S³-Mamba: Small-Size-Sensitive Mamba for Lesion Segmentation

Gui Wang[1,2], Yuexiang Li[3], Wenting Chen[4], Meidan Ding[1], Wooi Ping Cheah[2], Rong Qu[5], Jianfeng Ren[2*], Linlin Shen[1*]

[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China
[2]School of Computer Science, University of Nottingham Ningbo China, Ningbo, Zhejiang, China
[3]Medical AI ReSearch (MARS) Group, University Engineering Research Center of Digital Medicine and Healthcare, Guangxi Medical University, Nanning, Guangxi, China
[4]Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong
[5]School of Computer Science, University of Nottingham, Nottingham, United Kingdom
scxgw1@nottingham.edu.cn, yuexiang.li@sr.gxmu.edu.cn, wentichen7-c@my.cityu.edu.hk, dingmeidan2023@email.szu.edu.cn, wooi-ping.cheah@nottingham.edu.cn, rong.qu@nottingham.ac.uk, jianfeng.ren@nottingham.edu.cn, llshen@szu.edu.cn

Abstract

Small lesions play a critical role in early disease diagnosis and intervention of severe infections. Popular models often face challenges in segmenting small lesions, as it occupies only a minor portion of an image, while down-sampling operations may inevitably lose focus on local features of small lesions. To tackle the challenges, we propose a Small-Size-Sensitive Mamba (S³-Mamba), which promotes the sensitivity to small lesions across three dimensions: channel, spatial, and training strategy. Specifically, an Enhanced Visual State Space block is designed to focus on small lesions through multiple residual connections to preserve local features, and selectively amplify important details while suppressing irrelevant ones through channel-wise attention. A Tensor-based Cross-feature Multi-scale Attention is designed to integrate input image features and intermediate-layer features with edge features and exploit the attentive support of features across multiple scales, thereby retaining spatial details of small lesions at various granularities. Finally, we introduce a novel regularized curriculum learning to automatically assess lesion size and sample difficulty, and gradually focus from easy samples to hard ones like small lesions. Extensive experiments on three medical image segmentation datasets show the superiority of our S³-Mamba, especially in segmenting small lesions.

Introduction

As the earliest indicators of severe diseases, small lesions play a pivotal role in clinical outcomes, where timely and precise identification can dramatically improve patient prognosis (Qureshi et al. 2023). These lesions, despite their small size, often signal the onset of significant health issues, making their segmentation essential for effective treatment planning (de Teresa-Trueba et al. 2023; Ma and Wang 2023). Many images from different modalities in medical imaging (Greenwald et al. 2022; Luo et al. 2024) contain numerous lesions that occupy less than 10% of the image area (Duering et al. 2023; Ali et al. 2023), *e.g.*, early tumors, micrometastases in lymph nodes, and small vascular abnormalities.

[1]Corresponding Authors: Linlin Shen and Jianfeng Ren.

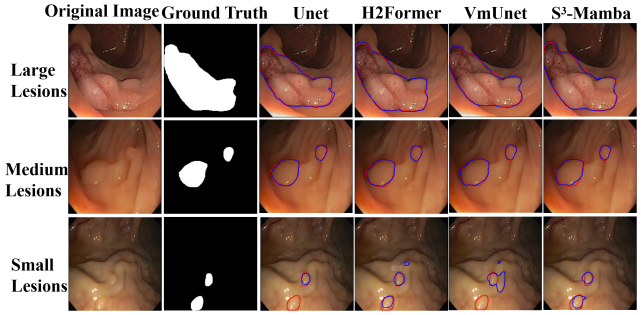

Figure 1: Segmentation results of CNN-based Unet (Ronneberger, Fischer, and Brox 2015), transformere-based H2Former (He et al. 2023), Mamba-based VmUnet (Ruan and Xiang 2024), and the proposed S^3-Mamba for lesions of varying sizes. All models perform well in segmenting large and medium lesions, but S^3-Mamba exhibits superior performance in segmenting small lesions, capturing fine details and contours of small lesions with higher accuracy.

Existing models for medical image segmentation such as U-Net (Ronneberger, Fischer, and Brox 2015), ViTSeg (Du et al. 2023), and MedT (Valanarasu et al. 2021; Tang et al. 2022) have made significant strides in segmenting large objects in medical images. However, these models often struggle with the segmentation of small lesions. Fig. 1 shows the results of several models for segmenting lesions. It is evident that small lesions are inaccurately segmented due to their subtle textures and morphology. A key challenge arises from the common down-sampling operations in network architectures, which unintentionally eliminates crucial local features associated with small lesions. As these lesions only occupy a small portion of the image, they may be overlooked or inadequately represented in the segmentation output, requiring more advanced techniques to handle small lesions.

To tackle the challenges, we propose a **S**mall-**S**ize-**S**ensitive **Mamba** (**S^3-Mamba**) to effectively and efficiently segment small lesions. In view of the balance performance of Mamba (Gu and Dao 2023) between accuracy and speed,

we choose Mamba as the baseline model. The proposed S^3-Mamba enhances the segmentation of small lesions through three novel techniques: an **En**hanced **V**isual **S**tate **S**pace (**EnVSS**) block to selectively amplify key features, an **T**ensor-based **C**ross-feature **M**ulti-scale **A**ttention (**TCMA**) to integrate and preserve critical spatial details, and a novel curriculum learning strategy to adapt the training process based on lesion size and sample difficulty.

More specifically, the proposed EnVSS block aims to tackle the problem of losing local features of small lesions during down-sampling through learnable channel weights and residual connections. Specifically, we insert the **En**hanced **C**hannel **F**eature Block (EnCFBlock) into the Visual State Space block (Liu et al. 2024), which integrates global context information through squeeze, excitation and scale operations to generate channel statistics and scale channel features, so that contribution of each channel is automatically adjusted and fine-tuned based on its relevance to small lesion features. In addition, two residual connections are applied to preserve important lesion fine details by amplifying crucial features while downplaying less relevant ones. The proposed EnVSS block enhances the model's ability to capture and emphasize the subtle characteristics of small lesions that are often overlooked in existing models.

The TCMA is designed to retain the spatial characteristics of small lesions across multiple granularities, exploiting the attentive support from feature maps of multiple scales to enhance the sensitivity to small lesions. The novel tensor-based attention provides an effective mechanism to not only integrate multi-modality features at multiple scales but also dynamically adjust the attention based on the interplay between small lesion areas and their surrounding background context. Unlike existing methods that often treat each feature scale independently or focus solely on the lesion (Fiaz et al. 2024; Huang et al. 2024), the proposed TCMA uniquely emphasizes the relationship between the lesion and the larger surrounding regions at multiple granularities. This dual focus well preserves the spatial integrity of small lesions while simultaneously using the broader context to ensure that critical small lesions are not overlooked.

Traditional random sampling strategies (Singh et al. 2021) often prioritize the accurate segmentation of larger lesions due to their dominance in the image, while neglecting small lesions. Although some workaround techniques such as data augmentation (Bosquet et al. 2023) have been explored to replicate small lesions to improve their representation, the bias towards larger objects remains a challenge. Inspired by (Wang, Chen, and Zhu 2021a), we propose a novel regularized curriculum learning strategy to dynamically adjust the training process based on lesion size and sample difficulty. More specifically, a Difficulty Measurer is designed to assess the the difficulty of samples, initialize sample weights based on lesion size, and dynamically update them according to loss values during training. A Training Scheduler with regularization constraints is designed to gradually shift the focus from simpler samples to more challenging samples such as small lesions as training converges. This strategy enables the model to prioritize more challenging cases and adaptively focus on small lesion segmentation.

