An Effective Algorithm for Skin Disease Segmentation Combining inter-channel Features and Spatial Feature Enhancement

ZunWang Ke
School of Software, Xinjiang University
Urumqi 830091, China
kzwang@xju.edu.cn

YinFeng Wang
School of Software, Xinjiang University
Urumqi 830091, China
wangyinfeng1999@163.com

Run Guo
Department of Dermatology, Guang'anmen Hospital China Academy of Chinese Medical Sciences
No. 5 Beixiange, Xicheng District, Beijing 100053, China
grlww25@sina.com

Minghua Du
Department of Emergency, the First Medical Center, Chinese PLA General Hospital
forrestdo@163.com

Ji-Sheng Zhou
Peking University Third Hospital
No. 49 Beihuayuan Road, Haidian District, Beijing 100191, China
piova@foxmail.com

Gang Wang
School of Computing and Data Engineering, NingboTech University
Ningbo 315100, China
wanggangnit@nit.zju.edu.cn

Yugui Zhang(✉)
Institute of Semiconductors, Chinese Academy of Sciences
Beijing 100083, China
zhangyugui@semi.ac.cn

Abstract

Skin lesion segmentation is essential for early disease detection and treatment planning in computer-aided diagnostic systems. However, U-Net faces challenges in handling long-distance dependencies and fully utilizing semantic information. Additionally, feature redundancy in channels and asymmetric supervised learning can lead to irrelevant features, resulting in suboptimal segmentation accuracy. To tackle these challenges, this paper presents a dermatological segmentation method that improves inter-channel and spatial features. The method introduces a compression excitation module and a channel mixing network, boosting both feature extraction capabilities and channel information exchange. Furthermore, the integration of a cross-region attention mechanism enhances the modeling of long-distance dependencies and spatial feature perception. The proposed approach also integrates a feature distillation loss function, which facilitates a balanced supervision mechanism between the encoder and decoder. This effectively minimizes redundant information within the U-Net architecture. Experiments conducted on the publicly available skin lesion datasets ISIC2016, ISIC2017, and ISIC2018 demonstrate that the proposed approach attains substantial performance enhancements in skin lesion image segmentation, showcasing its strong competitiveness.

Keywords: skin lesion segmentation,inter-channel fea-

ture enhancement,spatial feature enhancement,feature distillation loss function

1. Introduction

Skin diseases, such as skin cancer, psoriasis, and eczema, are prevalent health issues globally, significantly affecting patients' quality of life. According to statistics, millions of new cases are diagnosed each year, and this number continues to rise[1]. Early identification and treatment of skin disorders is critical to optimizing patient outcomes, yet the process presents numerous challenges. Traditional diagnostic methods rely on physicians' experience and subjective judgment, which may lead to misdiagnosis or missed diagnosis. Advancements in medical imaging technology have made image analysis a key tool in aiding skin disease diagnosis. Especially in images of skin lesions, accurate segmentation of the lesion area is important for quantifying lesion characteristics, evaluating treatment effects and formulating personalized treatment plans. However, due to the diversity of morphology and texture of skin lesions, traditional image segmentation methods are often difficult to achieve satisfactory accuracy[9, 8].

Recently, deep learning methods, particularly convolutional neural networks(CNNs), have accomplished significant advancements and demonstrated impressive results in medical image segmentation[33, 12, 45]. U-Net[36] is a notable instance of the CNN-based model, renowned for its uncomplicated architecture and scalability. A multitude of subsequent enhancements have been derived based on this U-shaped framework[10, 51, 38, 39, 40]. TransUnet[7], a pioneer in Transformer-based modeling, was the first to use a visual Transformer (ViT) in the encoding phase[13] to extract features and utilized CNNs in the decoding stage, showing its significant advantage in capturing global information. Subsequently, TransFuse[50] utilized a parallel architecture combining ViT and CNN to simultaneously capture both local and global features. In addition, SWin-UNet[5] integrates Swin Transformer[28] and U-shape architecture and proposes a U-shape model completely based on Transformer for the first time.

Models based on CNNs often confront issues in capturing long-range dependencies within images effectively, largely because their receptive fields are limited to local regions. This constrains their capacity to handle spatial features and inter-channel information. Moreover, while the U-Net structure excels in applications like medical image segmentation, it can suffer from issues such as feature redundancy and unbalanced supervision, which may negatively impact segmentation accuracy.

To solve these problems, this paper presents a dermatological segmentation strategy that combines inter-channel and spatial feature enhancement, incorporating various techniques optimized for medical image segmentation to showcase its potential in such tasks. Specifically, the method proposes an inter-channel feature enhancement module(CM-SSM) that combines compressive excitation and channel mixing, and a spatial feature enhancement module (CRA) based on cross-regional attention. The CM-SSM strengthens the network's feature extraction, channel information exchange, and local feature capture by learning the channel weights of the input feature map; CRA effectively combines regional information with channel attention, enhancing the network's capability to model long-range dependencies. Furthermore, this paper integrates Binary Cross-Entropy and Dice loss functions by incorporating cross-feature and internal feature distillation losses[44]. This approach aids in balancing supervision between the U-Net's encoder and decoder, while also reducing redundant information within the model.

In this study, a series of comprehensive experiments were carried out on segmentation tasks involving skin lesions, aiming to demonstrate the effectiveness of the proposed method in the field of medical image segmentation. Specifically, this study performed thorough evaluations on the standard datasets ISIC2016[21], ISIC2017[4], and ISIC2018[11]. The experimental results highlight the method's capability to achieve outstanding results.

The remainder of the paper is structured as follows: Section 2 summarizes the relevant research. Section 3 outlines the detailed implementation of the proposed method. Section 4 describes the setup and results of the experiment in detail. Section 5 summarizes the main findings of this study and explores possible paths for future research.

2. Related work

2.1. U-Net

Medical image segmentation has consistently posed challenges. In recent years, the application of deep learning technology in this area has been expanded significantly, leading to notable advancements. U-Net[36], as a benchmark network architecture in the field, is known for its encoder-decoder structure, which can efficiently extract and process image features. CE-Net[19] further integrates the contextual information encoding module to enhance the sensory field and semantic representation of the model. UNet++[51] introduces a nested U-Net architecture that enhances segmentation accuracy by integrating multi-scale feature fusion. Besides convolution-based methods, Transformer-based models have also gained significant attention. Vision Transformer[13] demonstrated the effectiveness of Transformer in image recognition tasks. Medical Transformer[43] and TransUNet[7] incorporated the Transformer architecture into the field, resulting in impressive performance. In addition, the attention mechanism[41] and multi-scale feature fusion[23] and other techniques have

also been extensively applied in tasks of the field. 3D segmentation models like multi-gated loop units[2] and efficient multi-scale 3D CNN[26] have also demonstrated notable success. Recently, Mamba[16] achieved a significant breakthrough by incorporating selection mechanisms and hardware-aware algorithms into previous work[17, 20, 32], enabling linear-time inference and an efficient training process. Building on the accomplishments of Mamba for vision applications, Vision Mamba[27] and VMamba[52] utilize bi-directional Vim blocks and cross-scan modules, each serving to capture global visual contexts that depend on data. Meanwhile, VM-UNet[37] and U-Mamba[31] demonstrate excellent performance in medical image segmentation.

