Multi-view Evidential Learning-based Medical Image Segmentation

Chao Huang[1], **Yushu Shi**[1], **Waikeung Wong***[2,3], **Chengliang Liu**[4], **Wei Wang**[1], **Zhihua Wang**[5], **Jie Wen***[6]

[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[2]School of Fashion and Textiles, Hong Kong Polytechnic University
[3]Laboratory for Artificial Intelligence in Design, Hong Kong
[4]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
[5]Department of Computer Science, City University of Hong Kong
[6]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
huangch253@mail.sysu.edu.cn, shiysh8@mail2.sysu.edu.cn, calvin.wong@polyu.edu.hk, liucl1996@163.com,
wangwei29@mail.sysu.edu.cn, zhihua.wang@my.cityu.edu.hk, wenjie@hit.edu.cn

Abstract

Medical image segmentation provides useful information about the shape and size of organs, which is beneficial for improving diagnosis, analysis, and treatment. Despite traditional deep learning-based models can extract domain-specific knowledge, they face a generalization bottleneck due to the limited embedded knowledge scope. Vision foundation models have been demonstrated to be effective in extracting generalizable knowledge, but they cannot extract domain-specific knowledge without fine-tuning. In this work, we propose a novel multi-view evidential learning-based framework, which can extract both domain-specific and generalizable knowledge from multi-view features by combining the advantages of traditional and vision foundation models. Specifically, a novel multi-view state space model (MV-SSM) is designed to extract task-related knowledge while removing redundant information within multi-view features. The proposed MV-SSM utilizes Mamba, a state space model, to model cross-view contextual dependencies between domain-specific and generalizable features. Additionally, evidential learning is adopted to quantify the segmentation uncertainty of the model for boundary. In special, variational Dirichlet is introduced to characterize the distribution of the result probabilities, parameterized with collected evidence to quantify uncertainty. As a result, the model can reduce the segmentation uncertainties of boundaries by optimizing the parameters of the Dirichlet distribution. Experimental results on three datasets show that our method obtains superior segmentation performance.

Introduction

Medical image segmentation is a critical component of computer-aided diagnosis, which plays a pivotal role in relative applications. It can provide useful information about the shape and size of organs, which is beneficial for improving diagnosis, analysis, and treatment. In recent years, deep learning techniques have been extensively applied to medical image segmentation and have achieved remarkable success (Qureshi et al. 2023). These methods (Butoi et al. 2023; Rahman and Marculescu 2023) obtained promising performance and significant improvement in disease diagnosis.

In order to facilitate the extraction of domain-specific knowledge, previous methods (Heidari et al. 2023; Roy et al. 2023) need to be trained on specific medical image datasets. In fact, these methods adopt the single view-based medical image segmentation framework, which extracts information for segmentation from a perspective of the original input. However, the single view-based framework restricts the scope of encoded knowledge for the model. Although some methods achieve information augmentation by introducing multi-scale representation learning (Zhan et al. 2023), they fail to consider multiple scales to balance the distribution of feature weights. Consequently, existing methods still cannot effectively explore sufficient information to overcome the generalization bottleneck caused by the limited embedded knowledge scope of the single view-based framework.

Drawing inspiration from biological cognitive mechanisms, visual information from different views (e.g., different observation and viewing angles) can be mutually correlated and complementary (Liu et al. 2022). Consequently, multi-view learning can be adopted to as a solution to overcome the generalization bottleneck. In fact, there are several works (Liu et al. 2024a) have attempted to achieve medical image segmentation based on multi-view learning. Despite achieving superior performance compared to single-view methods, these methods still exhibit limitations. Specifically, current multi-view methods are still trained on limited datasets. They remain incapable of extracting effective generalizable knowledge.

Recently, vision foundation models such as SAM (Kirillov et al. 2023) and CLIP (Radford et al. 2021) have been extensively applied to various vision tasks (Awais et al. 2023) and achieved remarkable success. These vision foundation models have strong feature extraction and zero-shot transfer capabilities, since they are usually trained on large scale datasets. Thus, vision foundation models can extract richer and generalizable information (Frintrop, Rome, and Christensen 2010). However, vision foundation models cannot extract domain-specific knowledge if directly applied to medical images segmentation due to domain differences (Zhang, Shen, and Jiao 2024). Thus, it is necessary to

[*]Corresponding Authors

fine-tune vision foundation models when applying them for medical image segmentation. Fine-tuning visual foundation models is a challenging task due to the extremely high cost of acquiring high-quality labeled data.

To address these challenges, we propose a novel multi-view evidential learning-driven framework for medical image segmentation. The proposed method leverages the strengths of two types of models (traditional deep learning model and vision foundation model) to extract domain-specific and generalizable knowledge from multi-view features. Specifically, we propose a novel multi-view state space model (MV-SSM). It is adopted to extract task-related knowledge while removing redundant information within multi-view features. Mamba, a state space model, is utilized in the proposed MV-SSM to model cross-view contextual dependencies between domain-specific and generalizable features. Moreover, we propose a novel Mamba block for domain-specific models, which is adopted to extract rich domain-specific knowledge during training through multi-scale representation. Additionally, the evidential learning is also adopted to quantify the segmentation uncertainty of the model for boundary. In detail, variational Dirichlet is introduced to characterize the distribution of the result probabilities and parameterized with collected evidence, which denotes the segmentation results and the uncertainty distribution. The variational Dirichlet can be adopted to efficiently quantify the boundary segmentation uncertainty. As a result, the model can reduce the segmentation uncertainties of boundaries by optimizing the parameters of the Dirichlet distribution. Our main contributions are summarized as follows:

- A novel multi-view evidential learning-driven framework is proposed for medical image segmentation, which combines the advantages of domain-specific and vision foundation models. Thus, the proposed method can extract both domain-specific and generalizable knowledge from multi-view features.
- A novel multi-view state space model is designed to extract task-related knowledge. It adopts Mamba to model cross-view contextual dependencies between domain-specific and generalizable features. Additionally, a novel Mamba block is proposed for the domain-specific model to extract rich domain-specific knowledge.
- The evidential learning is adopted to quantify the boundary segmentation uncertainty of the model. As a result, the model can reduce the segmentation uncertainties of boundaries by optimizing the relative parameters.