Our main contributions can be summarized as follows. 1) The proposed EnVSSBlock dynamically fine-tunes the contribution of each channel based on its relevance to small lesions through channel-wise attention and effectively preserves crucial lesion fine details through two residual connections. 2) The proposed TCMA employs a novel tensor-based multi-level attention strategy to integrate multi-modality features across scales and dynamically adjust the attention based on the interaction between small lesions and their surroundings, preserving small lesion integrity while leveraging surrounding context to better segment small lesions. 3) The proposed regularized curriculum learning assesses the sample difficulty through a Difficulty Measurer and gradual shifts from easier samples to more challenging ones such as small lesions, through a Training Scheduler.

Related Work

Medical Image Segmentation. Existing models for medical image segmentation can be broadly categorized into four groups based on their architecture: 1) Models based on convolutional neural networks (CNNs)(Ronneberger, Fischer, and Brox 2015; Milletari, Navab, and Ahmadi 2016; Chen et al. 2020; Xu et al. 2024; Zhang et al. 2023; Luo et al. 2023; Li and Shen 2018); 2) Transformer-based models such as (Shaker et al. 2024; Valanarasu et al. 2021; He et al. 2023); 3) Hybrid models combining CNN and Transformer architectures(Cao et al. 2022; Tang et al. 2022; **?**; Gao et al. 2022); and 4) State space models, like VmUNet (Ruan and Xiang 2024; Ma, Li, and Wang 2024; Xing et al. 2024). Although these models have made significant strides, small lesion segmentation remains challenging due to the down-sampling operations that commonly exist in these models, which may overlook small targets. In literature, there are some attempts for small object detection, *e.g.*, increase the input image size to generate high-resolution feature maps (Bosquet et al. 2023; Wang et al. 2024b), feature enhancement (Zhang et al. 2024b), feature pyramids (Gao et al. 2024; Mei et al. 2023) and tuning loss functions (Miao et al. 2023; Chen et al. 2022). These methods offer valuable insights for segmenting small lesions. However the unique characteristics of medical images, *e.g.*, low contrast, complex anatomical backgrounds, and diverse imaging modalities, often complicate segmentation tasks and increase the difficulty of accurately segmenting small lesions, highlighting the urgent need for tailored models to address these challenges.

Very recently, Mamba models (Zhu et al. 2024; Liu et al. 2024) have gained traction, especially for medical image segmentation. U-Mamba (Ma, Li, and Wang 2024) extends the Mamba to the U-Net architecture, aiming to capture long-range dependencies using hybrid CNN-SSM (State Space Model) blocks. VM-UNET (Ruan and Xiang 2024) incorporates the Visual State Space (VSS) block to capture contextual information, and its asymmetric encoder-decoder structure demonstrates the capabilities of SSM for segmentation. These Mamba-based models predominantly emphasize long-range dependencies and contextual information, while sacrificing the accuracy of segmenting small lesions. To bridge this gap, we propose an enhanced VSS block that

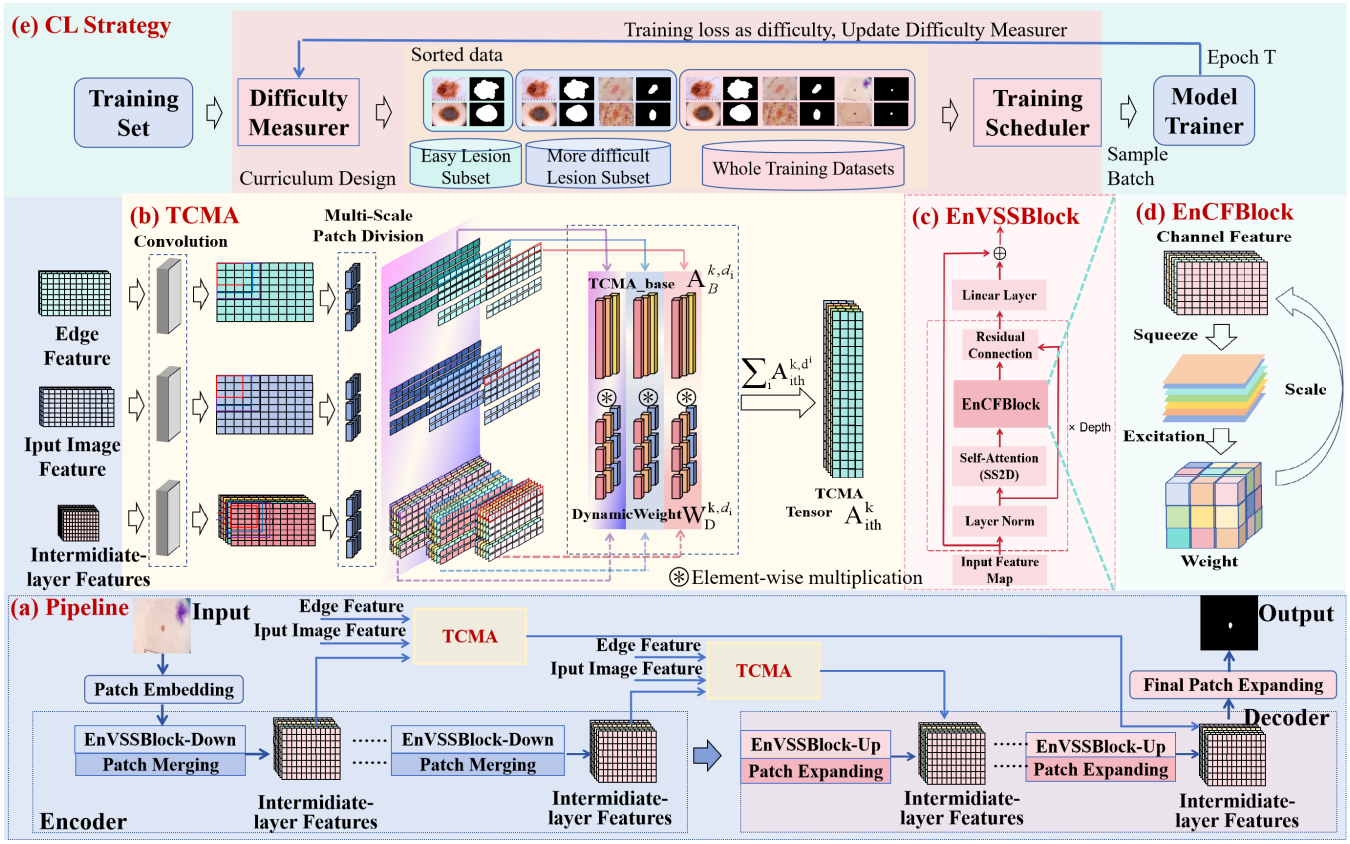

Figure 2: (a) Overview of the proposed S^3-Mamba with *TCMA* and *EnVSSBlock*. (b) Detailed architecture of the TCMA, where input image features, intermediate-layer features, and edge features are divided into patches of three different sizes. A tensor-based attention derives the dynamic weights of these patches of three different scales, exploits their interaction, and utilizes the TCMA features to modulate the features at decoder layers. (c) The detailed structure of the EnVSSBlock, which explicitly evaluates and adaptively adjusts the channel weights to enhance small lesion feature representation and preserves fine details through residual connections. (d) Detailed structure of the Enhanced Channel Feature Block (*EnCFBlock*), which enhances the feature interaction through channel-wise interaction. (e) The architecture of the regularized curriculum learning strategy.

reduces the computational complexity while improving the segmentation accuracy of small lesions.