2.2. State Space Modeling

The State Space Model (SSM) is a mathematical framework commonly employed in control theory and signal processing to represent the dynamic behavior of a system's state as it evolves. By defining a set of state variables and their alteration rules, SSM is capable of capturing the dynamic behavior of the system. Its computational complexity increases linearly in proportion to the extent of the input sequence, and it is globally perceptive, which renders SSM particularly efficacious in processing sequence data. Mamba Modeling [27]. is an application of SSM to natural language processing and vision tasks, which provides a new solution to visual perception tasks by combining the global information capture capability of SSM with the advantages of deep learning. In 2024, Liu et al. applied SSM to the field of vision and introduced Visual Mamba[27], a model that not only captures global information in images but also manages computational resources and time effectively. The application of SSM in vision tasks has been extended to several domains, including image classification[3, 6], video processing[48], event camera data processing[53, 46], etc. In the domain of medical image analysis, the use of SSM has resulted in substantial breakthroughs. An example is the U-Mamba model[31], which integrates the high efficiency of CNN in local feature extraction and the advantages of SSM in global information capture, thus improving the accuracy of segmentation. To reduce the computational resource consumption of SSM models, researchers have performed a series of optimizations. For example, the S4 model[18] implemented a diagonal architecture and a low-rank approach, while the S5[42] and H3 models[14] enhanced efficiency by utilizing parallel scanning techniques and optimizing hardware use. The S6 version of Mamba[27], on the other hand, optimizes the linear time-invariant features of the model by combining data-dependent parameters and demonstrates superior performance on large-scale datasets. ViM[52] and VMamba[27] and other models further integrated SSM into the visual backbone design and achieved results comparable to ViT and CNN by adapting the characteristics of image data through multiple scanning directions. These research results not only promote the development of SSM-based vision models but also demonstrate the potential of SSM for various applications within the domain of computer vision, creating new avenues for future research and applications.

3. Methods

3.1. Model Structure

The overall architecture of the framework is illustrated in Figure 1. More specifically, it concludes with a patch embedding layer, encoders, decoders, a final projection layer, and skip connections.

In this paper, the last up-sampling block in the model is defined as the last decoder block, and in the symbolic representation, this paper uses $\mathrm{E}_m^{(l)}/\mathrm{D}_m^{(l)}$ to denote the mth encoder/decoder block in the mth l layer. Accordingly, in the order from encoder to decoder, the corresponding feature map is denoted as F_m^l ($\mathrm{E}_1^{(1)}$, $\mathrm{E}_1^{(2)}$,..., $\mathrm{D}_3^{(2)}$, $\mathrm{D}_2^{(1)}$, $\mathrm{D}_2^{(2)}$).

The Patch Embedding layer segments the input image $x \in \mathrm{R}^{H\times W\times 3}$ into small blocks of size $r \times r$ and then maps the dimensions of the image to the number of channels C of default size 96 to generate the embedded output $x' \in R^{\frac{H}{r}\times\frac{W}{r}\times C}$. Then, this paper uses layer normalization[36] to normalize x' and input it into the encoder for extracting features. The framework consists of four encoder blocks, in which patch merging operations are applied in the first three encoder blocks to reduce the input feature height and width while increasing the number of channels. In this paper, CM-SSM blocks [2, 2, 2, 2] are used in four encoder blocks respectively, and the number of channels of each encoder block is [C, 2C, 4C, 8C].

Similarly, the framework contains four decoder blocks. In the final three stages of the decoder, a patch extension strategy is applied to decrease the density of the feature channels while enhancing the vertical and horizontal dimensions of the feature map. In this paper, CM-SSM blocks [2, 2, 2, 1] are used in four decoder blocks respectively, and the channel count for each decoder block is [8C, 4C, 2C, C]. The decoder module is followed by a final mapping layer, whose role is to adjust the feature dimensions back to the dimensions corresponding to the original image.

For the jump connection part, this paper uses the cross-region attention to obtain the location information of regions within the feature map., and then fuses this region information with the feature map, after fusing the region location information, the feature map can better capture the spatial structure of the image as a way to improve the segmentation accuracy.

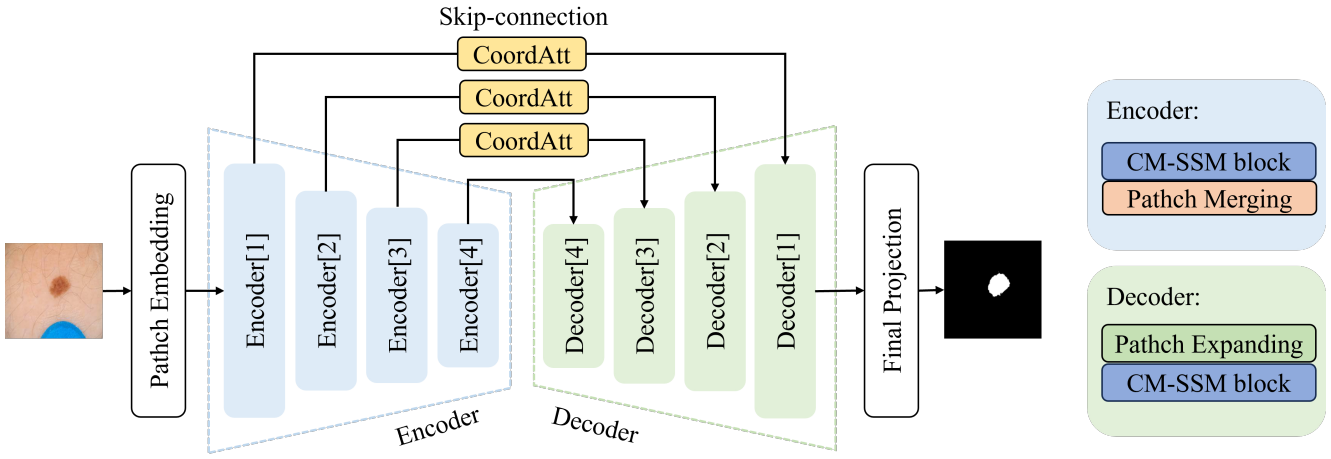

Figure 1. General Framework Diagram.

3.2. Interchannel Feature Enhancement Combining SE and CMN

The architectural framework of the CM-SSM block is shown in Figure 2 and consists of a 2D-Selective-Scan Vision Space State (2D-SSVS) component and a Channel Mixing Network (CMN).

