MCNMF-Unet: a mixture Conv-MLP network with multi-scale features fusion Unet for medical image segmentation

Lei Yuan [1], **Jianhua Song** [Corresp., 1], **Yazhuo Fan** [1]

[1] School of Physics and Information Engineering, Minnan Normal University, Zhangzhou, Fujian, China

Corresponding Author: Jianhua Song
Email address: songjianhua@mnnu.edu.cn

Recently, the medical image segmentation scheme combining Vision Transformer (ViT) and multilayer perceptron (MLP) has been widely used. However, one of its disadvantages is that the feature fusion ability of different levels is weak and lacks flexible localization information. To reduce the semantic gap between the encoding and decoding stages, we propose a mixture conv-MLP network with multi-scale features fusion Unet (MCNMF-Unet) for medical image segmentation. MCNMF-Unet is a U-shaped network based on convolution and MLP, which not only inherits the advantages of convolutional in extracting underlying features and visual structures, but also utilizes MLP to fuse local and global information of each layer of the network. MCNMF-Unet performs multi-layer fusion and multi-scale feature map skip connections in each network stage so that all the feature information can be fully utilized and the gradient disappearance problem can be alleviated. At the same time, MCNMF-Unet designed a solution for the adverse effects caused by image cropping and reduced the number of parameters and computational complexity. We evaluated the proposed model on BUSI, ISIC2018 and CVC-ClinicDB datasets. The experimental results show that the performance of our proposed model is superior to most existing networks, with an IoU of 84.72% and a F1-score of 91.39%.

MCNMF-Unet: A Mixture Conv-MLP Network with Multi-scale Features Fusion U-Net for Medical Image Segmentation

Lei Yuan, Jianhua Song, and Yazhuo Fan

Key Laboratory of Light Field Manipulation and System Integration Applications in Fujian Province, School of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China

Corresponding author:
Jianhua Song

Email address: songjianhua@mnnu.edu.cn

ABSTRACT

Recently, the medical image segmentation scheme combining Vision Transformer (ViT) and multilayer perceptron (MLP) has been widely used. However, one of its disadvantages is that the feature fusion ability of different levels is weak and lacks flexible localization information. To reduce the semantic gap between the encoding and decoding stages, we propose a mixture conv-MLP network with multi-scale features fusion U-Net (MCNMF-Unet) for medical image segmentation. MCNMF-Unet is a U-shaped network based on convolution and MLP, which not only inherits the advantages of convolutional in extracting underlying features and visual structures, but also utilizes MLP to fuse local and global information of each layer of the network. MCNMF-Unet performs multi-layer fusion and multi-scale feature map skip connections in each network stage so that all the feature information can be fully utilized and the gradient disappearance problem can be alleviated. At the same time, MCNMF-Unet designed a solution for the adverse effects caused by image cropping and reduced the number of parameters and computational complexity. We evaluated the proposed model on BUSI, ISIC2018 and CVC-ClinicDB datasets. The experimental results show that the performance of our proposed model is superior to most existing networks, with an IoU of 84.72% and a F1-score of 91.39%.

INTRODUCTION

In recent years, high-performance methods based on convolutional neural network (CNN) have demonstrated superior performance on many tasks (Kadry et al., 2022; Sun et al., 2022; Zamir et al., 2021; Li et al., 2021b; Ding et al., 2022b; Kalake et al., 2022). Benefiting from the development of CNN, computer vision techniques have been widely used in the field of medical image processing. Image semantic segmentation is an important component of medical image processing, especially accurate and robust medical image segmentation techniques can play a cornerstone role in computer-aided diagnosis and image-guided clinical surgery (Hatamizadeh et al., 2022; Valanarasu and Patel, 2022; Xie et al., 2022).

Image semantic segmentation can be formulated as a typical dense prediction problem, which aims at the pixel-level classification of feature maps. Existing CNN-based medical image segmentation methods mainly rely on fully convolutional neural network (FCNN) (Isensee et al., 2021; Jin et al., 2020). The most typical of these is the Unet (Ronneberger et al., 2015), which consists of a symmetric encoder-decoder and skip connections. With such an elegant structural design, Unet has achieved great success in medical image processing. Along this technical line, many algorithms have been developed for various types of medical image segmentation. The excellent performance of these Unet-based methods in medical image segmentation has demonstrated the strong ability of CNN to learn features. However, the inherent localization and weight sharing of the receiver domain in convolutional operations make it difficult for CNN-based methods to learn explicit global information and remote semantic information interactions (Xie et al., 2021), which to some extent cannot meet the stringent requirements for segmentation accuracy in the field of medical image segmentation. Many researchers have noticed this problem and designed