Curriculum Learning. Curriculum learning (Wang, Chen, and Zhu 2021a) was initially introduced as a training strategy to gradually increase sample complexity during training, mimicking the structured learning of humans and improving model generalization and convergence speed (Wang, Chen, and Zhu 2021b). Recent innovations in Curriculum Learning (CL) include dynamic CL (Wang et al. 2019), which adjusts task difficulty based on training progress to enhance generalization; MTL-based CL (Kong et al. 2021), which organizes tasks hierarchically to improve shared learning across tasks; RL-driven CL (Zhang et al. 2024a), which utilizes reinforcement learning to dynamically adjust the task difficulty for flexible and efficient learning. These methods primarily focus on dynamically adapting the task complexity to facilitate progressive learning. In this paper, a novel curriculum learning strategy with regularization is designed to help focus on more accurately segmenting challenging small lesions. The proposed approach improves upon existing methods by introducing a Difficulty Measurer to evaluate the sample difficulty based on lesion size and model training loss and implementing a Training Scheduler with regularization constraints to balance sample weights and prevent the model from overfocusing on specific samples.

Proposed Method

Overview of Proposed S^3-Mamba

As shown in Fig. 2(a), our S^3-Mamba is built upon a U-shape encoder-decoder architecture, including an Enhanced Visual State Space Block (*EnVSSBlock*) and a Tensor-based Cross-feature Multi-scale Attention (*TCMA*). A patch embedding operation (Liu et al. 2024) is first applied to the original image to generate initial feature maps. The encoder comprises a series of *EnVSSBlocks_Down* and patch merging operations, progressively capturing global dependencies, extracting small lesion features, and compressing spatial dimensions. The TCMA makes use of edge features, input image features, and intermediate-layer features, and divides them into patches of three distinct sizes, aiming to enrich

the diversity of extracted features and focus on small lesions. Then, novel tensor-based attention exploits the attentive supports from these multi-modal mult-scale patch features and merges them to produce the final TCMA features, which are applied to tune the features from decoder layers. The decoder consists of a series of *EnVSSBlocks_Up* and patch-expanding operations that progressively reconstruct the spatial dimensions while enhancing small lesion features. To further focus on small lesions, we introduce a regularized curriculum learning strategy, including a Difficulty Measurer to assess sample difficulty and a Training Scheduler to focus on segmenting challenge targets like small lesions.

Enhanced Visual State Space Block

VSSBlock is the core module of VMamba (Liu et al. 2024), capable of effectively capturing rich contextual information through deep convolution and *SS2D*, making it perform well in segmentation tasks. However, the *VSSBlock* has limitations in retaining fine details due to the local aggregation of deep convolution and the indiscriminate handling of channel features, causing these fine features to be excessively smoothed or overshadowed by more prominent background information. To tackle these challenges, we propose the *EnVSSBlock* to better preserve image fine details.

The improvements over *VSSBlock* are three-fold. 1) We remove the *DWConv* layer in *VSSBlock*, thereby avoiding excessive smoothing and reducing propagation loss of information, enabling the model to retain more detailed features. 2) We introduce two residual connections as shown in Fig. 2(c), one from the input directly to the output that helps maintain the integrity of the initial input features, preventing the information loss in a deep network, and the other highlighting the important features to small lesions while preserving local details.

3) An *EnCFBlock* is added to enhance channel features through adaptive reweighting, suppressing irrelevant background context, and highlighting the key features to more accurately represent small lesions. These improvements collectively help the model focus on segmenting small lesions by preserving local details and adaptively highlighting the features relevant to small lesions.

The *EnCFBlock* exploits the channel-wise attention as in (Hu, Shen, and Sun 2018). Specifically, a global average pooling is first applied to capture the global context of each channel; The squeezed features are then passed through two fully connected layers to further reduce the dimensionality and then restore it, followed by a Sigmoid activation function; These adaptive weights are then applied to the previous channel features, enhancing important features while suppressing less relevant ones; Finally, a residual connection is employed to to preserve locality. Hence, the *EnCFBlock* enhances the feature interaction through channel-wise attention to highlight local features relevant to small lesions.

Tensor-based Cross-feature Multi-scale Attention

Existing attention mechanisms such as Soft Attention (Omeroglu et al. 2023), Hard Attention (Jegham et al. 2023), Self-Attention (Cao et al. 2023), and Multi-Head Attention (Wang et al. 2024a) have demonstrated their power in various tasks, but they exhibit significant limitations when applied to small lesion segmentation, as they are not originally designed to effectively handle multi-scale information, nor can they fully integrate features of different types. What's more, attention weights are often fixed after computation (Luvembe et al. 2023), nor adaptive to varying contexts.

To tackle the challenges, we propose a Tensor-based Cross-feature Multi-scale Attention (TCMA) to enhance the model's ability to segment small lesions in three aspects. 1) The TCMA dynamically tunes the feature representations based on multi-feature multi-scale tensors through a novel tensor-based attention mechanism. The tensors encapsulate the input image features, intermediate prediction features, and edge features at multiple scales. The joint utilization of these features could help exploit the information relevant to segment small lesions in a wider spatial context and a broader range of modalities. 2) The novel tensor-based attention provides an effective mechanism to exploit the attentive supports from features of different modalities at different scales. This unique approach allows for simultaneous interaction between spatial, categorical (class-related), and edge features, enabling the model to dynamically focus on the most relevant features for precisely segmenting small lesions. 3) Finally, the integration of TCMA tensors from low-level features in encoder layers with high-level abstract features in decoder layers provides an effective mechanism to exploit both sets of features, which captures the key features of small lesions in both global context and fine details in local neighborhoods. As a result, the proposed TCMA leads to more precise segmentation of small lesions.

Generation of Multi-Feature Multi-Scale Tensors. We first construct a set of pyramid features, including the input image features embedded using a 7×7 convolutional layer and a 3×3 convolutional layer, the edge features extracted using a Sobel operator and embedded with a 3×3 convolutional layer, and intermediate prediction features. More specifically, Given the i-th level feature maps $\mathsf{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, a 1×1 convolutional layer $\mathcal{F}_{1\times1}$ with N_c filters converts F_i into per-category prediction features as $\mathsf{R}_i = \mathcal{F}_{1\times1}(\mathsf{F}_i)$, where $\mathsf{R}_i \in \mathbb{R}^{H_i \times W_i \times N_c}$ and N_c is the number of categories to predict. We hypothesize that $\mathsf{R}_i^k \in \mathbb{R}^{H_i \times W_i \times 1}$ (for $1 \leq k \leq N_c$) contains only the information associated with the k-th category. We then divide the feature pyramid $\mathsf{R} \in \mathbb{R}^{H \times W \times C}$ into patches of three different scales as,

$$\mathsf{P}^{d_i} = \mathcal{F}_\pi(\mathsf{R}, d_i), \tag{1}$$

where $\mathcal{F}_\pi$ is the image partition operation, d_i represents the size of patches, and $\mathsf{P}^{d_i} \in \mathbb{R}^{\frac{H_i W_i}{d_i^2} \times d_i^2 \times N_c}$. This multi-feature multi-scale tensor provides a detailed and comprehensive feature representation that enhances the model's ability to accurately identify and segment small lesions.