Related Work

Medical Image Segmentation

Medical image segmentation refers to identifying organ or lesion pixels from medical images (Yao et al. 2023). U-Net (Ronneberger, Fischer, and Brox 2015) is a widely adopted model for medical images. Various methods have been proposed to improve its performance. For instance, UNet++ (Zhou et al. 2019) adds dense skip connections to enhance feature propagation. Attention-UNet (Wang, Li, and Zhuang 2021) incorporated an attention mechanism to focus on relevant regions. TransUNet (Chen et al. 2021) and Swin-UNet (Cao et al. 2022) combined Transformer and UNet architectures for achieving better performance. However, these domain-specific models are typically trained with single-view data only. It limits the model's ability to encode comprehensive knowledge, although these models can extract domain-specific knowledge. Several methods have incorporated multi-view learning to enhance the scope of encoded knowledge for domain-specific models. However, these methods are still trained on a limited dataset, which cannot extract effective generalizable knowledge. In this work, we present a novel multi-view evidential learning framework that leverages the strengths of traditional deep learning model and vision foundation model. The proposed method enables the model to extract both domain-specific and generalizable knowledge from multi-view features.

Evidential Learning

Recently, some uncertainty estimation methods based on the Dirichlet distribution have been applied to the computer vision community. For example, PriorNet evidence networks (Sensoy, Kaplan, and Kandemir 2018) modeled the Dirichlet distribution using subjective logic. Ensemble distribution distillation (Malinin, Mlodozeniec, and Gales 2019) obtained the Dirichlet distribution by distilling from the predictions of multiple models. PostNet (Charpentier, Zügner, and Günnemann 2020) utilized normalized flows and Bayesian losses to estimate uncertainty and obtain the Dirichlet distribution during training. Median smoothing applied to the Dirichlet model significantly improves its ability to handle adversarial examples (Kopetzki et al. 2021). Motivated by these methods, we incorporate evidence learning to tackle the issue of model boundary segmentation uncertainty. Specifically, we quantify uncertainty by modeling the distribution of the result with a variational Dirichlet function and minimize the boundary segmentation uncertainty by optimizing the parameters of the Dirichlet distribution.

State Space Model

The concept of the state space model was initially introduced in the Structured State Space Sequence Model (Gu, Goel, and Ré 2021) (S4 model), presenting a distinctive architecture capable of effectively modeling global information compared to conventional CNN or Transformer. The subsequent Hungry Hungry Hippos (H3 model) (Mehta et al. 2022) further refined and expanded upon this foundation, enabling the model to perform competitively with Transformers in language model tasks. Mamba (Gu and Dao 2023) introduced an input adaptation mechanism. It significantly improves the inference speed of the State Space model. Thus, Mamba demonstrates superior overall metrics compared to transformers of equivalent scale. Vision Mamba (Zhu et al. 2024) and Vmaba (Liu et al. 2024b) expand the application of Mamba to visual tasks. These adaptations achieved impressive results in classification and segmentation, notably in medical image segmentation (Ma, Li, and Wang 2024; Zhang et al. 2024). The potential of this model in multi-view learning remains an area that has not been thoroughly explored. In this paper, a novel multi-view state space model (MV-SSM) is proposed to extract task-related knowledge from multi-view features. MV-SSM leverages Mamba to model cross-view contextual dependencies between domain-specific and generalizable features. Furthermore, a novel Mamba block is introduced to extract the domain-specific knowledge.

Methodology

Overview

The overall architecture of the proposed method is shown in Figure 1. Segment Anything Model (Kirillov et al. 2023) is adopted as the vision foundation model. Given a medical image $\mathbf{I}$, the SAM image encoder and domain-specific encoder are adopted to extract generalizable knowledge and domain-specific knowledge from the medical image. Then, the Multi-view State Space Model (MV-SSM) is adopted to model cross-view contextual dependencies and extract task-related knowledge between domain-specific and generalizable features. Additionally, a novel Mamba Residual Channel Cross-Fusion Transformer (Mamba Residual CCT) is designed to extract richer domain-specific knowledge. Then, a domain-specific decoder is adopted to collect evidence. We parameter a variational Dirichlet distribution based on the collected evidence, which is adopted to quantify the boundary segmentation uncertainty. Finally, the boundary segmentation uncertainty of the model can be reduced by optimizing the parameters of the Dirichlet distribution.

Preliminaries: State Space Model

Mamba is a State Space Model (SSM), which originates from continuous systems that transform a sequence $x(t) \rightarrow y(t)$ through a hidden state function $h(t) \in \mathbb{R}^N$. It can be expressed as the following equation:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a state matrix, and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^N$ are the projection matrixs.