Among them, The SS2D module serves as the central component of the CM-SSM block and consists of three main components: scan expansion, S6 block, and scan merging. The scan expansion technique divides the input image into individual sequences by four-way decomposition, a step that ensures extensive spatial information coverage and enables multi-directional feature capture. The S6 block then selects the parameters of the state-space model using a selectivity mechanism to accurately identify and extract useful information while filtering out irrelevant parts. Specifically, the block receives inputs in feature format [B, L, D], where B represents the batch size, L indicates the sequence length, and D represents the feature dimension. The features are first transformed through a linear layer, followed by the application of update and output equations from the state-space model to generate the final output features. Finally, scanning and merging operations reconfigure these transformed sequences to generate an output image that matches the dimensions of the original input image. Through this series of fine-grained operations, the SS2D module provides the CM-SSM block with powerful feature extraction and processing capabilities, which are particularly important for medical image segmentation.

According to the relevant literature[24, 35] , this paper integrates a Squeeze-Excitation (SE) after a 2D scanning[22] block. The restructured features are processed using the Squeeze-Excitation block to capture the interchannel relationships through global pooling and activation operations, and then use this information to re-weight the channels, enhancing important features and suppressing unimportant ones.

A departure from the traditional focus on token mixing in previous visual Mamba architectures[27, 52] , the design in this paper introduces a channel mixing network that consists of a deep convolution and two fully connected layers. The layers enable the model to exchange information between different channels by extending and recovering the channels, thus increasing exchange between different channels and enhancing the representation of features. Deep convolution helps to capture local dependencies of different features in the channel dimension by applying convolution independently on each channel and then mixing these channel features by convolution, while capturing more complex local features since it operates on the channel after extension.

3.3. Spatial Feature Enhancement Based on CRA

Through the aforementioned method, this paper enhances inter-channel features. Nevertheless, spatial features play a vital role in the realm of segmentation, given that they enable the model to better grasp the contextual information within the image, thereby improving segmentation accuracy. Skip connections are also an important part of the model, and they bridge the information between the encoder and decoder. Therefore, in the present research, we choose to augment the spatial features at the skip connections using the cross-region attention mechanism.

The Cross-region Attention mechanism (CRA) enhances the feature representation capability of deep neural networks by combining spatial region location information. As depicted in Figure 3, the cross-region attention mechanism first generates two feature maps with height-oriented and width-oriented input feature maps through two global pool-

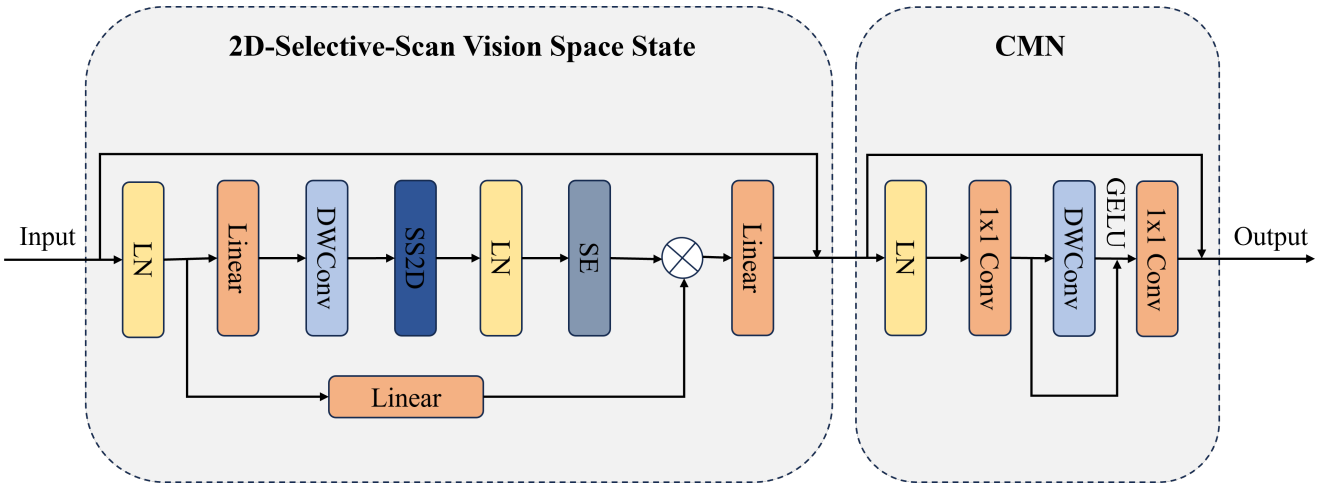

Figure 2. CM-SSM Block Framework Diagram.

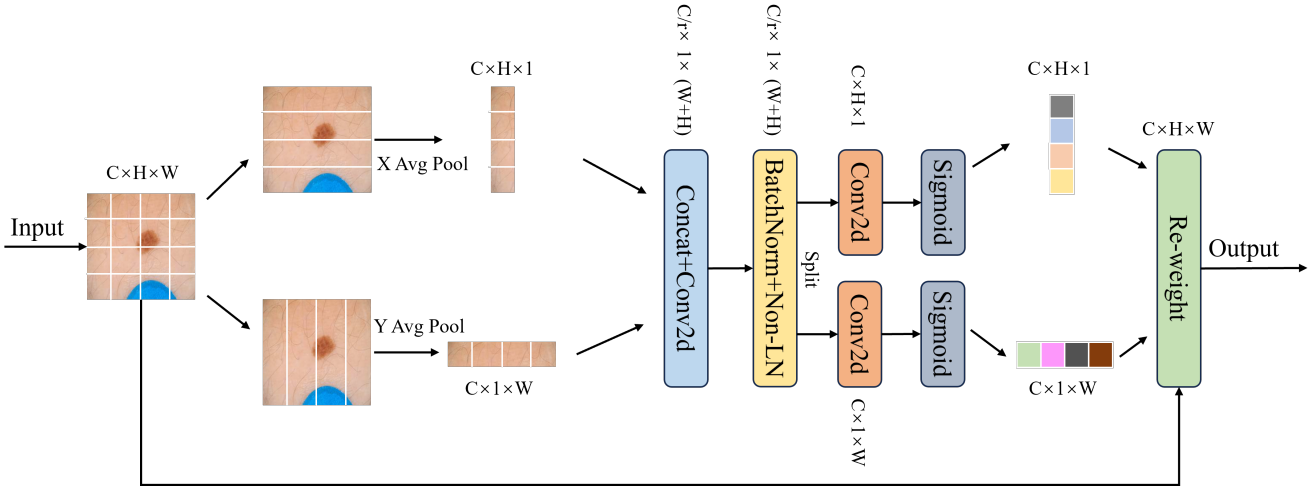

Figure 3. Framework Diagram of the cross-region Attention Mechanism.