some modules to solve it, including Residual learning (He et al., 2016), Dense connections (Huang et al., 2017), Self-attention mechanisms (Schlemper et al., 2019; Wang et al., 2018) and Image pyramids (Zhao et al., 2017). Nevertheless, these methods still have some limitations and cannot explicitly model dependencies over long-distance and often exhibit sub-segmentation results. Applying Transformer to computer vision (Wang et al., 2021b; Han et al., 2021; Zheng et al., 2021) can alleviate the long-distance dependencies to some extent compared to other traditional CNN-based methods. At the same time, Transformer has a powerful global relationship modeling capability that has yielded amazing results in medical image analysis tasks. Dosovitskiy et al. (2020) proposed Vision Transformer (ViT) to perform image recognition tasks. Taking 2D image patches with location markers as input and pre-trained on large datasets, ViT achieves comparable performance to CNN-based methods. Many Transformer-based architectures have also been proposed in the field of medical image segmentation, such as Trans-Unet (Chen et al., 2021b), Swin-Unet (Cao et al., 2023), ConViT (d'Ascoli et al., 2021) and ScaleFormer (Huang et al., 2022).

Many researchers have demonstrated the great potential of the structure on ViT-based image analysis (Azad et al., 2023; Dalmaz et al., 2022; Li et al., 2021a) and also promoted the study of Multi-Layer Perceptron (MLP) structures, such as MLP-Mixer (Tolstikhin et al., 2021), gMLP (Liu et al., 2021a),RepMLP (Ding et al., 2022a) and CycleMLP (Chen et al., 2021c). In particular, MLP-Mixer, an entirely MLP-based network, gives comparable performance to Transformer with less computation. In spite of the fact that existing Transformer-based and MLP-based methods have proven to be promising for image analysis tasks, including medical image segmentation, several daunting challenges remain: (1) the network only accepts a fixed image size, and it is necessary to divide the image into a fixed size, which may not capture the fine-grained spatial details of the image; (2) it will inevitably cause boundary artifacts when applied to larger images (Chen et al., 2021a); (3) it lacks detailed positioning information because the input is considered as a one-dimensional sequence and only global information is modeled at all stages. When performing semantic analysis, the ability to localize the location of interest may be lacking (Ni et al., 2022). These problems are the shortcomings of Transformer and MLP compared to CNN in extracting the underlying features and visual structure.

To solve the above-mentioned problems, we propose a Mixture Conv-MLP Network with Multi-scale Features Fusion U-Net for Medical Image Segmentation(MCNMF-Unet). In each module, this network combines the advantages of convolution in extracting low-level features and visual structure and the advantages of MLP in fusing local and global information. The core of MCNMF-Unet is the Conv-MLP module, which incorporates the characteristics of encoder and decoder in a U-shaped network and fuses MLP and convolution in each layer, allowing it to make good use of the advantages of both methods. In this network, the key technology is MLP Cross Gating (MCG) Block and Multi-axis and Multi-windows MLP (MsM) Block. MCG fuses the convolutional block information of different nodes through two paths. MsM uses multi-axis and multi-window to capture local and global information from multiple dimensions. The branches obtained are equally divided by channel, and the information is mixed by different mechanisms on the respective axes. The computational burden of MsM is linearly related to the size of the input feature map. MCNMF-Unet has high performance in medical image segmentation with smaller parameters and FLOPs. The main contributions of this paper are as follows:

(1) In order to combine the advantages of convolution and MLP to improve segmentation accuracy, we designed a general framework for medical image segmentation called MCNMF-Unet using the U-shaped encoder-decoder architecture.

(2) A Multi-axis and Multi-windows MLP (MsM) module is designed to capture feature map information from multiple layers and multiple dimensions, whose input does not require cropping of images, can receive images of arbitrary size, and always has a global receptive field.

(3) A Conv-MLP module is developed to perform feature cross-fusion on the two outputs of the convolution module, which is also a multi-path and multi-information interaction.

(4) MCNMF-Unet achieves some effect on lightweight while improving partitioning. Compared with most Unet-based improved networks, the performance of parameters and FLOPs indicators is better. It can achieve better segmentation results with less computing time and space spent.