Mamba discretizes the aforementioned continuous system by introducing a time scale parameter $\mathbf{\Delta A}$ and $\mathbf{\Delta B}$ that converts continuous parameters $\mathbf{A}$, $\mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}, \overline{\mathbf{B}}$. The zero-order hold is adopted as the discretization rule, defined as follows:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{J}) \cdot \Delta\mathbf{B}, \tag{2}$$

where $\mathbf{J}$ is a unit matrix.

After discretizing A and B, Eq. 1 can be written as follows:

$$h'(t) = \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t), \quad y(t) = \mathbf{C}h(t). \tag{3}$$

Furthermore, a global convolution is adopted to compute the output, defined as:

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{AB}}, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \quad y = x * \overline{\mathbf{K}}, \tag{4}$$

where L denotes the length of the input sequence and $\mathbf{K} \in \mathbb{R}^L$ is a structured convolution kernel.

Multi-view State Space Model

Previous medical image segmentation methods focuses on extracting domain-specific knowledge, which is essential for accurate segmentation. However, these methods cannot extract generalizable knowledge, which limits their performance on unseen data. Vision foundation models are effective in extracting generalizable knowledge since they are trained on large-scale datasets. Unfortulately, these models typically struggle to capture domain-specific knowledge, which results in suboptimal performance on the medical image segmentation task.

To address these limitations, we propose a multi-view state space module that can simultaneously the strengths of two types of models (traditional deep learning model and vision foundation model) to extract domain-specific and generalizable knowledge from multi-view features. Specifically, a novel Pixel-Window Channel Mamba (PWC-Mamba) is designed to extract task-related knowledge from both domain-specific and generalizable knowledge. PWC-Mamba will be detailed in the next subsection. The extracted features from the two knowledge sources are then concatenated in the channel dimension, defined as:

$$\mathbf{F}_V = \mathrm{Enc}_{\mathrm{SAM}}(\mathbf{I}), \mathbf{F}_{\mathrm{D}} = \mathrm{Enc}_{\mathrm{DSE}}(\mathbf{I}), \tag{5}$$

$$\mathbf{F} = [\mathrm{PPM}(\mathbf{F}_{\mathrm{V}}), \mathrm{PPM}(\mathbf{F}_{\mathrm{D}})], \tag{6}$$

where $\mathrm{Enc}_{\mathrm{SAM}}$ denotes the SAM image encoder, and $\mathrm{Enc}_{\mathrm{DSE}}$ is the domain-specific image encoder. The concatenated features are passed through L PWC-Mamba modules to further exploit the semantic relevance between domain-specific and generalizable knowledge and model cross-view contextual dependencies.

Pixel-Window Channel Mamba

Considering that global and local information in segmentation models contribute to accurately segmenting foreground objects, it is necessary to model global and local information simultaneously. Mamba combines the strengths of Convolution Neural Network (CNN) and Recurrent Neural Network (RNN), which provides a more efficient option to achieve this goal.

Therefore, inspired by previous work, the Pixel-Window Channel Mamba (PWC-Mamba) is designed to model global and local information simultaneously. Figure 1 shows details of the proposed PWC-Mamba. Specifically, it is achieved by simultaneously modeling pixel-level and patch-level information. The large window partition is adopted to split a feature with $H \times W$ resolution into many sub-windows. Therefore, Mamba can more effectively model the dependency relationships between local neighboring pixels by continuously inputting sub-windows. Considering the rich feature correlations within different channels in the same location of the feature map, we propose a novel window-based channel attention mechanism (WCA). It first duplicates the feature tensor and concats in channel splicing. Then, it applies a convolution with a large kernel to fuse the channel information. Each channel adopts different convolution kernel parameters to extract more diverse features. WCA models the dependency relationship between pixels in a large window