ing operations,

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{1}$$

Specifically, for any input x, we initially apply a pooled filter with dimensions (H, 1) or (1, W) to extract features from the height and width regions of each channel, respectively. Thus, the height of h of the first c channel's output can be formulated as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{2}$$

Likewise, the output of channel c with width w can be described as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{3}$$

The two transformations mentioned above apply distinct spatial operations to the input x: one targeting the height direction and the other focusing on the width direction, thereby aggregating feature information. These transformations yield two separate feature maps, each designed to capture features in a single dimension. We then convert the feature map into two weight maps with long distance dependent coding along the width and height directions, respectively.

$$f = \delta(F_1([z^h, z^w])) \tag{4}$$

In the equation, the symbol $[\cdot, \cdot]$ indicates the concatenation of features across spatial dimensions. The symbol

δ represents a nonlinear activation function, while f is an intermediate feature map with dimensions $\mathrm{R}^{C/r\times(H+W)}$, which encodes spatial information in both height and width directions. Here, r is the reduction ratio. Next, we split f throughout the height and width dimension into two tensors: f^h and f^w where $f^h \in \mathrm{R}^{C/r\times H}$ and $f^w \in \mathrm{R}^{C/r\times W}$. We then apply two 1×1 convolution transformations, T_h and T_w, to convert f^h and f^w into tensors that match the channels of the input x, we get:

$$g^h = \sigma(T_h(f^h)) \tag{5}$$

$$g^w = \sigma(T_w(f^w)) \tag{6}$$

δ is the sigmoid function, we then extend the output g^h and g^w to enable their use as attention weights. Finally, the response from the cross-region attention layer of this paper can be expressed as:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

In summary, the cross-region attention mechanism effectively merges region information with channel attention to boost the network's capacity to model long-range dependencies while keeping resource usage minimal.

3.4. Loss Function Complementary Optimization

In U-Net, the strength of the supervised signals differs between the encoder and decoder, resulting in certain encoder blocks learning information relevant to the segmentation task, while the decoder blocks in the last layer can more accurately understand the real segmentation region (the decoder can receive directly supervised signals from the outputs due to its closer proximity to the final segmentation result). Simultaneously, due to the over-parameterization of U-Net, there is a large amount of redundancy between the deeper feature channels, learning similar features, which can lead to performance degradation and unnecessary computational overhead.

To solve the above two problems, two loss functions, cross-feature distillation and internal feature distillation, are introduced in this paper. Among them, cross-feature distillation is used to solve the supervision asymmetry problem between encoder and decoder in U-Net. Since the D1 layer in the decoder has an accurate understanding of the ground is really segmented region, which contains the most semantic information, The feature maps from this layer serve to offer supplementary supervision for the other blocks within the U-Net, with the following formula:

$$\mathrm{T} = \frac{1}{|M-1||\mathcal{J}|}$$
$$\mathcal{L}_{CFD} = \mathrm{T}\sum_{m=1}^{M-1}\sum_{I\sim\mathcal{J}} \parallel \mathrm{RCS}(F_m^i) - \mathrm{AvgPool}(F_{final}) \parallel^2 \tag{8}$$

Where M denotes the count of blocks, and J represents the quantity of images, and F_{final} is the feature mapping located in the last decoded block (D_1) of the feature mapping, and F_m^i is all the feature mappings of layer i located in the Mth block ($\mathbf{E}_1^{(1)}$, $\mathbf{E}_1^{(2)}$, ..., $\mathbf{D}_3^{(2)}$, $\mathbf{D}_2^{(1)}$, $\mathbf{D}_2^{(2)}$) except for D_1. To align the features in both channel and spatial dimensions, this paper employs average pooling along with a random channel selection (RCS) operation.

Internal feature distillation addresses redundancy in deep feature channels by transferring information from shallow features to deeper ones. This is achieved using the L2 paradigm[7, 21, 41] penalty to encourage deeper features to acquire valuable contextual information, ultimately enhancing their sensitivity to context and improving the model's overall performance and accuracy. Specifically, it can be formulated as:

$$\mathcal{L}_{IFD} = \frac{1}{|M||\mathcal{J}|}\sum_{m=1}^{M}\sum_{I\sim\mathcal{J}} \parallel \widetilde{F_m^\iota} - \overline{F_m^\iota} \parallel^2 \tag{9}$$

Where M denotes the count of blocks, and J indicates the amount of images, and F_i^k denotes all feature pictures located in layer i of the mth block ($\mathrm{E}_1^{(1)}$, $\mathrm{E}_1^{(2)}$, ...,$\mathrm{D}_2^{(1)}$, $\mathrm{D}_2^{(2)}$). $\tilde{F}$ represents the channel feature from deep-level layers, and $\bar{F}$ refers to the channel feature from shallow-level layers. In this paper, the channels are divided into upper and lower halves and are bounded by this to ensure that the shallow and deep channels have the same number of features.

By the preceding explanation, the study integrates the cross-feature distillation loss with the intrinsic feature distillation loss to formulate the composite distillation loss, termed FDLoss (Feature Distillation Loss). The FDLoss is formulated as follows:

$$\mathcal{L}_{fd} = \lambda_1\mathcal{L}_{CFD} + \lambda_2\mathcal{L}_{IFD} \tag{10}$$

of which $\mathcal{L}_{CFD}$ and $\mathcal{L}_{IFD}$ are cross-characteristic distillation and internal characteristic distillation loss functions, respectively. λ_1 and λ_2 are the equilibrium parameters, and (0.015,0.015) is usually chosen as the default parameter.

Finally, the characteristic distillation loss (FDLoss), binary cross entropy and Dice coefficient loss are fused. This function is used to evaluate the difference between the estimated results and the true values. The detailed formulation is presented below:

$$
\begin{aligned}
L_{Bce} &= -\frac{1}{N}\sum_{1}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \\
L_{\text{Dice}} &= 1 - \frac{2|X \cap Y|}{|X| + |Y|} \\
L_{\text{BceDice}} &= \lambda_1 L_{\text{Bce}} + \lambda_2 L_{\text{Dice}} \\
L_{\text{total}} &= L_{fd} + L_{\text{BceDice}}
\end{aligned} \tag{11}
$$

Where N signifies the total sample size. y_i and $\hat{y}_i$ denote the true values and the estimated results, respectively.$|X|$ and $|Y|$ represent the ground truth and estimated values. λ_1 and λ_2 are the weights of the loss function with default values of 0.3 and 0.7, respectively.

4. Experiments

In this section, the paper conducts extensive experiments to assess the proposed approach's effectiveness. Specifically, the performance of the proposed methods is evaluated on the ISIC2016, ISIC2017, and ISIC2018 datasets.

