, with decoder blocks reversing this process. Skip connections link the encoder and decoder. Channel counts in the Convolution and Tok-KAN Phases are determined by hyperparameters C_1 to C_3 and D_1 to D_3, respectively.

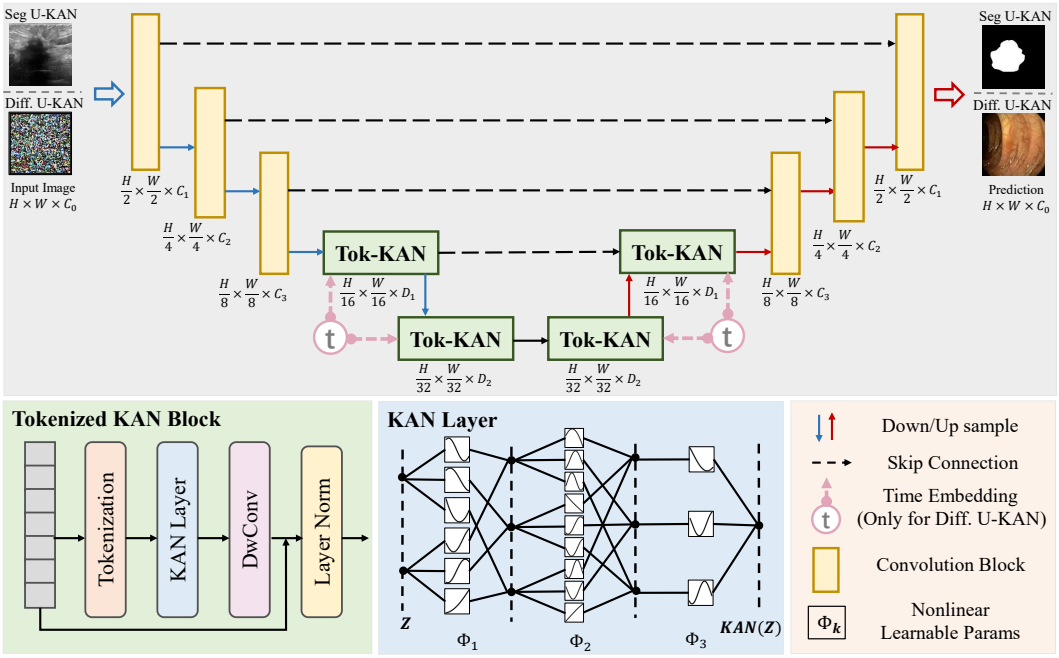

Figure 1: Overview of U-KAN pipeline. After feature extraction by several convolution blocks in Convolution Phrase, the intermediate maps are tokenized and processed by stacked Tok-KAN blocks in Tokenized KAN Phrase. The time embedding is only injected into the KAN blocks when applied for Diffusion U-KAN.

KAN as Efficient Embedder

This research aims to incorporate Kolmogorov–Arnold Networks (KANs) into the U-Net framework. The basis of this approach is the proven high efficiency and interpretability of KANs as outlined in (Liu et al. 2024d). A Multi-Layer Perceptron (MLP) comprising K layers can be described as an interplay of transformation matrices W and activation functions σ. This can be mathematically expressed as:

$$\mathrm{MLP}(\mathbf{Z}) = (W_{K-1} \circ \sigma \circ W_{K-2} \circ \sigma \circ \cdots \circ W_1 \circ \sigma \circ W_0)\,\mathbf{Z}, \tag{1}$$

where it strives to mimic complex functional mappings through a sequence of nonlinear transformations over multiple layers. Despite its potential, the inherent obscurity within this structure significantly hampers the model's interpretability, thus posing considerable challenges to intuitively understanding the underlying decision-making mechanisms. In an effort to mitigate the issues of low parameter efficiency and limited interpretability inherent in MLPs, Liu *et al.* (Liu et al. 2024d) proposed the Kolmogorov-Arnold Network (KAN), drawing inspiration from the Kolmogorov-Arnold representation theorem (Kolmogorov 1961).