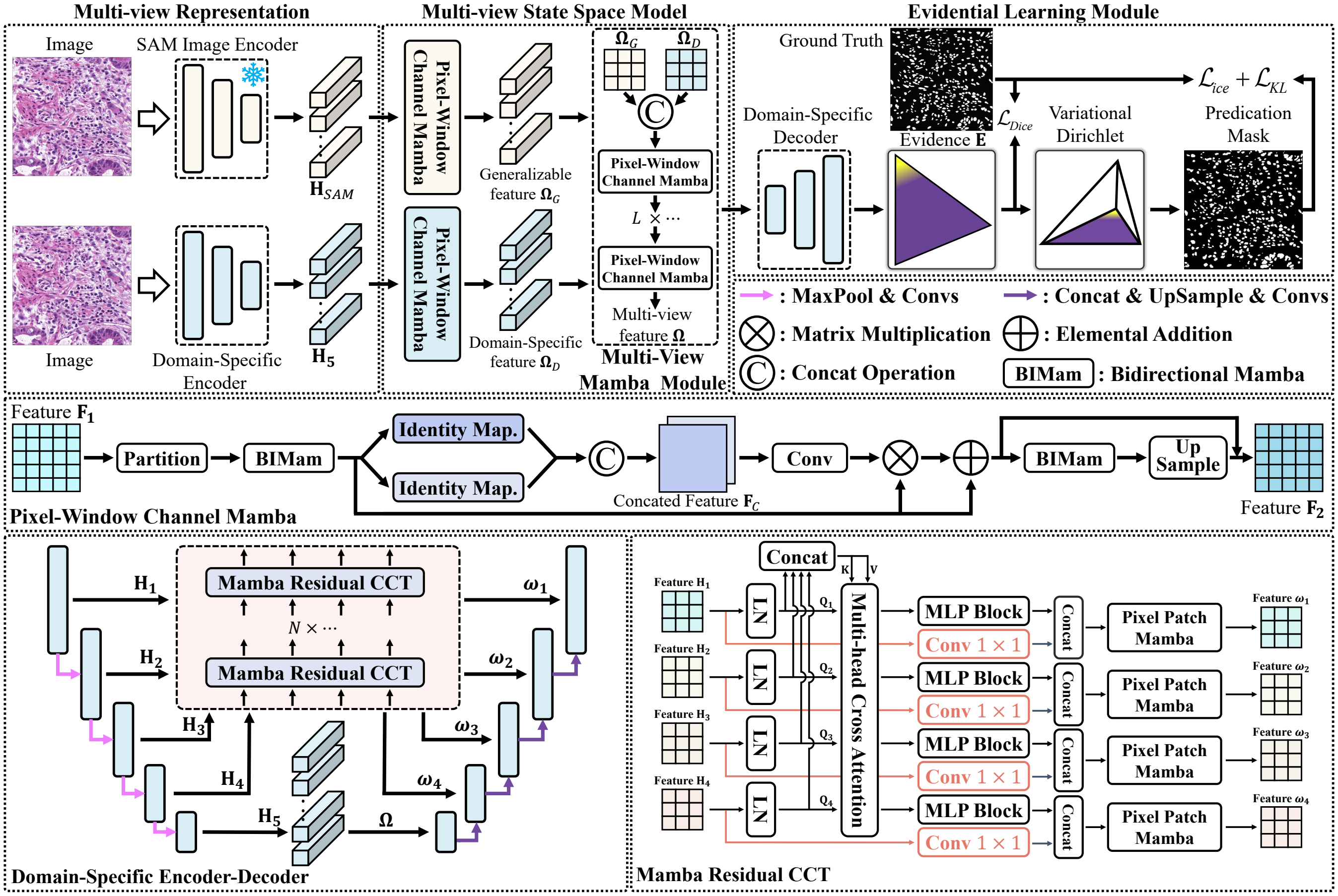

Figure 1: Overview of the proposed method, 'Snow' symbol indicates frozen parameters. Our method uses a SAM image encoder and a domain-specific encoder to respectively extract generalizable and domain-specific knowledge from medical images. The Multi-view State Space Model (MV-SSM) is designed to model cross-view contextual dependencies and extract task-related knowledge. The Mamba Residual Channel Cross-Fusion Transformer (Mamba Residual CCT) is designed to further enrich domain-specific knowledge. Then our method collects evidence with a domain-specific decoder and parameterizes a variational Dirichlet distribution to quantify the boundary segmentation uncertainty. Finally, the parameters of Dirichlet are optimized to reduce the boundary segmentation uncertainty of the model.

while preserving local features. Finally, WCA is applied to the origin feature tensor through element-wise multiplication and addition. The element-wise multiplication and addition with the original feature tensor ensure that the fine-grained details captured by pixel-level modeling are not lost. The window-based channel attention is defined as:

$$\mathbf{M}_1 = \text{Identify}(\mathbf{F}_\text{o}), \mathbf{M}_2 = \text{Identify}(\mathbf{F}_\text{o}), \tag{7}$$

$$\mathbf{M} = \text{Conv}_{7\times 7}([\mathbf{M}_1, \mathbf{M}_2]), \mathbf{F} = \mathbf{M} \times \text{F}_\text{o} + \text{F}_\text{o}, \tag{8}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are the identify mapping results, respectively. $\mathbf{F}_o$ is the pixel-level feature, $[\cdot]$ denotes the concatenation operation, $\mathbf{M}$ denotes the output after a 7×7 convolution kernel, and $\mathbf{F}$ denotes the channel attention output.

Then, a Mamba block is adopted for larger scale global information modeling. Meanwhile, the residual connection is also adopted to preserve the original information. Moreover, the Mamba block suffers from information loss for earlier input elements. Thus, we adopt the Bidirectional Mamba (BiMam) (Wang et al. 2024) which performs both forward and backward scans to avoid information loss.

Mamba Residual Cross-Fusion Transformer

The Channel-wise Cross fusion Transformer (CCT) (Wang et al. 2022) has been demonstrated to be effective in fusing multi-scale features and achieving information enhancement. However, it still has some limitations. For example, it does not explicitly preserve the original information before fusion, which results in information loss. Moreover, it only uses a single feature extraction strategy, which is not optimal for all medical image segmentation tasks. As shown in figure 1, we propose a Mamba Residual CCT model, which combines the advantages of CCT and the proposed PWC-Mamba. Specifically, a 1×1 convolution is adopted as a residual connection to preserve the original information before fusion. It avoids information loss by extracting the rich channel information. PWC-Mamba is applied to simultaneously model cross-scale global and local information, which enables the model to more effectively mine domain-specific knowledge and helps the model to obtain more accurate segmentation results.

Evidential Learning-Driven Segmentation Uncertainty Reduction

Existing medical image segmentation methods typically adopt a Softmax function to convert the output of decoder into the predicted probability. However, the Softmax function is prone to be overconfident on prediction results. The boundary ambiguity in medical images poses a challenge to accurate boundary segmentation. In other words, boundary segmentation remains a high degree of uncertainty.