4.1. Datasets

4.1.1 ISIC2016

ISIC2016 is a dataset of dermoscopic images from the International Symposium on Biomedical Imaging (ISBI) focusing on skin lesion analysis to tackle challenges in melanoma detection. The dataset comprises 1,279 dermoscopy images, each accompanied by corresponding segmentation mask labels. Adhering to the methodology of prior research, this manuscript partitions the dataset, allocating 70% for model training and reserving 30% for validation. The detailed allocation is detailed as: For the ISIC2016 dataset, 900 instances are allocated for training, while 379 are earmarked for the validation phase.

4.1.2 ISIC2017

ISIC2017 is a dermoscopy image dataset designed for skin lesion analysis at the ISBI, focusing on three types of disease challenges: melanoma, seborrheic keratosis, and benign nevi. It contains 2,150 dermoscopy images with corresponding segmentation mask labels. Similarly, the research applies a 7:3 split for the ISIC2017 dataset, creating separate subsets for training and evaluation. To be precise, the training subset encompasses 1,500 images, and the evaluation subset consists of 650 images.

4.1.3 ISIC2018

ISIC2018 is a dermoscopy image dataset for skin lesion analysis at the ISBI for seven types of skin disease challenges such as actinic keratoses, basal cell carcinomas, and benign keratoses. For the segmentation task, the ISIC2018 dataset encompasses 2694 dermatoscopic images paired with their segmentation masks. Adhering to the prior dataset division approach, this research similarly apportioned the data into a 7:3 training-to-testing ratio. In detail, the training subset encompasses 1,886 images, while the test subset comprises 808 images.

4.2. Evaluation Indicators

In order to evaluate the efficacy of the suggested approach, this study employs three key performance indicators for semantic segmentation: the average Intersection over Union (IoU), the Dice Coefficient(DSC), and accuracy (Acc). The mIoU and DSC are measures of consistency between the predicted mask and the true annotation, while Acc is used to assess the overall accuracy of the model. the mathematical representations of mIoU, DSC, and Acc are summarized below:

$$
mIoU = \frac{TP}{TP + FP + FN} \tag{12}
$$

$$
DSC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{13}
$$

$$
Acc = \frac{TN + TP}{N} \tag{14}
$$

Here, TP denotes true positives, FP denotes false positives, TN denotes true negatives, and FN denotes false negatives. The variable N represents the aggregate number of instances.

4.3. Main Results

In line with prior research[38, 5], the images from the ISIC2016, ISIC2017, and ISIC2018 datasets have been adjusted to a dimension of 256 by 256 pixels. To mitigate the possibility of overfitting, this analysis incorporated data augmentation techniques such as random horizontal flipping and rotational adjustments. The loss function delineated in section 3.4 serves as the criterion for this dataset. For this analysis, batches of 32 examples were processed, and the AdamW[29] optimizer was chosen for the training phase, initiating the learning rate at 0.001. A Cosine Annealing Learning Rate Scheduler[30] was implemented to modulate the learning rate, with a cap at 50 cycles and a gradual reduction to a nadir of 1e-5. The training regimen spanned 300 epochs, and the computational experiments were executed on a solitary NVIDIA RTX 4090 GPU.

In order to assess the efficacy of the suggested approach, this research conducted empirical tests across the ISIC2016, ISIC2017, and ISIC2018 datasets, and juxtaposed the findings against those of contemporary leading techniques.

Model	mIoU (%)↑	DSC (%)↑	ACC (%)↑
U-Net	84.01	91.16	95.06
Attention-UNet	84.30	91.37	95.26
LV-UNet	85.95	92.35	95.25
VM-UNet	86.41	92.71	95.28
Ours	**88.16**	**93.71**	**95.52**

Table 1. Comparative experimental results of different methods (bold represents the best) on ISIC2016 Dataset.

4.3.1 ISIC2016

This paper addresses the ISIC2016 dataset in U-Net[36], Attention-UNet[34], LV-UNet[25], VM-UNet[37] Exhaustive experiments were performed on these methods. The qualitative visualization results are shown in Figure 4.

The visualization results show that the method proposed in this study has significant advantages in the image segmentation task of ISIC2016 dataset. This approach effectively extracts and utilizes both channel and spatial features by incorporating a channel mixing network and a cross-region attention mechanism. Additionally, feature redundancy and supervision imbalance are addressed through the integration of the feature distillation loss function. The method's performance is compared with other advanced techniques, and the results of the experiments are summarized in Table 1.

The data indicate that the proposed method demonstrates a significant performance enhancement compared to other existing techniques. Specifically, when compared to VM-UNet, this method achieves improvements of 1.75% in mIoU and 1% in the DSC on the ISIC2016 dataset. This improvement is mainly owing to the efficient extraction and effective use of channel and spatial features, as well as the resolution of feature redundancy and supervision imbalance within the model.

4.3.2 ISIC2017

For the ISIC2017 dataset, this paper is presented in U-Net[36], UTNetV2[15], TransFuse[50], MALUNet[38], MedMamba[49], HC-Mamba[47] and VM-UNet[37] Exhaustive experiments were performed on these methods. The qualitative visualization results are shown in Figure 5.

Figure 5 illustrates that the proposed approach achieves excellent image segmentation performance on the ISIC2017 dataset. The introduction of the channel mixing network and cross-region attention mechanism optimizes the extraction and utilization of both channel and spatial features, effectively addressing feature redundancy and supervision imbalance through the integration of the feature distillation loss function. The performance of this method is further compared with other advanced techniques, with the experimental results detailed in Table 2.

Model	mIoU (%)↑	DSC (%)↑	ACC (%)↑
U-Net	76.98	86.99	95.65
UTNetV2	77.35	87.23	95.84
TransFuse	79.21	88.40	96.17
MALUNet	78.78	88.13	96.18
MedMamba	78.82	88.15	95.01
HC-Mamba	79.27	88.18	95.17
VM-UNet	79.54	88.60	96.29
Ours	**81.48**	**89.79**	**96.61**

Table 2. Comparative experimental results of different methods (bold represents the best) on ISIC2017 Dataset.

The results indicate that this method demonstrates significant advantages across all evaluation metrics. Specifically, on the ISIC2017 dataset, the proposed method achieves mIoU and Dice scores that are 1.94% and 1.19% higher than those of VM-UNet, respectively. This further confirms the superiority and reliability of the proposed approach for image segmentation tasks. These improvements stem from the optimized extraction of channel and spatial features, as well as the effective resolution of feature redundancy and supervision imbalance within the model.

4.3.3 ISIC2018

For the ISIC2018 dataset, the methods compared in this paper are U-Net[36], Unet++[51], UTNetV2[15], MALUNet[38], MedMamba[49], HC-Mamba[47] and VM-UNet[37]. The results of the qualitative visualization of the proposed method in this paper at ISIC2018 are shown in Figure 6.