Similar to an MLP, a K-layer KAN can be characterized as a nesting of multiple KAN layers:

$$\mathrm{KAN}(\mathbf{Z}) = (\mathbf{\Phi}_{K-1} \circ \mathbf{\Phi}_{K-2} \circ \cdots \circ \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0)\,\mathbf{Z}, \tag{2}$$

where $\mathbf{\Phi}_i$ signifies the i-th layer of the entire KAN network. Each KAN layer, with n_{in} -dimensional input and n_{out} -dimensional output, $\mathbf{\Phi}$ comprises $n_{in} \times n_{out}$ learnable activation functions ϕ:

$$\mathbf{\Phi} = \{\phi_{q,p}\}, \quad p = 1, 2, \cdots, n_{\text{in}}, \quad q = 1, 2 \cdots, n_{\text{out}}, \tag{3}$$

The computation result of the KAN network from layer k to layer $k+1$ can be expressed in matrix form $\mathbf{Z}_{k+1} = \mathbf{\Phi}_k \mathbf{Z}_k$, where:

$$\mathbf{\Phi}_k = \begin{pmatrix} \phi_{k,1,1}(\cdot) & \cdots & \phi_{k,1,n_k}(\cdot) \\ \phi_{k,2,1}(\cdot) & \cdots & \phi_{k,2,n_k}(\cdot) \\ \vdots & \vdots & \vdots \\ \phi_{k,n_{k+1},1}(\cdot) & \cdots & \phi_{k,n_{k+1},n_k}(\cdot) \end{pmatrix} \tag{4}$$

In conclusion, KANs innovate by using learnable and parametrized activation functions as both edges and weights, eliminating linear weight matrices. This design achieves comparable or superior performance with smaller models, while enhancing interpretability. Consequently, KANs offer a versatile solution for various applications.

U-KAN Architecture

Convolution Phrase Each convolution block is constructed of the components as follows: a convolutional layer (Conv), a batch normalization layer (BN), and a ReLU activation function. We apply a kernel size of 3x3, a stride length of 1, and a padding quantity of 1. The convolution blocks within the encoder integrate a max-pooling layer with a size of 2x2. Formally, given an image $\boldsymbol{X}_0 = \boldsymbol{I} \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, the output of each convolution block can be elaborated as:

$$\boldsymbol{X}_\ell = \mathrm{Pool}\big(\mathrm{Conv}(\boldsymbol{X}_{\ell-1})\big), \tag{5}$$

where $\boldsymbol{X}_\ell \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell}$ represents the output feature maps at ℓ-th layer. Given the configuration that there are L blocks in the Convolution Phrase, the final output is derived as $\boldsymbol{X}_L$.

Tokenized KAN Phrase In the tokenized KAN block, we first perform tokenization (Dosovitskiy et al. 2021) by reshaping the output feature of convolution phrase $\boldsymbol{X}_L$ into a sequence of flattened 2D patches $\{\boldsymbol{X}_L^i \in \mathbb{R}^{P^2 \cdot C_L} | i = 1, .., N\}$, where each patch is of size $P \times P$ and $N = \frac{H_L \times W_L}{P^2}$ is the number of feature patches. We then map the vectorized patches into a latent D-dimensional embedding space using a trainable linear projection $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C_L) \times D}$, as:

$$\boldsymbol{Z}_0 = [\boldsymbol{X}_L^1\mathbf{E};\ \boldsymbol{X}_L^2\mathbf{E}; \cdots ;\ \boldsymbol{X}_L^N\mathbf{E}], \tag{6}$$

The linear projection $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C_L) \times D}$ uses a 3x3 convolution layer. As shown in (Xie et al. 2021), this effectively encodes positional information and outperforms standard positional encoding techniques. Unlike traditional methods that may require interpolation for different resolutions, this design keeps robust performance across varying input sizes.

Given the obtained tokens, we pass them into a series of KAN layers ($N = 3$). Followed each KAN layers, the features are passed through a efficient depth-wise convolutional layer (DwConv) and a bacth normalization layer (BN) and a ReLU activation. We use a residual connection here and add the original tokens as residuals. We then apply a layer normalization (LN) and pass the output features to the next block. Formally, the output of k-th Tokenized KAN block is:

$$\boldsymbol{Z}_k = \mathrm{LN}(\boldsymbol{Z}_{k-1} + \mathrm{DwConv}(\mathrm{KAN}(\boldsymbol{Z}_{k-1}))), \tag{7}$$

where $\boldsymbol{Z}_k \in \mathbb{R}^{H_k \times W_k \times D_k}$ is the output feature maps at k-th layer. Given the setup that there are K blocks in the Tokenized KAN Phrase, the final output is derived as $\boldsymbol{Z}_K$. In our implementation, we set $L = 3$ and $K = 2$.