Tensor-based Attention. While the tensor P^{d_i} encapsulates multi-feature multi-scale information, it is difficult to exploit the attentive support through the interaction among the tensor features. To achieve this, a novel tensor-based attention is designed to dynamically tune the weights of the tensor features. Specifically, an `einsum` operation is firstly adopted

to exploit the relations between edge features, input image features, and the k-th category-related features as,

$$\mathsf{A}_{\mathrm{B}}^{k,d_i} = \sum_{j=1}^{d_i^2} \mathsf{P}_{\mathrm{O}}^{d_i}(b,i,j,c) \cdot \mathsf{P}_{\mathrm{I}}^{k,d_i}(b,i,j,k) \cdot \mathsf{P}_{\mathrm{E}}^{d_i}(b,i,j,e), \tag{2}$$

where $\mathsf{P}_{\mathrm{O}}^{d_i}(b,i,j,c)$ represents the tensor value derived from the original input image features for the b-th sample in the batch, the i-th patch, the j-th pixel and the c-th channel; $\mathsf{P}_{\mathrm{I}}^{k,d_i}(b,i,j,k)$ represents the tensor value related to the k-th category from the intermediate prediction features; $\mathsf{P}_{\mathrm{E}}^{d_i}(b,i,j,e)$ represents the tensor value of edge features for the e-th channel. By summing over all pixel positions within the patch, we aggregate the information from all pixels of a patch into the TCMA base tensor $\mathsf{A}_{\mathrm{B}}^{k,d_i}$.

The element-wise multiplication between $\mathsf{P}_{\mathrm{O}}^{d_i}(b,i,j,c)$, $\mathsf{P}_{\mathrm{I}}^{k,d_i}(b,i,j,k)$, and $\mathsf{P}_{\mathrm{E}}^{d_i}(b,i,j,e)$ encapsulates the combined influence of the spatial, categorical (class-related), and edge features, which are crucial for small lesion segmentation. In particular, the spatial features provide location-based context and the general appearance of the lesion, the categorical features ensure the model focuses on the relevant class (*e.g.*, small lesions), and the edge features help to accurately delineate the boundaries of the lesion, which is often challenging due to blurring or low contrast. By combining all of them, the model generates a more robust and accurate representation $\mathsf{A}_{\mathrm{B}}^{k,d_i}$. $\mathsf{P}_{\mathrm{O}}^{d_i}$, $\mathsf{P}_{\mathrm{I}}^{k,d_i}$, and $\mathsf{P}_{\mathrm{E}}^{d_i}$ are then concatenated along the last dimension to form a comprehensive tensor $\mathsf{A}_{\mathrm{C}}^{k,d_i} \in \mathbb{R}^{B\times P\times d_i^2\times(C+N_c+E)}$, and processed through an MLP to compute dynamic weights,

$$\mathsf{W}_{\mathrm{D}}^{k,d_i} = \mathcal{F}_{\mathrm{MLP}}(\mathcal{F}_{\mathrm{Flatten}}(\mathsf{A}_{\mathrm{C}}^{k,d_i})) \in \mathbb{R}^{B\times P\times d_i^2\times O}, \tag{3}$$

where $\mathcal{F}_{\mathrm{Flatten}}(\cdot)$ is the flatten operation, $\mathcal{F}_{\mathrm{MLP}}(\cdot)$ represents fully connected layers with batch normalization, ReLU activation, and sigmoid activation functions, and O is the output dimension. The weighted TCMA tensor is then obtained by aggregating over the last dimension of the channel as,

$$\mathsf{A}_{\mathrm{ith}}^{k,d_i} = \sum \mathsf{A}_{\mathrm{B}}^{k,d_i} \cdot \mathsf{W}_{\mathrm{D}}^{k,d_i}. \tag{4}$$

The TCMA tensor $\mathsf{A}_{\mathrm{ith}}^{k}$ is obtained by integrating the tensors from different scales as $\mathsf{A}_{\mathrm{ith}}^{k} = \sum_i \mathsf{A}_{\mathrm{ith}}^{k,d_i}$, ensuring the model's robustness for different sizes of lesions.

The combination of different types of features helps more accurately segment small lesions, which are often subtle and difficult to distinguish from the background. The utilization of multi-scale features allows the model to capture both fine details and broader contextual information, critical for handling lesions of varying sizes. The resulting tensor $\mathsf{A}_{\mathrm{ith}}^{k}$ enables the model to dynamically adjust its focus on the most relevant features, leading to more precise segmentation.

Integration of TCMA Tensors with Decoder Features.

The obtained TCMA tensor $\mathsf{A}_{\mathrm{ith}}^{k}$ is then applied to modulate the output of each *EnVSSLayer_up*, enhancing the segmentation of small objects. Given a predicted mask $\hat{\mathsf{M}} \in \mathbb{R}^{H\times W\times N_c}$ of an *EnVSSLayer_up* layer, it is first resized using fixed-size average pooling to match the dimensions of an intermediate feature map, and reshaped to $\tilde{\mathsf{M}} \in \mathbb{R}^{H_i\times W_i\times N_c}$ to match the dimensions of TCMA tensor. Then, $\mathsf{A}_{\mathrm{ith}}^{k}$ modulates the decoder features as,

$$\hat{\mathsf{M}}_{\mathrm{ith}}^{k} = \mathsf{A}_{\mathrm{ith}}^{k} \otimes \tilde{\mathsf{M}}^{k}, \tag{5}$$

where $\otimes$ denotes the `einsum` operation. This operation integrates the multi-modal multi-scale image fine details conveyed by the TCMA with the global context in the decoder features, refining the focus on small lesions in the network.

Regularized Curriculum Learning Strategy

We further refine the training strategy to fully unlock the model's potential in segmenting small lesions through curriculum learning shown in Fig. 2(e). We establish a Difficulty Measurer based on the size of the lesions. During training, we dynamically update the Difficulty Measurer based on the loss values, assigning higher weights to samples with lower losses and vice versa. As training progresses and loss stabilizes, the Training Scheduler shifts focus to difficult samples with lower weights. Specifically, given a dataset $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ represents features and $\mathbf{y}_i$ denotes labels for each instance i, the model $\mathcal{F}_\phi$, parameterized by ϕ, generates a prediction $\mathcal{F}_\phi(\mathbf{x}_i)$ for each $\mathbf{x}_i$ and computes the corresponding loss $l_i = \mathcal{L}(\mathcal{F}_\phi(\mathbf{x}_i),\mathbf{y}_i)$, where $\mathcal{L}(\cdot)$ denotes the loss function. The primary objective is to minimize the empirical loss across the entire training set,

$$\min_{\phi} \sum_{i=1}^{N} v_i l_i + g(\mathbf{v}), \tag{6}$$

where $\mathbf{v} = [v_1, v_2, \ldots, v_N] \in [0,1]^N$ represents the vector of weights for training samples, determined by the Difficulty Measurer. Different from traditional curriculum learning (Wang, Chen, and Zhu 2021a), we introduce a novel regularization term $g(\mathbf{v})$ to influence the selection of v_i as,

$$g(\mathbf{v}) = \lambda \sum_{i=1}^{N} \left(\frac{1}{\mathcal{F}_{\mathrm{rank}}(l_i)}\right) + (1-\lambda)\sum_{i=1}^{N} v_i^2, \tag{7}$$

where λ is a weighting factor to balance the influence of weight assignment using loss ranking and the strength of the weight squared sum regularization, and $\mathcal{F}_{\mathrm{rank}}(\cdot)$ is a rank function. Both terms serve as regularization constraints, with the former to regularize the weight distribution, ensuring the model focuses on easier samples first and gradually transit to more difficult ones such as small lesions, and the latter to regularize the sample weights v_i to prevent excessively large weights, ensuring a balanced distribution of weights.