As illustrated in Figure 6, on the ISIC2018 dataset, the method proposed in this study also shows excellent segmentation results. The success is attributed to the use of channel mixing networks and cross-zone attention mechanisms, which enhance the extraction and utilization of both channel and spatial features. Additionally, the application of the feature distillation loss function effectively mitigates feature redundancy and supervision imbalance within the model. In order to further verify the efficiency of this method, we compare it with many advanced methods. The relevant experimental data are shown in Table 3.

As shown above, our proposed method achieves optimal performance on all major evaluation indicators. Specifically, in the evaluation of the ISIC2018 dataset, our approach showed a significant performance advantage by improving mIoU and DSC by 0.97% and 0.59%, respectively, compared to the VM-UNet approach. These results are primarily due to enhancements in optimizing the extraction of channel and spatial features, as well as effectively addressing feature redundancy and supervision imbalance within

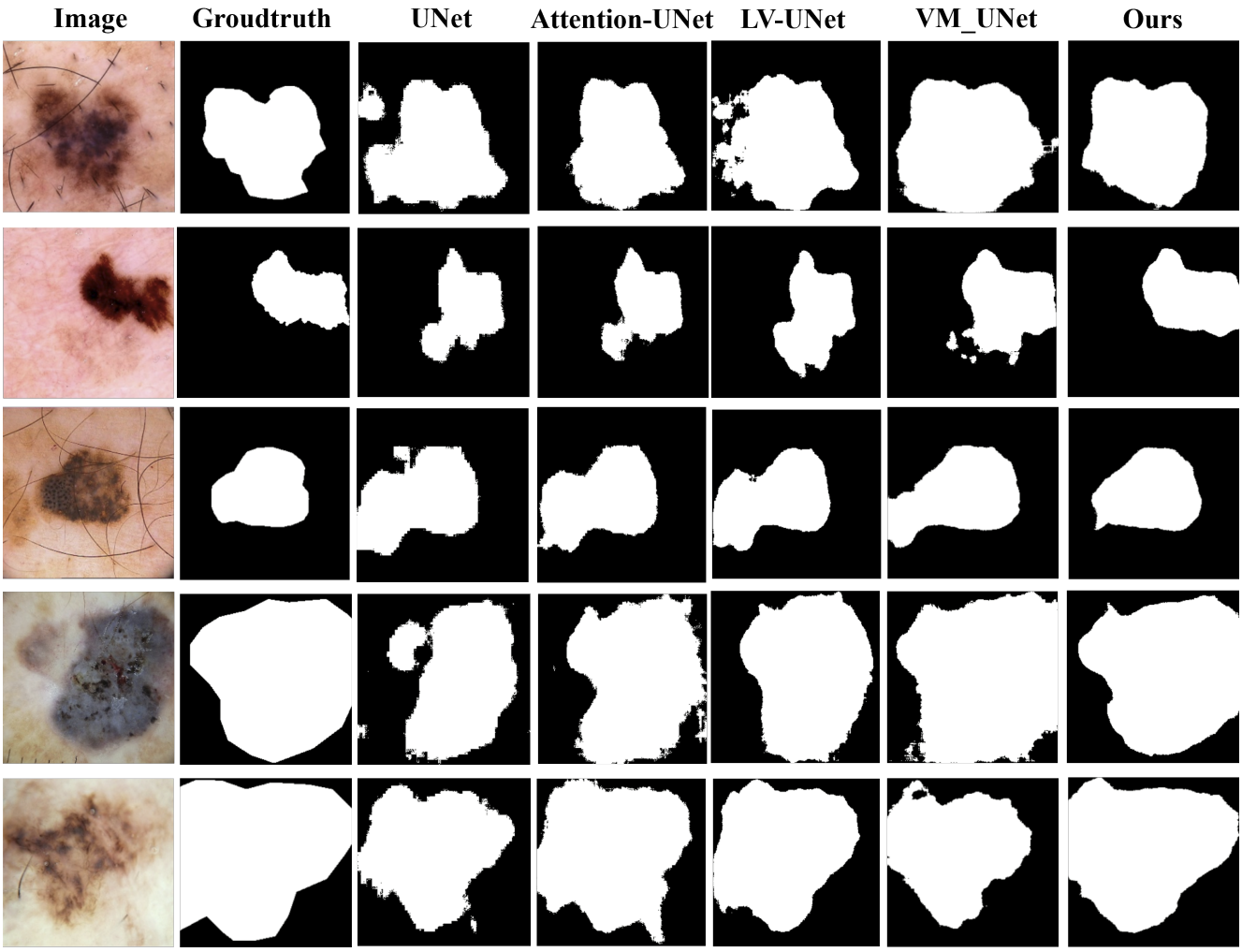

Figure 4. Comparative experimental results of different methods (bold represents the best) on ISIC2016 Dataset.

Model	mIoU (%)↑	DSC (%)↑	ACC (%)↑
U-Net	77.86	87.55	94.05
Unet++	78.31	87.83	94.02
UTNetV2	78.97	88.25	94.32
MedMamba	79.13	88.35	94.23
MALUNet	80.25	89.04	94.62
HC-Mamba	80.60	89.25	94.84
VM-UNet	80.28	89.06	94.57
Ours	**81.25**	**89.65**	**94.94**

Table 3. Comparative experimental results of different methods (bold represents the best) on ISIC2018 Dataset.

the model. Such findings not only underscore the method's high efficiency and accuracy in dermatological image segmentation but also reinforce its leading position in the field of medical image segmentation.

4.4. Ablation Experiments

4.4.1 Results under different module combinations

Within this segment, we conduct dismantling studies focusing on the three components: Squeeze-and-Excitation Networks (SE) and Channel Mixing Network (CMN) and Cross Region Attention (CRA) proposed in this paper. These experiments aim to verify the contributions of these modules to the overall effectiveness of the method.

As shown in Table 4, compared to the baseline metrics, SE improved the mIoU and Dice coefficients by 0.9% and 0.56% on the ISIC2017 dataset, indicating that the SE module could effectively enhance the important features and suppress the unimportant features; the Channel Mixing Network (CMN) enhances the mIoU and Dice coefficients on the ISIC2017 dataset by 1.77% and 1.09% respectively. This indicates that CMN effectively boosts the exchange