U-KAN Decoder We construct U-KAN following the widely adopted U-shaped architecture with dense skip connections. U-Net and its variations have demonstrated remarkable efficiency in medical image segmentation tasks (Li et al. 2022a; Xu et al. 2024), leveraging skip connections to recover low-level details while using an encoder-decoder structure for high-level information extraction. Given skip-connected feature $\boldsymbol{Z}_k$ from layer-k in KAN Phrase and feature $\boldsymbol{Z}'_{k+1}$ from the last up-sample block, the output feature $\boldsymbol{Z}'_k$ of k-th up-sample block is:

$$\boldsymbol{Z}'_k = \mathrm{Cat}\left(\boldsymbol{Z}'_{k+1}, (\boldsymbol{Z}_k)\right), \tag{8}$$

where $\mathrm{Cat}(\cdot)$ denotes the feature concatenation operation. Likewise, given skip-connected feature $\boldsymbol{X}_\ell$ from layer-ℓ in Convolution Phrase and feature $\boldsymbol{X}'_{\ell+1}$ from the last up-sample block, the output $\boldsymbol{X}'_\ell$ of ℓ-th up-sample block is:

$$\boldsymbol{X}'_\ell = \mathrm{Cat}\left(\boldsymbol{X}'_{\ell+1}, (\boldsymbol{X}_\ell)\right), \tag{9}$$

In the context of semantic segmentation tasks, the final segmentation map can be derived from the output feature maps $X'_0 \in \mathbb{R}^{H_0 \times W_0 \times C_Y}$ at layer-0, where C_Y is the number of semantic categories and $\boldsymbol{Y}$ denotes the ground-truth segmentation and. As a result, the segmentation loss can be:

$$\mathcal{L}_{\text{Seg}} = CE\big(\boldsymbol{Y}, \text{U-KAN}(\boldsymbol{I})\big). \tag{10}$$

where CE denotes the pixel-wise cross-entropy loss.

Extending U-KAN to Diffusion Models

Building on our discussion of U-KAN for segmentation mask generation, we now introduce Diffusion U-KAN, an extension that harnesses the generative capacity of KANs. Inspired by Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain, and Abbeel 2020), Diffusion U-KAN generates images from random Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ through gradual noise removal. This process is achieved by predicting noise given a noisy input: $\epsilon_t = \text{U-KAN}(\boldsymbol{I}_t, t)$, where $\boldsymbol{I}_t$ is image I corrupted by Gaussian noise ϵ_t, $t = [1, T], T = 1000$ is the time-step controlling noise intensity, and $\boldsymbol{I}_T \sim \mathcal{N}(0, 1)$.

To this end, we conduct two modifications based on the Segmentation U-KAN to lift it to the diffusion version. First, different from only propagating features among different hidden layers, we inject learnable time embedding into each block to enable the network time-aware (see the dashed-line "Time Embedding" in Fig 1) and remove the DwConv and residual connections, thereby changing Eq. 7 into the following format for the goal of generative tasks:

$$\boldsymbol{Z}_k = \mathrm{LN}(\mathrm{KAN}(\boldsymbol{Z}_{k-1})) + \mathcal{F}(\mathrm{TE}(t)), \tag{11}$$

where $\mathcal{F}$ represents the linear projection, and $\mathrm{TE}(t)$ denotes the time embedding for the given time step t (Ho, Jain, and Abbeel 2020). The second modification alters the prediction objective to enable diffusion-based image generation. Rather than predicting segmented masks from images, Diffusion U-KAN aims to predict noise ϵ_t given a noise-corrupted image I_t and a random time-step $t = \text{Uniform}(1, T)$. This prediction is optimized using MSE loss as follows:

$$\mathcal{L}_{\text{Diff}} = ||\epsilon_t - \text{U-KAN}(\boldsymbol{I}_t, t)||_2. \tag{12}$$

After optimization using the described loss function, we generate images using the DDPM sampling algorithm (Ho, Jain, and Abbeel 2020), which employs the trained Diffusion U-KAN for denoising.

Experiments

Datasets

We evaluated our method on three diverse datasets, each with unique characteristics, sizes, and image resolutions. These widely-used datasets for image segmentation and generation tasks provide a comprehensive testing ground for assessing the efficacy and adaptability of our approach.

BUSI. The dataset (Al-Dhabyani et al. 2020) comprises ultrasound images of normal, benign, and malignant breast cancer cases with corresponding segmentation maps. We used 647 images of benign and malignant breast tumors, resized to 256×256. This comprehensive collection aids in tumor detection and differentiation, providing valuable insights for medical professionals and researchers.