Experimental Results

Datasets. The ISIC2018 dataset (Azad et al. 2019) contains 2,694 dermoscopy images specifically designed for lesion segmentation. The CVC-ClinicDB dataset (Jha et al. 2019), a benchmark for colonoscopy image analysis, includes 612 high-resolution colonoscopy images with corresponding polyp annotations, focusing on polyp detection

		MODELS	MIOU			DSC			ACC			SPE			SEN		
			S	M	L	S	M	L	S	M	L	S	M	L	S	M	L
ISIC2018	CNN	UNet (MICCAI'15)	64.36	79.31	83.60	77.78	88.46	91.07	98.16	97.13	92.65	98.30	97.99	96.01	94.37	90.86	88.11
		UNETR++ (TMI'24)	55.72	74.17	82.20	71.56	85.17	90.23	97.44	96.22	92.02	97.55	97.14	95.96	94.46	89.60	86.69
	Trans	MedT (MICCAI'21)	56.19	73.19	81.57	71.93	84.52	89.85	97.52	96.04	91.62	97.67	96.95	94.91	33.00	89.38	87.18
		H2Former (TMI'23)	63.51	77.61	83.26	77.68	87.39	90.87	98.18	96.85	92.50	98.36	97.76	96.03	93.06	90.24	87.73
		ViTSeg (MICCAI'23)	55.12	73.83	80.50	71.07	84.94	89.19	97.51	96.18	91.20	97.79	97.16	95.47	89.76	89.05	85.42
	Hybrid	SwinUnet (ECCV'22)	62.89	79.36	81.38	77.21	88.49	89.74	98.14	97.19	91.67	98.34	98.31	96.14	92.53	89.10	85.62
		SwinUNETR (CVPR'22)	71.36	78.63	82.98	83.29	88.03	90.70	98.73	97.08	92.25	98..93	98.21	94.83	93.08	88.83	88.77
		MedFormer (arXiv'23)	60.78	77.00	84.52	75.60	87.00	91.61	97.91	96.64	92.68	98.01	97.17	93.74	**95.01**	**92.84**	**91.66**
	Mam	VmUnet (MICCAI'24)	64.04	78.90	**84.67**	78.07	88.21	**91.70**	98.21	97.03	**93.07**	98.37	97.76	95.42	93.67	91.75	89.91
		S^3-Mamba (Ours)	**77.13**	**81.36**	83.28	**87.09**	**89.72**	90.87	**99.07**	**97.55**	92.60	**99.34**	**98.81**	**97.01**	91.66	88.39	86.65
CVC-ClinicDB	CNN	UNet (MICCAI'15)	69.17	78.13	79.87	81.78	87.72	88.81	99.02	98.52	96.88	99.40	99.33	99.02	84.80	86.19	84.44
		UNETR++ (TMI'24)	36.07	51.89	55.98	53.02	68.33	71.78	97.00	96.09	92.13	97.84	97.86	96.22	65.43	68.96	68.32
	Trans	MedT (MICCAI'21)	23.56	39.05	47.15	38.13	56.16	64.08	94.67	93.59	90.33	95.50	95.31	95.72	63.42	67.13	58.91
		H2Former (TMI'23)	65.82	81.02	86.61	79.39	89.52	92.82	98.85	98.73	97.93	99.21	99.41	99.11	85.40	88.41	91.12
		ViTSeg (MICCAI'23)	13.48	22.42	30.50	23.76	36.62	49.06	91.57	90.03	87.50	92.66	92.83	90.54	50.72	47.09	41.10
	Hybrid	SwinUnet (ECCV'22)	24.06	39.43	53.89	38.79	56.56	70.04	94.56	93.55	91.90	95.30	95.18	96.59	66.60	68.59	64.61
		SwinUNETR (CVPR'22)	49.89	65.05	71.33	66.57	78.82	83.27	98.07	97.36	95.52	98.70	98.48	98.87	74.22	80.21	76.03
		MedFormer (arXiv'23)	60.89	74.35	80.78	75.69	85.29	89.37	98.61	98.16	96.98	99.01	98.88	98.77	83.66	87.11	86.55
	Mam	VmUnet (MICCAI'24)	58.26	82.24	**87.94**	73.63	90.25	**93.58**	98.37	98.79	98.14	98.66	99.36	99.14	87.73	**91.83**	**92.34**
		S^3-Mamba (Ours)	**75.40**	**83.81**	85.86	**85.97**	**91.19**	92.39	**99.25**	**99.46**	**99.08**	**99.53**	**99.81**	**99.58**	**88.78**	89.61	91.37

Table 1: Performance comparison on ISIC2018 and CVC-ClinicDB datasets. **S**, **M** and **L** represent small, medium, and large lesions respectively. Our S^3-Mamba performs the best for most metrics on most settings, especially for small lesions.

and segmentation. The in-house Lymph dataset consists of 5,344 multi-modal MRI images of prostate cancer lymph nodes, including 2,733 T2-weighted (T2WI) and 2,611 T1-weighted (T1WI) images, annotated by three skilled radiologists with cross-validation to ensure reliable ground truth.

We then analyze lesion sizes across the three datasets. We sort the lesion pixel distributions from smallest to largest in the ISIC2018 and CVC-ClinicDB datasets, dividing them into three groups: the smallest 30% as small lesions, 30%-60% as medium lesions, and 60% and above as large lesions. We randomly select 30% from each group to create three separate sets for model testing, each focusing on small, medium, and large lesions, respectively. The remaining 70% samples are used for model training. The Lymph dataset predominantly consists of very small and uniformly distributed objects. We utilize 70% for training and 30% for testing without further subdivision.

Compared Methods. Nine state-of-the-art segmentation models are compared, falling into four groups based on their architecture. 1) Traditional CNN-based models such as UNet (Ronneberger, Fischer, and Brox 2015); 2) Attention-based models, including UNETR++ (Shaker et al. 2024), MedT (Valanarasu et al. 2021), H2Former (He et al. 2023), and ViTSeg (Du et al. 2023); 3) Hybrid models combining CNN and Transformer architectures, *e.g.*, SwinUNet (Cao et al. 2022), SwinUNETR (Tang et al. 2022), and MedFormer (Gao et al. 2022); and 4) State space models such as VmUNet (Ruan and Xiang 2024).