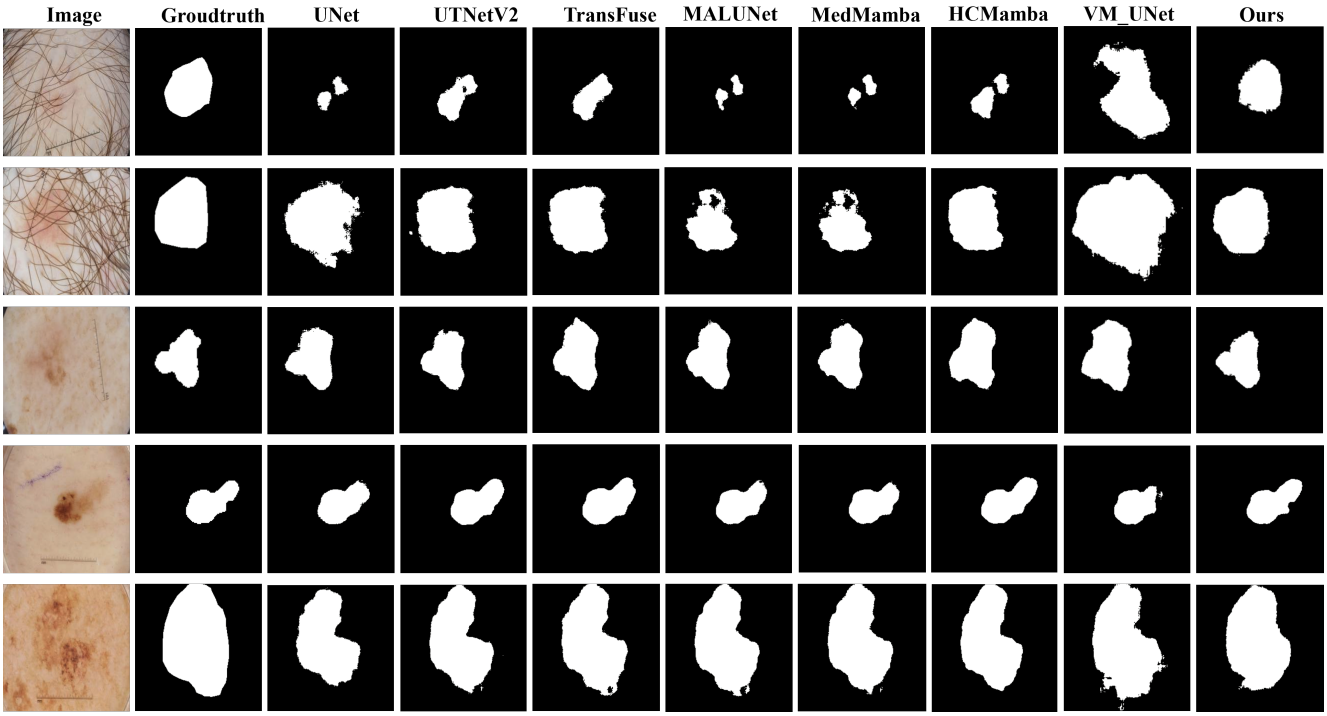

Figure 5. Examples of segmentation results for different methods on ISIC2017 Dataset.

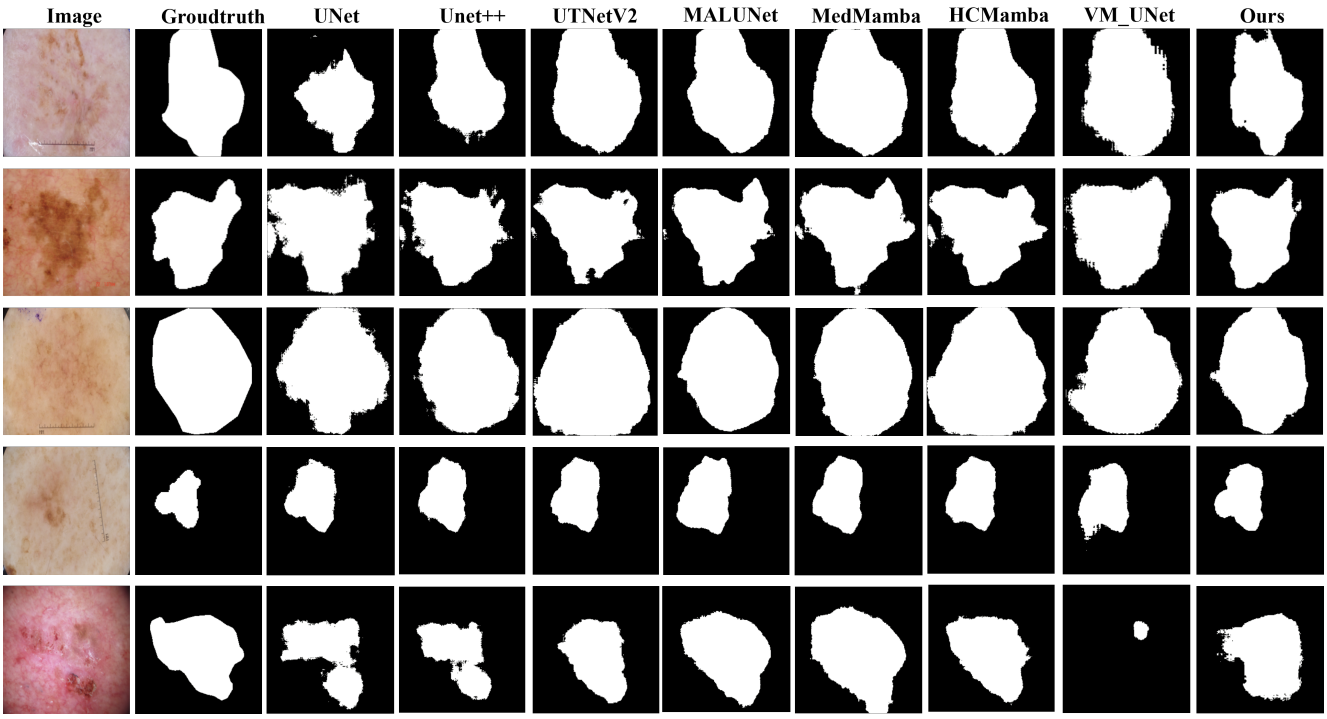

Figure 6. Examples of segmentation results for different methods on ISIC2018 Dataset.

of channel information and the capture of channel features. Meanwhile, the Cross Region Attention (CRA) module im-

SE	CMN	CRA	Evaluation	
			mIoU(%)↑	DSC(%)↑
×	×	×	79.54	88.60
✓	×	×	80.44	89.16
×	✓	×	79.98	88.88
×	×	✓	80.97	89.45
✓	✓	×	81.31	89.69
✓	×	✓	80.13	88.97
×	✓	✓	80.39	89.13
✓	✓	✓	**81.48**	**89.79**

Table 4. Comparative experimental results of different methods (bold represents the best).

$\mathcal{L}_{CFD}$	$\mathcal{L}_{IFD}$	Evaluation	
		mIoU(%)↑	DSC(%)↑
×	×	80.66	89.26
✓	×	81.35	89.57
×	✓	81.24	89.44
✓	✓	81.48	89.79

Table 5. Ablation results of different loss functions on ISIC2017 dataset.

proves the mIoU and DSCby 1.43% and 0.85%, respectively, demonstrating its ability to enhance the network's modeling of long-range dependencies, thereby improving segmentation accuracy. When the three modules are enabled at the same time, the proposed method achieves the highest performance on the ISIC2017 dataset, with mIoU and Dice coefficients increasing to 81.48% and 89.79%, respectively. This result fully proves the importance of the synergistic effect of the three modules to improve the model performance. The combination of the three modules not only optimizes the feature representation of the image, but also enhances the recognition ability of the model to the key areas, so as to achieve more accurate image segmentation.