To address this issue, we incorporate an activation function (Softplus) after the decoder to ensure non-negative values. These non-negative outputs are interpreted as evidence $\mathbf{E} = \mathrm{SoftPlus}(\mathbf{O})$, from which a Dirichlet distribution $D(\boldsymbol{p} \mid \boldsymbol{\alpha})$ can be derived. The Dirichlet distribution $D(\boldsymbol{p} \mid \boldsymbol{\alpha})$ is considered as the conjugate prior to the multinomial distribution. It provides a predictive distribution for the segmentation results, defined as follows:

$$D(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \begin{cases} \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} & for \boldsymbol{p} \in \mathcal{Q}_C \\ 0 & otherwise \end{cases} \tag{9}$$

where $\boldsymbol{\alpha} = \left[\alpha^1, \ldots, \alpha^c\right]$ is the Dirichlet distribution parameters, $\boldsymbol{p} = [p_1, \ldots p_k]$ are the parameters of the multinomial distribution, $\beta(\boldsymbol{\alpha})$ is the high-dimensional multinomial beta function, and $\mathcal{Q}_C$ is the $K-1$ dimensional unit simplex, defined as:

$$\mathcal{Q}_C = \left\{ \boldsymbol{p} \left| \sum_{c=1}^{C} p^c = 1 \text{ and } 0 \leq p^1, \ldots, p^C \leq 1 \right. \right\}. \tag{10}$$

After obtaining the Dirichlet distribution $D(\boldsymbol{p} \mid \boldsymbol{\alpha})$, we focus on quantifying uncertainty and optimizing parameters to minimize segmentation uncertainty. Therefore, Subjective Logic (SL) (Jsang 2018) is adopted to optimize the parameters of Dirichlet distribution. In multi-class segmentation, subjective logic provides a theoretical framework for the relationship between Dirichlet distribution parameters for confidence and uncertainty. Specifically, it provides a mass belief and uncertainty for the segmentation results. Therefore, for a two-dimensional input image $\mathbf{I}$ and decoder output $\mathbf{O}$ without the Softmax function, all $C+1$ quality values are non-negative and sum to 1, defined as follows:

$$u_{i,j}^c + \sum_{c=1}^{C} b_{i,j}^c = 1, \tag{11}$$

where $u_{i,j}^c$ and $b_{i,j}^c$ mean the segmentationn probability of the pixel (i, j) for the c-th class and the overall uncertainty of the pixel (i, j) in the input image $\mathbf{I}$, respectively. Subsequently, SL links the evidence of pixel (i, j) to the Dirichlet distribution parameters $\alpha_{i,j}^c = e_{i,j}^c + 1$, where $\alpha_{i,j}^c \in \boldsymbol{\alpha}$ and $e_{i,j}^c \in \mathbf{E}$. Therefore, the mass of belief and uncertainty for the pixel (i, j) can be expressed as follows:

$$b_{i,j}^c = \frac{e_{i,j}^c}{Q} = \frac{\alpha_{i,j}^c - 1}{Q} \quad \text{and} \quad u_{i,j} = \frac{C}{Q}, \tag{12}$$

where $Q = \sum_{c=1}^{C} \left(e_{i,j}^c + 1\right)$ is known as the Dirichlet strength. It describes that the higher the allocated belief mass, the more evidence is obtained for pixel (i, j). Conversely, the less evidence obtained, the greater the overall uncertainty for the segmentation of pixel (i, j).

Considering that on the simplex, the ideal Dirichlet distribution should concentrate on the vertex corresponding to the class label. This means that the Dirichlet distribution parameters should be as close as possible to 1 except for the correctly labeled parameters. Therefore, it is necessary to design a loss function that optimizes the model parameters by optimizing the Dirichlet distribution parameters to minimize the segmentation uncertainty of model. It allows the boundary segmentation results in the image to collect more evidence. Specifically, we adopt the cross-entropy loss function $\mathcal{L}_{ce} = \sum_{c=1}^{C} -y^c \log\left(p^c\right)$, associating the Dirichlet distribution with the belief distribution within the subjective logical framework. Based on the evidence collected from the decoder, probabilities for different classes and uncertainties for different pixels are obtained. Furthermore, the cross-entropy loss can be reformulated as:

$$\mathcal{L}_{ice} = \int \left[\sum_{c=1}^{C} -y^c \log(p^c) \right] \frac{1}{\beta(\alpha)} \prod_{c=1}^{C} \left(p^c\right)^{\alpha^c - 1} d\mathbf{p} \tag{13}$$

$$= \sum_{c=1}^{C} y^c \left(\psi\left(\mathcal{Q}_C\right) - \psi\left(\alpha^c\right)\right), \tag{14}$$

where $\psi(\cdot)$ is digamma function, and y^c is the ground truth labels. Additionally, we introduce the Kullback-Leibler (KL) divergence loss function to ensure that incorrect labels generate less evidence and avoid penalizing the Dirichlet parameter of the ground-truth class to $\mathbf{1}$, defined as follows:

$$\mathcal{L}_{KL} = \log \left(\frac{\Gamma\left(\sum_{c=1}^{C} \tilde{\alpha}^c\right)}{\Gamma(C) \sum_{c=1}^{C} \Gamma\left(\tilde{\alpha}^c\right)} \right) \tag{15}$$

$$+ \sum_{c=1}^{C} \left(\tilde{\alpha}^c - 1\right) \left[\psi\left(\tilde{\alpha}^c\right) - \psi\left(\sum_{c=1}^{C} \tilde{\alpha}^c \right) \right], \tag{16}$$

where $\Gamma(\cdot)$ is the gamma function, $\tilde{\boldsymbol{\alpha}}^c = y^c + (1 - y^c) \odot \boldsymbol{\alpha}^c$.