Implementation Details. The input image size is 256×256 pixels. Common data augmentation techniques (Garcea et al. 2023) are applied to boost the performance. The backbone is initialized using VMamba-S pre-trained on ImageNet-1k. The AdamW optimizer is used with an initial learning rate of 0.0001 and a cosine scheduler. The batch size is 16. The maximum number of epochs is 600. Experiments are conducted on a Tesla V100 GPU with 32GB memory.

Evaluation Metrics. Five evaluation metrics are adopted: Mean Intersection over Union (mIoU), Dice Similarity Coefficient (DSC), Accuracy (ACC), Specificity (SPE), and Sensitivity (SEN), same as in (Hirling et al. 2024).

	MODELS	MIOU	DSC	ACC	SPE	SEN
CNN	UNet (MICCAI'15)	48.56	65.37	99.91	99.98	56.88
Trans	UNETR++ (TMI'24)	48.68	65.48	99.92	99.98	57.33
	MedT (MICCAI'21)	23.77	38.41	99.86	99.95	32.68
	H2Former (TMI'23)	47.52	64.42	99.92	99.98	54.88
	ViTSeg (MICCAI'23)	14.94	26.00	99.79	99.89	27.12
Hybrid	SwinUnet (ECCV'22)	19.64	32.83	99.85	99.95	26.51
	SwinUNETR (CVPR'22)	45.03	62.09	99.91	99.97	54.02
	MedFormer (arXiv'23)	55.25	71.18	99.92	99.97	68.78
Mam	VmUnet (MICCAI'24)	52.27	68.66	99.92	99.97	62.44
	S^3-Mamba (Ours)	**61.19**	**75.93**	**99.94**	**99.97**	**75.18**

Table 2: Performance comparison on the Lymph dataset.

Comparison with State-of-the-Art Methods

ISIC2018 Dataset. The comparison results on the ISIC2018 dataset are summarized in Table 1. The following can be observed. 1) Our S^3-Mamba achieves the highest mIoU of

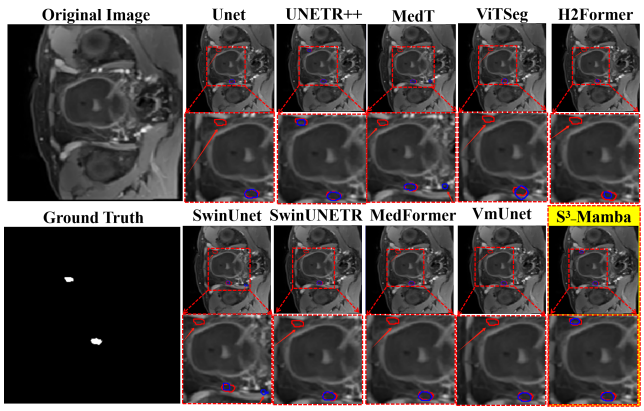

Figure 3: Segmentation results on the Lymph dataset. Red outlines represent the ground truth segmentation masks, and the blue marks indicate the model predictions.

77.13% and 81.36% for small and medium lesions respectively, and a competitive mIoU of 83.28% for large lesions. This robustness across lesion sizes is attributed to the EnVSS block's dynamic fine-tuning of channel contributions and the TCMA's multi-level attention strategy, both of which enhance the model's ability to preserve small lesion details while leveraging the surrounding context. 2) The VmUnet (Ruan and Xiang 2024), a predecessor of the Mamba series, excels in large lesion segmentation with a mIoU of 84.67% and a DSC of 91.70%, but it is less effective with small and medium lesions. This suggests that the VmUnet prioritizes broader spatial context, potentially at the cost of fine-grained details necessary for smaller lesions. The enhancements in our S^3-Mamba, particularly in attention mechanisms, likely contribute to its improved performance across different lesion sizes. 3) UNet (Ronneberger, Fischer, and Brox 2015) achieves an mIoU of 64.36%, 79.31%, 83.60% for small, medium and large lesions respectively. While it is a foundational model in medical image segmentation, it lags behind newer models, particularly on small lesions, highlighting the need for ongoing innovation in segmentation techniques. 4) Transformer-based models like MedT achieves an mIoU of 56.19%, 73.19%, and 81.57% and H2Former achieves an mIoU of 63.51%, 77.61%, and 83.26% respectively for small, medium and large lesions. They exhibit varied performance, with H2Former particularly strong in small lesion segmentation. This suggests that transformers are highly dependent on their specific configuration and integration.

CVC-ClinicDB Dataset. The following can be observed on the CVC-ClinicDB dataset. 1) Some models such as UNETR++ (Shaker et al. 2024), MedT (Valanarasu et al. 2021) and ViTSeg (Du et al. 2023) experience a significant performance drop on this dataset compared to their performance on the ISIC2018 dataset, possibly due to the generally small lesions in this dataset, with a lesion pixel ratio ranging from 0.44% to 50.17% only. In addition, the ISIC2018 dataset utilizes high-resolution dermoscopic images, while the CVC-ClinicDB dataset consists of colonoscopic images where the visual characteristics of lesions differ. 2) In contrast, our S^3-Mamba continues to excel on the CVC-ClinicDB dataset, because of its optimization for small lesion segmentation, *e.g.*, the EnVSS block dynamically adjusts channel contributions, and the multi-level tensor-based attention integrates features across scales, preserving lesion integrity while leveraging the background information. 3) While the proposed S^3-Mamba exhibits strong overall performance, its sensitivity (SEN) is not the highest across both datasets. The slightly lower sensitivity could be due to the model's balanced focus on both sensitivity and specificity.

Methods	MIOU	DSC	ACC	SPE	SEN
Baseline	52.27	68.66	99.92	99.97	62.44
+EnCF	55.70	71.55	99.93	99.97	67.06
+TCMA	55.66	71.51	99.93	99.97	68.67
+CL	55.15	71.09	99.92	99.97	67.82
+TCMA+EnCF	57.14	72.73	99.93	99.97	69.99
+EnCF+CL	57.71	73.19	99.93	99.96	72.82
+TCMA+CL	59.62	74.70	99.93	99.97	**77.60**
S^3-Mamba	**61.19**	**75.93**	**99.94**	**99.97**	75.18

Table 3: Ablation of key components on the Lymph dataset.

Private Lymph Dataset. The comparison results for ultra-small lesions on the private prostate lymph dataset are summarized in Table 2. Our S^3-Mamba outperforms all the models, achieving an mIoU of 61.19%, DSC of 75.93%, accuracy of 99.94%, and sensitivity of 75.18%, respectively. While the performance of all models decreases on the Lymph dataset compared to that on the ISIC2018 and CVC-ClinicDB datasets, our S^3-Mamba still leads by about 5% in several metrics. Fig. 3 shows sample segmentation results, particularly focusing on small lesions. S^3-Mamba demonstrates a closer alignment with the ground truth and produces fewer false positives compared to other models, as evident in the zoomed-in sections of the image. These observations demonstrate that S^3-Mamba's design innovations significantly contribute to its leading performance on the Lymph dataset, particularly in small lesion segmentation.