4.4.2 Results under different combinations of loss functions

In order to further verify the effectiveness of the optimization of loss function in this paper, ablation experiments were carried out on the introduced cross characteristic distillation loss function ($\mathcal{L}_{CFD}$) and internal characteristic distillation loss function ($\mathcal{L}_{IFD}$), and the results were shown in Table 5. The results show that $\mathcal{L}_{CFD}$ and $\mathcal{L}_{IFD}$ can effectively enhance the feature extraction capability of the model and improve the segmentation accuracy of the model.

5. Conclusion

This study introduces an innovative technique for dermatological medical image segmentation that improves the delineation of skin lesions by utilizing compressed excitation modules and cross-area attention mechanisms to enhance both channel and spatial features. Additionally, the model's feature extraction efficiency is further boosted by incorporating a channel mixing network and a feature distillation loss function. When compared to current state-of-the-art methods and recent Mamba-based algorithms, this approach demonstrates a significant enhancement in performance. Moving forward, we intend to thoroughly investigate the full potential of the method presented here, aiming to increase its segmentation accuracy in skin imaging. Simultaneously, we will explore the applicability of this technique on datasets acquired through other medical imaging methods.

Acknowledgement

This work was partially supported by Xinjiang Autonomous Region Natural Science Foundation General Program under Grant 2024D01C53, Finance Science and Technology Project of Xinjiang Uyghur Autonomous Region under Grant 2023B01029-1, Xinjiang Uygur Autonomous Region Key Research and Development Task Special Project under Grant 2022B02052-3, Ningbo Key R&D Program under Grant No.2023Z231, the China Postdoctoral Science Foundation under Grant No.2022M723863 and the Researchers Supporting Project number (RSPD2024R968), King Saud University Riyadh. The authors would like to express their heartfelt gratitude to those people who have helped with this manuscript and to the reviewers for their comments on the manuscript.

References

[1] S. Andermatt, S. Pezold, and P. Cattin. Multi-dimensional gated recurrent units for the segmentation of biomedical 3d-data. In *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, pages 142–151. Springer, 2016. 2

[2] S. Andermatt, S. Pezold, and P. Cattin. Multi-dimensional gated recurrent units for the segmentation of biomedical 3d-data. In *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, pages 142–151. Springer, 2016. 3

[3] A. Behrouz, M. Santacatterina, and R. Zabih. Mambamixer: Efficient selective state space models with dual token and channel selection. *arXiv preprint arXiv:2403.19888*, 2024. 3

[4] M. Berseth. Isic 2017-skin lesion analysis towards melanoma detection. *arXiv preprint arXiv:1703.00523*, 2017. 2

[5] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical

image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2, 7

[6] C.-S. Chen, G.-Y. Chen, D. Zhou, D. Jiang, and D.-S. Chen. Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning. *arXiv preprint arXiv:2402.15761*, 2024. 3

[7] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 6

[8] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79:102444, 2022. 2

[9] S. P. Choy, B. J. Kim, A. Paolino, W. R. Tan, S. M. L. Lim, J. Seo, S. P. Tan, L. Francis, T. Tsakok, M. Simpson, et al. Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. *NPJ Digital Medicine*, 6(1):180, 2023. 2

[10] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2

[11] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 2

[12] K. Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007. 2

[13] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022. 3

[15] Y. Gao, M. Zhou, D. Liu, and D. Metaxas. A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arxiv 2022. *arXiv preprint arXiv:2203.00131*, 2022. 8

[16] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 3

[17] A. Gu, K. Goel, A. Gupta, and C. Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022. 3

[18] A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

[19] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019. 2

[20] A. Gupta, A. Gu, and J. Berant. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022. 3

[21] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2, 6

[22] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[23] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020. 2

[24] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. 4

[25] J. Jiang, M. Wang, H. Tian, L. Cheng, and Y. Liu. Lv-unet: A lightweight and vanilla model for medical image segmentation. *arXiv preprint arXiv:2408.16886*, 2024. 8

[26] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017. 3

[27] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu. Vmamba: Visual state space model, 2024. 3, 4

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[29] I. Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[30] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[31] J. Ma, F. Li, and B. Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 3

[32] H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*, 2022. 3

[33] G. Muhammad, M. S. Hossain, and N. Kumar. Eeg-based pathology detection for home health monitoring. *IEEE Journal on Selected Areas in Communications*, 39(2):603–610, 2020. 2

[34] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla,

B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 8

[35] X. Pei, T. Huang, and C. Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. 4

[36] O. Ronneberger, P. Fischer, and T. Brox. 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*, pages 234–241. Springer, 2015. 2, 3, 8

[37] J. Ruan and S. Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024. 3, 8

[38] J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1150–1156. IEEE, 2022. 2, 7, 8

[39] J. Ruan, M. Xie, J. Gao, T. Liu, and Y. Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 481–490. Springer, 2023. 2

[40] J. Ruan, M. Xie, S. Xiang, T. Liu, and Y. Fu. Mew-unet: Multi-axis representation learning in frequency domain for medical image segmentation. *arXiv preprint arXiv:2210.14007*, 2022. 2

[41] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019. 2, 6

[42] J. T. Smith, A. Warrington, and S. W. Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 3

[43] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. 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*, pages 36–46. Springer, 2021. 2

[44] H. Wang, P. Cao, J. Wang, and O. R. Zaiane. 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, pages 2441–2449, 2022. 2

[45] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023. 2

[46] Z. Wang, Z. Wan, H. Han, B. Liao, Y. Wu, W. Zhai, Y. Cao, and Z.-j. Zha. Mambapupil: Bidirectional selective recurrent model for event-based eye tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5762–5770, 2024. 3

[47] J. Xu. Hc-mamba: Vision mamba with hybrid convolutional techniques for medical image segmentation. *arXiv preprint arXiv:2405.05007*, 2024. 8

[48] J. X. Yang, J. Zhou, J. Wang, H. Tian, and A. W. C. Liew. Hsimamba: Hyperpsectral imaging efficient feature learning with bidirectional state space for classification. *arXiv preprint arXiv:2404.00272*, 2024. 3

[49] Y. Yue and Z. Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024. 8

[50] Y. Zhang, H. Liu, and Q. Hu. Transfuse: Fusing transformers and cnns 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*, pages 14–24. Springer, 2021. 2, 8

[51] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2, 8

[52] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3, 4

[53] N. Zubic, M. Gehrig, and D. Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5819–5828, 2024. 3