GlaS. The dataset (Valanarasu et al. 2021) consists of 612 Standard Definition (SD) frames (384×288) from 31 sequences, collected from 23 patients at the Hospital Clinic in Barcelona, Spain. Sequences were recorded using Olympus Q160AL and Q165L devices with an Extra II video processor. Following common practice (Liu et al. 2024c), we used 165 images, resized to 512×512.

Methods	BUSI (Al-Dhabyani et al. 2020)		GlaS (Valanarasu et al. 2021)		CVC (Bernal et al. 2015)	
	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑
U-Net (Ronneberger, Fischer, and Brox 2015)	57.22±4.74	71.91±3.54	86.66±0.91	92.79±0.56	83.79±0.77	91.06±0.47
Att-Unet (Oktay et al. 2018)	55.18±3.61	70.22±2.88	86.84±1.19	92.89±0.65	84.52±0.51	91.46±0.25
U-Net++ (Zhou et al. 2018)	57.41±4.77	72.11±3.90	87.07±0.76	92.96±0.44	84.61±1.47	91.53±0.88
U-NeXt (Valanarasu and Patel 2022)	59.06±1.03	73.08±1.32	84.51±0.37	91.55±0.23	74.83±0.24	85.36±0.17
Rolling-UNet (Liu et al. 2024c)	61.00±0.64	74.67±1.24	86.42±0.96	92.63±0.62	82.87±1.42	90.48±0.83
U-Mamba (Ma, Li, and Wang 2024)	61.81±3.24	75.55±3.01	87.01±0.39	93.02±0.24	84.79±0.58	91.63±0.39
Seg. U-KAN (Ours)	**63.38±2.83**	**76.40±2.90**	**87.64±0.32**	**93.37±0.16**	**85.05±0.53**	**91.88±0.29**

Table 1: Comparison with state-of-the-art segmentation models on three heterogeneous medical scenarios. The average results with standard deviation over three random runs are reported.

Methods	Average Seg.		Efficiency	
	IoU↑	F1↑	Gflops	Params (M)
U-Net (Ronneberger, Fischer, and Brox 2015)	75.89±2.14	85.25±1.52	524.2	34.53
Att-Unet (Oktay et al. 2018)	75.51±1.77	84.85±1.26	533.1	34.9
U-Net++ (Zhou et al. 2018)	76.36±2.33	85.53±1.74	1109	36.6
U-NeXt (Valanarasu and Patel 2022)	72.80±0.54	83.33±0.57	4.58	1.47
Rolling-UNet (Liu et al. 2024c)	76.76±1.01	85.92±0.89	16.82	1.78
U-Mamba (Ma, Li, and Wang 2024)	77.87±1.47	86.73±1.25	2087	86.3
Seg. U-KAN (Ours)	**78.69±1.27**	**87.22±1.15**	14.02	6.35

Table 2: Overall comparison with state-of-the-art segmentation models w.r.t. efficiency and segmentation metrics.

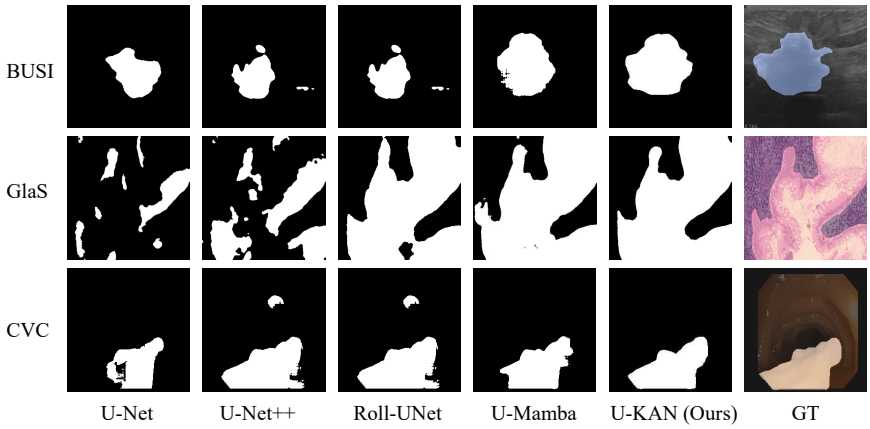

Figure 2: Visualized segmentation results of U-KAN against other state-of-the-arts over three heterogeneous medical scenarios.

CVC-ClinicDB. The dataset (Bernal et al. 2015) is a public resource for polyp diagnosis in colonoscopy videos. It contains 612 images (384×288) from 31 colonoscopy sequences, offering diverse polyp instances for detection algorithm development and evaluation. All images were resized to 256×256 for consistency with other datasets in our study.