Loss function

The loss function $\mathcal{L}$ used in this work is consist of Dice loss $\mathcal{L}_{Dice}$, $\mathcal{L}_{ice}$, and $\mathcal{L}_{KL}$, formulated as:

$$\mathcal{L} = \mathcal{L}_{Dice} + \mathcal{L}_{ice} + \lambda_t \mathcal{L}_{KL}. \tag{17}$$

where λ_t is a hyperparameter, which is adopted to balance the expected $\mathcal{L}_{ice}$ and $\mathcal{L}_{KL}$. To prevent the network from overemphasizing $\mathcal{L}_{KL}$ in the early stages, which would result in insufficient exploration of the parameter space, and a nearly flat uniform distribution is finally obtained, a small value of λ_t is employed and gradually increase it during training.

The Dice loss $\mathcal{L}_{ice}$ is adopted to maximize the correspondence between the segmentation result and the ground-truth label, defined as follows:

$$L_{Dice} = 1 - \sum_{i=1}^{N} \sum_{j=1}^{C} \frac{1}{NC} \cdot \frac{2|d_{ij} \cap o_{ij}|}{(|d_{ij}| + |o_{ij}|)}, \tag{18}$$

Method	Venue	MoNuSeg		GlaS		TNBC		DRIVE	
		Dice	MIoU	Dice	MIoU	Dice	MIoU	Dice	MIoU
UNet	MICCAI'2015	76.45	62.86	85.45	74.78	80.64	67.62	81.41	68.64
UNet++	TMI'2017	79.49	66.04	87.36	79.03	81.19	68.44	81.14	68.27
DCA	EAAI'2023	79.50	65.97	89.90	81.68	82.45	70.14	76.66	62.16
AttUNet	MICCAI'2021	76.67	63.47	88.80	80.69	81.25	68.59	80.39	67.21
Swin-UNet	ECCV'2022	77.69	63.77	89.58	82.06	72.50	58.10	64.24	47.48
TransUNet	ArXiv'2021	78.53	65.05	88.40	80.40	78.30	64.37	78.15	64.20
UCTransNet	AAAI'2022	79.09	66.68	89.76	81.91	82.07	70.33	82.67	70.54
TGANet	MICCAI'2022	70.09	61.63	87.96	82.74	—	—	—	—
VMUNet	ArXiv'2024	78.85	69.20	89.85	82.53	82.22	69.98	82.61	70.41
MedSAM	Nat. Commun'2024	—	21.40	—	54.79	75.74	61.42	—	—
VMUNet v2	ArXiv'2024	63.89	47.02	87.88	79.31	68.01	52.40	79.02	65.34
LViT-T	TMI'2023	80.15	67.00	90.02	82.68	—	—	—	—
SAM	ICCV'2023	30.24	18.21	58.46	42.81	20.44	12.53	—	—
Ours		**81.88**	**69.45**	**90.59**	**83.70**	**83.75**	**72.23**	**84.12**	**72.67**

Table 1: Quantitative comparison of our method and other SOTA methods on MoNuSeg, GlaS, TNBC, and DRIVE datasets.

where N is the number of pixels in the image, C is the number of classes, o_{ij} and d_{ij} are the ground truth and predicted segmentation output, respectively.

Experiments

Datasets and Evaluation Metrics

The proposed method is evaluated on four medical image segmentation datasets: MoNuSeg (Kumar et al. 2017), GlaS (Sirinukunwattana et al. 2017), TNBC (Naylor et al. 2018), and DRIVE (Staal et al. 2004). MoNuSeg contains 30 digital microscopic tissue images of several patients. The training, validation, and test sets are organized as (Wang et al. 2022). GlaS has 85 images for training and 80 images for testing. TNBC (Naylor et al. 2018) contains a total of 50 images with a total of 4,022 annotated cells. DRIVE (Staal et al. 2004) is a retinal vessel segmentation dataset that contains 40 images. DRIVE dataset is divided into 20 images for training and 20 images for testing. Dice score and IoU (Wang et al. 2022) are used to evaluate the performance.

Implementation Details

Our model is optimized by AdamW (Loshchilov and Hutter 2017) and the cosine annealing learning rate scheduler is adopted. The learning rate is set to 1e-3. The size of input image and batch size are 224×224 and 2, respectively. All experiments are conducted on a single NVIDIA A100 GPU with 80GB memory. We select several State-of-the-art methods as baselines, including U-Net (Ronneberger, Fischer, and Brox 2015), DoubleUnet (Ates, Mohan, and Celik 2023), AttUNet (Wang, Li, and Zhuang 2021), Swin-UNet (Cao et al. 2022), TransUNet (Chen et al. 2021), UCTransNet (Wang et al. 2022), MedSAM (Ma et al. 2024), LViT (Li et al. 2023) and TGANet (Tomar et al. 2022), SAM (Kirillov et al. 2023), VMUNet (Ruan and Xiang 2024), and VMUNet v2 (Zhang et al. 2024). More details are reported in supplementary materials.

Comparsion with State-of-the-art Methods

The quantitative comparison results are presented in Table 1. It shows that the proposed method consistently outperforms domain-specific and vision foundation models across all datasets. Specifically, on the MoNuSeg dataset, the proposed method obtains an 81.88% Dice score and 69.45% mIoU, which surpasses the previous best results. Similarly, on the GlaS dataset, our method has a 90.59% Dice score and 83.70% mIoU. Moreover, our method obtains an 83.75% Dice score and 72.23% mIoU on the TNBC dataset. The performance improvements are 1.3% and 1.9%, respectively. Additionally, our method obtains an 84.12% Dice score and 72.67% mIoU on the DRIVE dataset. Figure 2 shows the visualization of segmentation results. It shows that our method effectively achieves superior boundary segmentation. These remarkable results can be attributed to the carefully designed multi-view learning framework, which can combine the advantages of traditional and vision foundation models to extract both domain-specific and generalizable knowledge from multi-view features. Besides, variational Dirichlet is introduced to characterize the distribution of the result probabilities, parameterized with collected evidence to quantify uncertainty. Thus, the model can reduce the segmentation uncertainties of boundaries, which leads to further performance improvement.