Ablation Study of Key Components

We ablate the key components of the proposed S^3-Mamba model using the Lymph and ISIC2018 datasets. The baseline model is trained without the EnCFBlock, TCMA, or curriculum learning (CL). We then assess the impact of adding these components individually and in combination. Specifically, we evaluate the baseline with EnCFBlock (+EnCF), TCMA (+TCMA), and CL (+CL), as well as their combinations, such as EnCFBlock + TCMA (+TCMA+EnCF), EnCFBlock + CL (+EnCF+CL), and TCMA + CL (+TCMA+CL). The ablation results on the Lymph dataset and ISIC2018 dataset are summarized in Table 3 and Table 4 respectively.

As shown in Table 3, the baseline achieves an mIoU of 52.27%. Adding EnCFBlock increases the mIoU to 55.70%, while TCMA results in a similar improvement (55.66%). Combining TCMA and EnCFBlock (+TCMA+EnCF) further increases the mIoU to 57.14%, indicating a synergistic effect between these two components. When curriculum

Methods	Size	MIOU	DSC	ACC	SPE	SEN
Baseline	S	64.04	78.07	98.21	98.37	93.67
	M	78.90	88.21	97.03	97.76	91.75
	L	84.67	91.70	93.07	95.42	89.91
+EnCF	S	68.51	81.32	98.54	98.74	93.04
	M	82.14	90.19	97.61	98.55	90.79
	L	84.71	91.72	93.05	94.88	90.57
+TCMA	S	67.97	80.93	98.49	98.65	94.02
	M	81.70	89.93	97.70	98.26	92.03
	L	84.62	91.67	92.98	94.63	90.76
+CL	S	66.73	80.05	98.43	98.64	92.44
	M	81.47	89.79	97.52	98.57	89.93
	L	85.60	92.24	93.51	95.64	90.64
+TCMA+EnCF	S	74.89	85.64	98.95	98.18	92.33
	M	81.71	89.94	97.56	98.58	90.13
	L	85.41	92.13	93.46	96.00	90.02
+EnCF+CL	S	70.02	82.36	98.63	98.79	94.04
	M	81.74	89.95	97.51	98.27	92.02
	L	85.43	92.14	93.40	95.16	91.02
+TCMA+CL	S	72.23	83.87	98.76	98.89	94.96
	M	80.58	89.25	97.31	98.00	92.26
	L	83.75	91.15	92.68	95.67	88.65
S^3-Mamba	S	77.13	87.09	99.07	99.34	91.66
	M	81.36	89.72	97.55	98.81	88.39
	L	83.28	90.87	92.60	97.01	86.65

Table 4: Ablation results on the ISIC2018 dataset for Small (S), Medium (M) and Large (L) lesions.

learning (CL) is incorporated (+CL), the mIoU improves slightly to 55.15%, suggesting its role in helping the model focus on small lesions. The combination of TCMA and CL (+TCMA+CL) further improves the mIoU to 59.62%. Finally, the complete S^3-Mamba model, achieves the highest performance across all metrics, with an mIoU of 61.19%.

The ablation results on the ISIC2018 dataset with lesions of different sizes (Small (S), Medium (M), and Large (L)) are shown in Table 4. The baseline model performs relatively poorly on small lesions (S), with an mIoU of only 64.04%. EnCFBlock improves this to 68.51%, while TCMA and CL achieve an mIoU of 67.97% and 66.73%, respectively. The combination of TCMA and EnCFBlock (+TCMA+EnCF) significantly improves the performance to 74.89%. The full model achieves the best result for small lesions with an mIoU of 77.13%, demonstrating that integrating all components enhances the segmentation accuracy and robustness.

Model Complexity Analysis

Fig. 4 presents a scatter plot comparing various segmentation models based on FLOPs (Floating Point Operations per Second), DSC (Dice Similarity Coefficient), and Model Size (indicated by circle size). S^3-Mamba stands out as the most balanced model, achieving the highest DSC of 75.93% with a moderate computational demand of 27.58G FLOPs and a compact model size of 4.64M parameters. It outperforms other models like UNETR++ (DSC 65.48%, 87.78G FLOPs) and ViTSeg (DSC 26.0%, 91.88G FLOPs), which are both computationally expensive and less accurate. While VmUnet also performs well, it requires more resources, *e.g.*, 39.77G FLOPs and 6.4M parameters, while S^3-Mamba's requires 27.58G FLOPs and 4.64M parameters. Both VmUnet and S^3-Mamba are based on the Mamba framework, showcasing the effectiveness of this architecture. Finally, our S^3-Mamba achieves a better segmentation accuracy of 75.93% than 68.66% for VmUnet while being more computationally efficient, making it the preferred choice for applications where both performance and efficiency are critical.

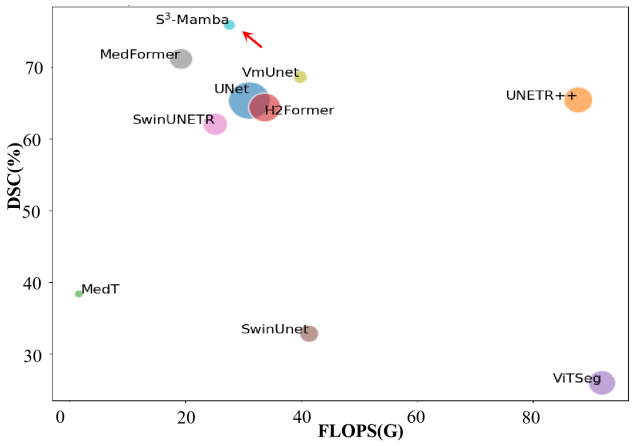

Figure 4: Comparison with other models in terms of FLOPs, DSC, and model size represented by the circle size on the Lymph dataset.

Conclusion

In this paper, we propose Small-Size-Sensitive Mamba (S^3-Mamba), which significantly improves small lesion segmentation by introducing the EnVSSBlock, Tensor-based Cross-feature Multi-scale Attention, and a novel regularized curriculum learning strategy. In particular, the EnVSSBlock enhances the feature representation by adding residual connections to preserve local features and adaptively adjusting channel weights to amplify important features, particularly the subtle ones from small lesions through the channel-wise attention. The TCMA captures multi-scale information from patches of different modalities of different scales through a novel tensor-based attention mechanism, combining fine details with global context, thus improving the model's capability in segmenting small lesions. Finally, the novel regularized curriculum learning strategy enables the model to gradually learn features from large to small lesions, enhancing the segmentation accuracy and robustness. The combination of these strategies achieves significant performance gains over state-of-the-art models on three benchmark datasets, significantly enhancing small lesion segmentation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 82261138629 and 12326610; Guangdong Provincial Key Laboratory under Grant 2023B1212060076, and Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030.

References

Ali, S.; Jha, D.; Ghatwary, N.; Realdon, S.; Cannizzaro, R.; Salem, O. E.; Lamarque, D.; Daul, C.; Riegler, M. A.; Anonsen, K. V.; et al. 2023. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10(1): 75.

Azad, R.; Asadi-Aghbolaghi, M.; Fathy, M.; and Escalera, S. 2019. Bi-directional ConvLSTM U-Net with densley connected convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 10–20.

Bosquet, B.; Cores, D.; Seidenari, L.; Brea, V. M.; Mucientes, M.; and Del Bimbo, A. 2023. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognition*, 133: 108998.