Implementation Details

Segmentation U-KAN. We implemented U-KAN using PyTorch on an NVIDIA RTX 4090 GPU. For all datasets, we used a batch size of 8 and an initial learning rate of 1e-4. We employed the Adam optimizer with a cosine annealing learning rate scheduler (minimum 1e-5). The loss function combined binary cross-entropy (BCE) and dice loss. Datasets were randomly split into 80% training and 20% validation subsets. Results are reported over three random runs. We applied vanilla data augmentations (random rotation and flipping) and trained for 400 epochs. Evaluation metrics include IoU, F1 Score, GFLOPs, and parameter count.

Diffusion U-KAN. For unconditional generation, images were cropped and resized to 64×64. All methods were benchmarked using the same training settings: 1e-4 learning rate,

1000 epochs, Adam optimizer, and cosine annealing learning rate scheduler. We generated 2048 image samples from random Gaussian noise for each method. Evaluation metrics include Fréchet Inception Distance (FID)(Parmar, Zhang, and Zhu 2021) and Inception Score (IS)(Saito, Matsumoto, and Saito 2017), providing insights into the diversity and quality of generated images.

Performance Comparison on Image Segmentation

Tab. 1 presents the results of our proposed U-KAN against various comparison methods across all benchmarking datasets. We evaluated U-KAN against recent popular frameworks for medical image segmentation, including convolutional baselines like U-Net (Ronneberger, Fischer, and Brox 2015) and U-Net++ (Zhou et al. 2018), attention-based models such as Att-UNet (Oktay et al. 2018), and the state-of-the-art efficient transformer variant, U-Mamba (Ma, Li, and Wang 2024). Additionally, we compared against advanced MLP-based segmentation networks like U-NeXt (Valanarasu and Patel 2022) and Rolling-UNet (Liu et al. 2024c). Performance was evaluated using standard metrics: Intersection over Union (IoU) and F1 scores. The results demonstrate that U-KAN outperforms all other methods across all datasets. Furthermore, Tab. 2 highlights the efficiency of our method as a baseline. We report the parameter count (M), GFLOPs, and segmentation accuracy across various datasets. The results indicate that our method not only surpasses most segmentation methods in accuracy but also demonstrates significant advantages or comparable levels in efficiency, with the exception of U-Next. Overall, U-KAN exhibits superior performance in balancing segmentation accuracy and efficiency.

Fig. 2 further presents qualitative comparisons across all datasets. Pure CNN-based approaches like U-Net and U-Net++ show tendencies for over- or under-segmentation, indicating limitations in global context encoding and semantic discrimination. Our U-KAN yields fewer false positives and exhibits finer boundary details compared to Transformer-based and MLP-based architectures. These observations highlight U-KAN's superior capability for refined segmentation while preserving intricate shape information, corroborating the advantages of incorporating the KAN layer and supporting our quantitative findings.

Performance Comparison on Image Generation

We evaluated U-KAN's potential as a backbone for generative tasks, comparing it with various diffusion variants based on conventional U-Nets. Tab. 3 presents FID (Parmar, Zhang, and Zhu 2021) and IS (Saito, Matsumoto, and Saito 2017) metrics across three datasets. Lower FID indicates better resemblance to real images, while higher IS suggests better image quality. Our results demonstrate U-KAN's superior generative performance compared to other state-of-the-art models. This indicates that U-KAN's architecture is particularly well-suited for generative tasks, offering an effective and efficient approach to producing high-quality images. These findings highlight U-KAN's versatility in both segmentation and image generation tasks.

Fig. 3 showcases visualizations of our generated results. Our method produces realistic and diverse content across multiple datasets, demonstrating its versatility and effectiveness in high-quality image generation. These visual results reinforce U-KAN's significant advantage in generative tasks, positioning it as a promising candidate for future research and development in this field.

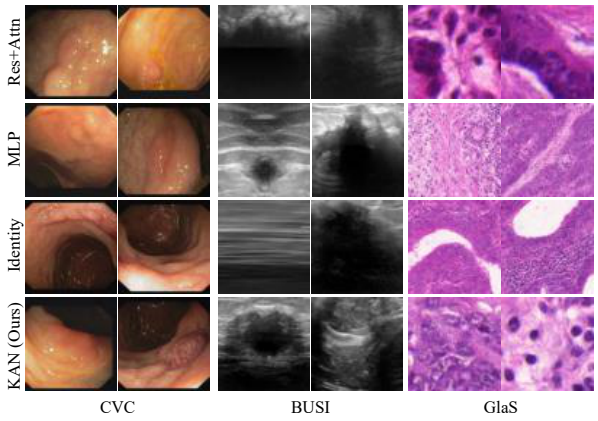

Figure 3: Generated images by proposed Diffusion U-KAN in three heterogeneous medical scenarios.