Ablation Study

Effectiveness of the proposed components Table 2 reports the ablation results of the effectiveness of each proposed component. These results indicate that these components are effective for the proposed method. Specifically, introducing the Dirichlet distribution explicitly improves the performance of our model across all datasets. On the MoNuSeg dataset, the Dice score and MIoU score respectively increase by 0.55% and 0.15%. On the GlaS dataset, these metrics exhibit respective gains of 0.22% and 0.52%. It demonstrates the effectiveness of minimizing boundary

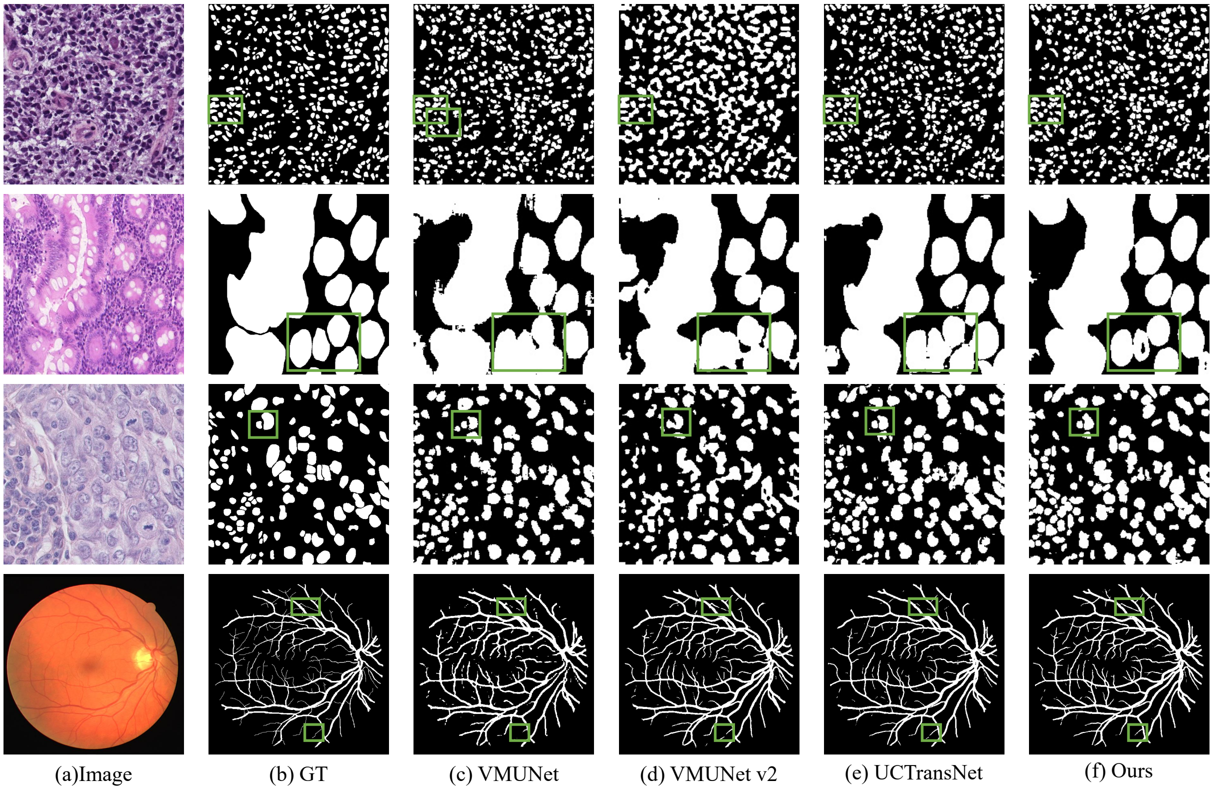

Figure 2: Qualitative comparison of the proposed method and other SOTA methods on MoNuSeg and GlaS datasets.

UCTransNet	Uncertainty	Mamba Residual CCT	MV-SSM	MoNuSeg		GlaS	
				Dice	MIoU	Dice	MIoU
✓				79.09	66.68	89.76	81.91
✓	✓			79.64	66.33	89.98	82.43
✓	✓	✓		80.22	67.19	90.11	82.45
✓	✓		✓	81.30	68.61	90.12	82.94
✓	✓	✓	✓	**81.88**	**69.45**	**90.67**	**83.67**

Table 2: Ablation study on the effectiveness of the proposed components on MoNuSeg and GlaS datasets.

segmentation uncertainty through Dirichlet distribution parameter optimization. Notably, the results show that integrating the proposed MV-SSM yields more performance gains. Specifically, the Dice score and MIoU score increase by 1.66% and 2.28% on MoNuSeg, respectively. Similarly, the Dice score and MIoU score increase by 0.14% and 0.51%. The performance improvement is due to the multi-view learning framework effectively integrating domain-specific and generalization knowledge. Moreover, the MV-SSM module can effectively extract task-relevant knowledge through the exploration of cross-view contextual information, which is beneficial for performance improvement.