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.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Chen, W.; Yu, S.; Ma, K.; Ji, W.; Bian, C.; Chu, C.; Shen, L.; and Zheng, Y. 2022. TW-GAN: Topology and width aware GAN for retinal artery/vein classification. *Medical Image Analysis*, 77: 102340.

Chen, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Chu, C.; Shen, L.; and Zheng, Y. 2020. TR-GAN: Topology ranking GAN with triplet loss for retinal artery/vein classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, 616–625. Springer.

de Teresa-Trueba, I.; Goetz, S. K.; Mattausch, A.; Stojanovska, F.; Zimmerli, C. E.; Toro-Nahuelpan, M.; Cheng, D. W.; Tollervey, F.; Pape, C.; Beck, M.; et al. 2023. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nature Methods*, 20(2): 284–294.

Du, S.; Bayasi, N.; Hamarneh, G.; and Garbi, R. 2023. AViT: Adapting Vision Transformers for Small Skin Lesion Segmentation Datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 25–36. Springer.

Duering, M.; Biessels, G. J.; Brodtmann, A.; Chen, C.; Cordonnier, C.; de Leeuw, F.-E.; Debette, S.; Frayne, R.; Jouvent, E.; Rost, N. S.; et al. 2023. Neuroimaging standards for research into small vessel disease—advances since 2013. *The Lancet Neurology*, 22(7): 602–618.

Fiaz, M.; Noman, M.; Cholakkal, H.; Anwer, R. M.; Hanna, J.; and Khan, F. S. 2024. Guided-attention and gated-aggregation network for medical image segmentation. *Pattern Recognition*, 110812.

Gao, J.; Geng, X.; Zhang, Y.; Wang, R.; and Shao, K. 2024. Augmented weighted bidirectional feature pyramid network for marine object detection. *Expert Systems with Applications*, 237: 121688.

Gao, Y.; Zhou, M.; Liu, D.; Yan, Z.; Zhang, S.; and Metaxas, D. N. 2022. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*.

Garcea, F.; Serra, A.; Lamberti, F.; and Morra, L. 2023. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152: 106391.

Greenwald, N. F.; Miller, G.; Moen, E.; Kong, A.; Kagel, A.; Dougherty, T.; Fullaway, C. C.; McIntosh, B. J.; Leow, K. X.; Schwartz, M. S.; et al. 2022. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, 40(4): 555–565.

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

He, A.; Wang, K.; Li, T.; Du, C.; Xia, S.; and Fu, H. 2023. H2Former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9): 2763–2775.

Hirling, D.; Tasnadi, E.; Caicedo, J.; Caroprese, M. V.; Sjögren, R.; Aubreville, M.; Koos, K.; and Horvath, P. 2024. Segmentation metric misinterpretations in bioimage analysis. *Nature Methods*, 21(2): 213–216.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, H.; Chen, Z.; Zou, Y.; Lu, M.; Chen, C.; Song, Y.; Zhang, H.; and Yan, F. 2024. Channel prior convolutional attention for medical image segmentation. *Computers in Biology and Medicine*, 178: 108784.

Jegham, I.; Alouani, I.; Khalifa, A. B.; and Mahjoub, M. A. 2023. Deep learning-based hard spatial attention for driver in-vehicle action monitoring. *Expert Systems with Applications*, 219: 119629.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Johansen, D.; De Lange, T.; Halvorsen, P.; and Johansen, H. D. 2019. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–2255. IEEE.

Kong, Y.; Liu, L.; Wang, J.; and Tao, D. 2021. Adaptive curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5067–5076.

Li, Y.; and Shen, L. 2018. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2): 556.

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

Luo, X.; Liu, C.; Wong, W.; Wen, J.; Jin, X.; and Xu, Y. 2023. MVCINN: multi-view diabetic retinopathy detection using a deep cross-interaction neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8993–9001.

Luo, X.; Xu, Q.; Wang, Z.; Huang, C.; Liu, C.; Jin, X.; and Zhang, J. 2024. A Lesion-Fusion Neural Network for Multi-View Diabetic Retinopathy Grading. *IEEE Journal of Biomedical and Health Informatics*.

Luvembe, A. M.; Li, W.; Li, S.; Liu, F.; and Xu, G. 2023. Dual emotion based fake news detection: A deep attention-weight update approach. *Information Processing & Management*, 60(4): 103354.

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

Ma, J.; and Wang, B. 2023. Towards foundation models of biological image segmentation. *Nature Methods*, 20(7): 953–955.

Mei, Y.; Fan, Y.; Zhang, Y.; Yu, J.; Zhou, Y.; Liu, D.; Fu, Y.; Huang, T. S.; and Shi, H. 2023. Pyramid attention network for image restoration. *International Journal of Computer Vision*, 131(12): 3207–3225.

Miao, J.; Zhou, S.-P.; Zhou, G.-Q.; Wang, K.-N.; Yang, M.; Zhou, S.; and Chen, Y. 2023. SC-SSL: Self-correcting Collaborative and Contrastive Co-training Model for Semi-Supervised Medical Image Segmentation. *IEEE Transactions on Medical Imaging*.

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

Omeroglu, A. N.; Mohammed, H. M.; Oral, E. A.; and Aydin, S. 2023. A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Engineering Applications of Artificial Intelligence*, 120: 105897.

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.

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.

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

Shaker, A. M.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2024. UNETR++: delving into efficient and accurate 3D medical image segmentation. *IEEE Transactions on Medical Imaging*.

Singh, D. K.; Rai, S. N.; Joseph, K.; Saluja, R.; Balasubramanian, V. N.; Arora, C.; Subramanian, A.; and Jawahar, C. 2021. Order: Open world object detection on road scenes. In *Proc. NeurIPS Workshops*, volume 1, 3.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740.

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.

Wang, J.; Lai, C.; Wang, Y.; and Zhang, W. 2024a. EMAT: Efficient feature fusion network for visual tracking via optimized multi-head attention. *Neural Networks*, 172: 106110.

Wang, P.; Ma, Z.; Dong, B.; Liu, X.; Ding, J.; Sun, K.; and Chen, Y. 2024b. Generative data augmentation by conditional inpainting for multi-class object detection in infrared images. *Pattern Recognition*, 153: 110501.

Wang, X.; Chen, Y.; and Zhu, W. 2021a. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4555–4576.

Wang, X.; Chen, Y.; and Zhu, W. 2021b. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4555–4576.

Wang, Y.; Gan, W.; Yang, J.; Wu, W.; and Yan, J. 2019. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5017–5026.

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

Xu, Q.; Luo, X.; Huang, C.; Liu, C.; Wen, J.; Wang, J.; and Xu, Y. 2024. HACDR-Net: Heterogeneous-Aware Convolutional Network for Diabetic Retinopathy Multi-Lesion Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6342–6350.

Zhang, J.; Ren, J.; Zhang, Q.; Liu, J.; and Jiang, X. 2023. Spatial Context-Aware Object-Attentional Network for Multi-Label Image Classification. *IEEE Transactions on Image Processing*, 32: 3000–3012.

Zhang, J.; Yang, X.; He, W.; Ren, J.; Zhang, Q.; Zhao, Y.; Bai, R.; He, X.; and Liu, J. 2024a. Scale Optimization Using Evolutionary Reinforcement Learning for Object Detection on Drone Imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 410–418.

Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; and Yan, J. 2024b. FFCA-YOLO for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.

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*.