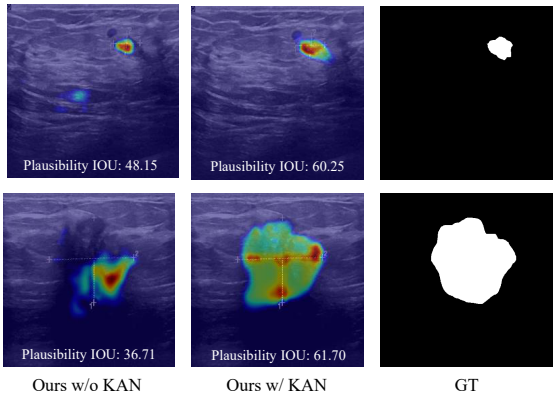

Figure 4: Explainability of U-KAN with channel activation.

Ablation Studies

To thoroughly evaluate the proposed U-KAN framework and validate its performance under different settings, we conducted a series of ablation studies as follows.

The Number of KAN Layer. This ablation study assessed the impact of varying the number of KAN Layers in U-KAN. We tested configurations with one to five KAN Layers, as shown in Tab. 4. The default setup of U-KAN is denoted. The results indicate that the configuration with three KAN Layers yielded the best performance. These findings demonstrate that strategically integrating an optimal number of KAN Layers within U-KAN effectively captures intricate segmentation details, validating the benefit of incorporating these highly efficient embeddings.

Impact on Using KAN Layer v.s. MLP. To assess the impact

Methods	Middle Blocks	BUSI (Al-Dhabyani et al. 2020)		GlaS (Valanarasu et al. 2021)		CVC (Bernal et al. 2015)	
		FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
Diffusion U-Net	ResBlock+Attn	116.52	2.54	42.65	2.45	49.30	2.65
	Identity	124.46	2.71	42.63	2.41	50.42	2.49
	MLP	104.95	2.59	44.21	2.43	51.16	2.69
Diffusion U-KAN (Ours)	KANBlock	**101.93**	**2.76**	**41.55**	**2.46**	**46.34**	**2.75**

Table 3: Comparison with standard U-Net based diffusion models on three heterogeneous medical scenarios. Results by different variants of Diffusion U-Net is provided for comprehensive evaluation.

#KAN	IoU↑	F1↑	Gflops
1 Layer	64.20	77.81	13.97
2 Layer	64.56	78.01	14.00
3 Layer	65.26	78.75	14.02
4 Layer	64.72	78.35	14.05
5 Layer	64.86	78.42	14.07

Table 4: Ablation studies on number of used KAN layers.

KAN vs. MLP	IoU↑	F1↑	Gflops
KAN×3	65.26	78.75	14.02
MLP+KAN+KAN	64.12	77.86	14.29
KAN+MLP+KAN	63.82	77.58	14.29
KAN+KAN+MLP	64.30	77.95	14.29
MLP×3	63.49	77.07	14.84

Table 5: Ablation studies on using KAN layers against MLPs.

Model Scale	C_1	C_2	C_3	IoU↑	F1↑	Gflops
U-KAN-S	64	96	128	64.62	78.28	3.740
U-KAN	128	160	256	65.26	78.75	14.02
U-KAN-L	256	320	512	66.01	79.09	55.11

Table 6: Ablation studies on model scaling by using different channel settings in U-KAN.

of KAN layers, we conducted ablation experiments (Tab. 5) by replacing KAN layers with traditional multilayer perceptrons (MLPs) in our model. We modified models by substituting one or more KAN layers with MLPs, then retrained these variants using identical datasets and parameters. Performance was evaluated across various tasks. Results showed a notable performance decline when KAN layers were replaced, especially in complex tasks requiring robust feature extraction. This decline was particularly evident in tasks demanding high representational capacity. These findings highlight the crucial role of KAN layers in enhancing the model's expressive capabilities and overall performance.