Conclusion

In this work, we proposes a novel multi-view evidential learning framework, which leverages the advantage of both traditional deep learning and vision foundation models. Our proposed framework can extract both domain-specific and generalizable knowledge from multi-view features. Specifically, a novel multi-view state space model (MV-SSM) is designed to extract task-relevant knowledge. The MV-SSM utilizes Mamba to model the cross-view contextual dependencies between domain-specific and generalizable features. Additionally, evidential learning is adopted to quantify the segmentation uncertainty of the model for boundary. Specifically, Variational Dirichlet is adopted to characterize the distribution of result probabilities, parameterized with collected evidence to quantify uncertainty. It enables the model to reduce segmentation uncertainties by optimizing the parameters of the Dirichlet distribution. Experimental results on medical image segmentation datasets indicate that our method outperforms other state-of-the-art methods.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62301621), and Shenzhen Science and Technology Program (No. 20231121172359002, No. 202412023000572).

References

Ates, G. C.; Mohan, P.; and Celik, E. 2023. Dual Cross-Attention for medical image segmentation. *Engineering Applications of Artificial Intelligence*, 126: 107139.

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; and Khan, F. S. 2023. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*.

Butoi, V. I.; Ortiz, J. J. G.; Ma, T.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2023. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21438–21451.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Charpentier, B.; Zügner, D.; and Günnemann, S. 2020. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33: 1356–1367.

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 preprint arXiv:2102.04306*.

Frintrop, S.; Rome, E.; and Christensen, H. I. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1): 1–39.

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

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Heidari, M.; Kazerouni, A.; Soltany, M.; Azad, R.; Aghdam, E. K.; Cohen-Adad, J.; and Merhof, D. 2023. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 6202–6212.

Jsang, A. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kopetzki, A.-K.; Charpentier, B.; Zügner, D.; Giri, S.; and Günnemann, S. 2021. Evaluating robustness of predictive uncertainty estimation: Are Dirichlet-based models reliable? In *International Conference on Machine Learning*, 5707–5718. PMLR.

Kumar, N.; Verma, R.; Sharma, S.; Bhargava, S.; Vahadane, A.; and Sethi, A. 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7): 1550–1560.

Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; and Hong, Q. 2023. Lvit: language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*.

Liu, C.; Liu, H.; Zhang, X.; Guo, J.; and Lv, P. 2024a. Multi-scale and multi-view network for lung tumor segmentation. *Computers in Biology and Medicine*, 172: 108250.

Liu, D.; Gao, Y.; Zhangli, Q.; Han, L.; He, X.; Xia, Z.; Wen, S.; Chang, Q.; Yan, Z.; Zhou, M.; et al. 2022. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 485–495. Springer.

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

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

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

Malinin, A.; Mlodozeniec, B.; and Gales, M. 2019. Ensemble Distribution Distillation. In *International Conference on Learning Representations*.

Mehta, H.; Gupta, A.; Cutkosky, A.; and Neyshabur, B. 2022. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*.

Naylor, P.; Laé, M.; Reyal, F.; and Walter, T. 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2): 448–459.

Qureshi, I.; Yan, J.; Abbas, Q.; Shaheed, K.; Riaz, A. B.; Wahid, A.; Khan, M. W. J.; and Szczuko, P. 2023. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90: 316–352.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rahman, M. M.; and Marculescu, R. 2023. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6222–6231.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. 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*, 234–241. Springer.

Roy, S.; Koehler, G.; Ulrich, C.; Baumgartner, M.; Petersen, J.; Isensee, F.; Jaeger, P. F.; and Maier-Hein, K. H. 2023. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 405–415. Springer.

Ruan, J.; and Xiang, S. 2024. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Sirinukunwattana, K.; Pluim, J. P.; Chen, H.; Qi, X.; Heng, P.-A.; Guo, Y. B.; Wang, L. Y.; Matuszewski, B. J.; Bruni, E.; Sanchez, U.; et al. 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35: 489–502.

Staal, J.; Abràmoff, M. D.; Niemeijer, M.; Viergever, M. A.; and Van Ginneken, B. 2004. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4): 501–509.

Tomar, N. K.; Jha, D.; Bagci, U.; and Ali, S. 2022. TGANet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 151–160. Springer.

Wang, H.; Cao, P.; Wang, J.; and Zaiane, O. R. 2022. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2441–2449.

Wang, J.; Chen, J.; Chen, D.; and Wu, J. 2024. Large Window-based Mamba UNet for Medical Image Segmentation: Beyond Convolution and Self-attention. *arXiv preprint arXiv:2403.07332*.

Wang, S.; Li, L.; and Zhuang, X. 2021. AttU-Net: attention U-Net for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, 302–311. Springer.

Yao, W.; Bai, J.; Liao, W.; Chen, Y.; Liu, M.; and Xie, Y. 2023. From CNN to Transformer: A Review of Medical Image Segmentation Models. *arXiv preprint arXiv:2308.05305*.

Zhan, B.; Song, E.; Liu, H.; Gong, Z.; Ma, G.; and Hung, C.-C. 2023. CFNet: A medical image segmentation method using the multi-view attention mechanism and adaptive fusion strategy. *Biomedical Signal Processing and Control*, 79: 104112.

Zhang, M.; Yu, Y.; Gu, L.; Lin, T.; and Tao, X. 2024. Vm-unet-v2 rethinking vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2403.09157*.

Zhang, Y.; Shen, Z.; and Jiao, R. 2024. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 108238.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6): 1856–1867.

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 preprint arXiv:2401.09417*.