Model Scaling. We conducted an ablation study on various sizes of U-KAN, examining Small and Large configurations alongside our default model. The primary distinction lies in their channel settings (C1-C3) from the first to third KAN layer, as detailed in Tab. 6. We observed that larger models correlate with enhanced performance, aligning with the scaling law characteristics of KAN-integrated models. To balance performance and computational costs, we opted for the default base model in our experiments.

Explainability Analysis. We further explore the interpretability of KAN layers by analyzing activated patterns, as depicted in Fig. 4. When utilizing MLP layers (1th column), the model struggles to identify appropriate activation regions essential with an unsatisfactory Plausibility IoU, which is a metric provided in (Cambrin et al. 2024) that calculates IoU between thresholded activation maps and GT masks. In contrast, with integrating KAN layer (2nd column), there is a marked improvement in the ability to precisely locate the region of interest and activate the boundaries that align closely with the ground truth (3rd column). This underscores the pivotal role of KAN layers in enhancing the explainable decision-making of deep models, especially for mask prediction, which is also aligned with the observation in KAN (Liu et al. 2024d).

Conclusion

This paper introduces U-KAN, demonstrating the potential of Kolmogorov-Arnold Networks (KANs) in enhancing U-Net-like backbones for visual applications. By integrating KAN layers into U-Net, we create a robust network offering improved accuracy, efficiency, and interpretability for vision tasks. We empirically evaluate our method on several medical image segmentation tasks. Additionally, U-KAN shows promise as an alternative to U-Net for noise prediction in diffusion models. These findings highlight the importance of exploring non-traditional network structures like KANs to advance a broader range of vision applications.

Acknowledgment. This work was supported by the Hong Kong Research Grants Council (RGC) General Research Fund under Grant 14220622, and 14204321

References

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in brief*, 28: 104863.

Andermatt, S.; Pezold, S.; and Cattin, P. C. 2016. Multi-dimensional Gated Recurrent Units for the Segmentation of Biomedical 3D-Data. In *DLDAs*, 142–151. Springer, Cham.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for

accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG*, 43: 99–111.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv:1809.11096*.

Cambrin, D. R.; Poeta, E.; Pastor, E.; Cerquitelli, T.; Baralis, E.; and Garza, P. 2024. KAN You See It? KANs and Sentinel for Effective and Explainable Crop Field Segmentation. arXiv:2408.07040.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv:2102.04306*.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv:2108.02938*.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Ding, Z.; Dong, Q.; Xu, H.; Li, C.; Ding, X.; and Huang, Y. 2022. Unsupervised Anomaly Segmentation for Brain Lesions Using Dual Semantic-Manifold Reconstruction. In *International Conference on Neural Information Processing*, 133–144. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*.

Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *TMI*, 38(10): 2281–2292.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H. R.; and Xu, D. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, 272–284. Springer.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 574–584.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NIPS*.

Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5): 359–366.

Huang, G.-B.; Zhao, L.; and Song, Y. 2014. Deep architecture of Kolmogorov-Arnold representation. In *IJCNN*, 1001–1008. IEEE.

Huang, G.-B.; Zhao, L.; and Xing, Y. 2017. Towards theory of deep learning on graphs: Optimization landscape and train ability of Kolmogorov-Arnold representation. *Neurocomputing*, 251: 10–21.

Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-L.; and Xu, W. 2020. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In *ICASSP*, 1055–1059. IEEE.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jabri, A.; Fleet, D. J.; and Chen, T. 2022. Scalable Adaptive Computation for Iterative Generation. *arXiv:2212.11972*.

Kamnitsas, K.; Ledig, C.; Newcombe, V. F.; Simpson, J. P.; Kane, A. D.; Menon, D. K.; Rueckert, D.; and Glocker, B. 2017. Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *MIA*, 36: 61–78.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*.

Kolmogorov, A. N. 1957. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *American Mathematical Society Translations*, 28: 55–59.

Kolmogorov, A. N. 1961. *On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables*. American Mathematical Society.

Li, C.; Li, W.; Liu, H.; Liu, X.; Xu, Q.; Chen, Z.; Huang, Y.; Yuan, Y.; et al. 2024a. Flaws can be Applause: Unleashing Potential of Segmenting Ambiguous Objects in SAM. In *NIPS*.

Li, C.; Lin, M.; Ding, Z.; Lin, N.; Zhuang, Y.; Huang, Y.; Ding, X.; and Cao, L. 2022a. Knowledge condensation distillation. In *European Conference on Computer Vision*, 19–35. Springer.

Li, C.; Liu, H.; Liu, Y.; Feng, B. Y.; Li, W.; Liu, X.; Chen, Z.; Shao, J.; and Yuan, Y. 2024b. Endora: Video Generation Models as Endoscopy Simulators. *arXiv:2403.11050*.

Li, C.; Ma, W.; Sun, L.; Ding, X.; Huang, Y.; Wang, G.; and Yu, Y. 2022b. Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. *Neural Computing and Applications*, 1–14.

Li, C.; Zhang, Y.; Li, J.; Huang, Y.; and Ding, X. 2021a. Unsupervised anomaly segmentation using image-semantic cycle translation. *arXiv:2103.09094*.

Li, C.; Zhang, Y.; Liang, Z.; Ma, W.; Huang, Y.; and Ding, X. 2021b. Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, 61–65. IEEE.

Li, Z.; Guan, B.; Wei, Y.; Zhou, Y.; Zhang, J.; and Xu, J. 2024c. Mapping New Realities: Ground Truth Image Creation with Pix2Pix Image-to-Image Translation. *arXiv:2404.19265*.

Liang, X.; Zhao, L.; and Huang, G.-B. 2018. Deep Kolmogorov-Arnold representation for learning dynamics. *Access*, 6: 49436–49446.

Liu, Y.; Li, C.; Yang, C.; and Yuan, Y. 2024a. EndoGaussian: Gaussian Splatting for Deformable Surgical Scene Reconstruction. *arXiv:2401.12561*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024b. Vmamba: Visual state space model. *arXiv:2401.10166*.

Liu, Y.; Zhu, H.; Liu, M.; Yu, H.; Chen, Z.; and Gao, J. 2024c. Rolling-Unet: Revitalizing MLP's Ability to Efficiently Extract Long-Distance Dependencies for Medical Image Segmentation. In *AAAI*, volume 38, 3819–3827.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024d. Kan: Kolmogorov-arnold networks. *arXiv:2404.19756*.

Ma, J.; Li, F.; and Wang, B. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv:2401.04722*.

Mehta, S.; Mercan, E.; Bartlett, J.; Weaver, D.; Elmore, J. G.; and Shapiro, L. 2018. Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, 893–901. Springer.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 565–571. Ieee.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv:1804.03999*.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2021. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv:2104.11222*, 5: 14.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, 234–241.

Ruan, J.; and Xiang, S. 2024. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *arXiv:2402.02491*.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022. Palette: Image-to-image diffusion models.

Saito, M.; Matsumoto, E.; and Saito, S. 2017. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, 2830–2839.

Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; and Rueckert, D. 2019. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. *MIA*, 53: 197–207.

Sun, L.; Li, C.; Ding, X.; Huang, Y.; Chen, Z.; Wang, G.; Yu, Y.; and Paisley, J. 2022. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in biology and medicine*, 140: 105067.

Torbunov, D.; Huang, Y.; Yu, H.; Huang, J.; Yoo, S.; Lin, M.; Viren, B.; and Ren, Y. 2023. Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In *WACV*, 702–712.

Valanarasu, J. M. J.; Oza, P.; Hacihaliloglu, I.; and Patel, V. M. 2021. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, 36–46. Springer.

Valanarasu, J. M. J.; and Patel, V. M. 2022. Unext: Mlp-based rapid medical image segmentation network. In *MICCAI*, 23–33. Springer.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *NIPS*.

Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; and Nandi, A. K. 2022. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5): 1243–1267.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34: 12077–12090.

Xing, Y.; Zhao, L.; and Huang, G.-B. 2018. Kolmogorov-Arnold representation based deep learning for time series forecasting. In *SSCI*, 1483–1490. IEEE.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv:2401.13560*.

Xu, H.; Li, C.; Zhang, L.; Ding, Z.; Lu, T.; and Hu, H. 2024. Immunotherapy efficacy prediction through a feature re-calibrated 2.5 D neural network. *CMPB*, 249: 108135.

Yu, Y.; Buchanan, S.; Pai, D.; Chu, T.; Wu, Z.; Tong, S.; Haeffele, B.; and Ma, Y. 2024. White-Box Transformers via Sparse Rate Reduction. *NeurIPS*, 36.

Zhang, Y.; Li, C.; Lin, X.; Sun, L.; Zhuang, Y.; Huang, Y.; Ding, X.; Liu, X.; and Yu, Y. 2021. Generator versus segmentor: Pseudo-healthy synthesis. In *MICCAI*, 150–160. Springer.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep learning in MIA and multimodal learning for clinical decision support*, 3–11. Springer, Cham.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv:2401.09417*.