RELATED WORK

CNN of U-shaped Semantic Segmentation methods

Most of the early semantic segmentation research was based on traditional machine learning algorithms of contours and regions (Zhang et al., 2022; Tsai et al., 2003). In recent years, with the development of deep learning, the Unet architecture proposed by borrowing the Fully convolutional neural network (FCN) has rapidly become the baseline network for computer vision tasks by virtue of its elegant design and superior performance. Based on Unet's U-shaped architecture, various networks are proposed for semantic segmentation. Zhou et al. (2019) proposed the Unet++ network,which integrates Unet structures of different sizes into one network, captures features at different layers, and integrates them into a shallower Unet structure by feature superposition, resulting in smaller scale differences in the feature maps during fusion. Inspired by deep residual learning (ResNet), Zhang et al. (2018) proposed ResUnet. The proposed architecture uses a series of stacked residual units instead of ordinary neural units as the basic blocks to build deep ResUnet, which allows the network training layers to be deepened effectively. Oktay et al. (2018) proposed Attention Unet. The network proposes an attention gates (AGs) mechanism, which implicitly generates soft region suggestions and highlighting salient features useful for a specific task. By suppressing features of irrelevant regions, the sensitivity and accuracy of the model for dense label prediction are improved. Çiçek et al. (2016) proposed 3D Unet and introduced the structure into the field of 3D medical image segmentation.

Vision Transformers

Transformers, first proposed by Vaswani et al. (2017), is applied to natural language processing (NLP) tasks. Its multi-headed self-attentive and feedforward MLP layers are stacked to capture non-local interactions between words. In the field of NLP, the method has achieved optimal performance in various tasks. Inspired by the success of Transformers, Dosovitskiy et al. (2020) pioneered its introduction into computer vision and proposed ViT, which is the first pure Transformer architecture for image recognition. Recently, many researchers have tried to introduce ViT into medical image processing tasks and developed many Transformer-based models (Heidari et al., 2023; Huang et al., 2021; Wang et al., 2021a). Chen et al. (2021b) proposed TransUNet, which integrates the advantages of Transformers and Unet. TransUnet uses the local information encoded by CNN and the global context encoded by Transformer to encode tokenized image blocks from the CNNs feature map as an input sequence for extracting the global context. The decoder upsamples the encoded features and then combines them with the high-resolution CNN feature map for accurate localization. Cao et al. (2023) proposed Swin-UNet inspired by the shift window mechanism of Swin Transformer (Liu et al., 2021b), and their work is the first pure U-based architecture based on Transformer, where the encoder, bottleneck and decoder are built based on Swin-Transformer blocks. In the encoder, self-attentiveness from local to global is implemented; in the decoder, global features are upsampled to the input resolution for the corresponding pixel-level segmentation prediction.

MLP Vision

Recently, many researchers have been considering the necessity of a self-attention mechanism with a tremendous amount of computation in the Transformer architecture in computer vision. Tolstikhin et al. (2021) proposed the MLP-Mixer architecture (Mixer), an architecture developed entirely based on multilayer perceptrons (MLPs), which enables the interaction of two input dimensions through the interleaving of channel-mixing MLPs and token-mixing MLPs. MLP-Mixer architecture completely replaces the self-attention mechanism in ViT. Liu et al. (2021a) proposed gMLP, which is constructed from a basic MLP layer with gating. As an alternative to Transformer without self-attentiveness, gMLP consists only of channel and spatial projections with static parameterization. By comparing experimental results, it has been proven that the self-attention mechanism is not important for ViT, and gMLP can achieve the same accuracy in critical language and vision applications. Ding et al. (2022a) proposed RepMLP, which incorporates local prior knowledge into a fully connected (FC) layer while reducing the number of parameters and inference time. The architecture takes full account of the global representation capability of the fully connected (FC) layer and the local capture property of the convolutional layer, using convolution to strengthen FC and make it possess locality and globality. Therefore, RepMLP is more suitable for computer vision tasks.

METHOD

The proposed model is a novel architecture combining the advantages of convolution and MLP for medical image segmentation. In this section, we introduce the backbone network of MCNMF-Unet and also describe the main component modules of the network, including Encoder Conv-MLP (ECM), Decoder Conv-MLP (DCM), Bottleneck Conv-MLP (BCM), Output Conv-MLP (OCM), Multi-axis and Multi-windows MLP (MsM) and MLP Cross Gating (MCG).

The main backbone of MCNMF-Unet

The main framework of MCNMF-Unet follows the most classical U-like architecture, as shown in Fig. 1. For medical image segmentation, many studies have demonstrated that the Unet-like network is still a very competitive infrastructure (Wang et al., 2022; Wu et al., 2023; Azad et al., 2022). Most of the research on Unet-like networks is based on the CNN method, which has great advantages in extracting underlying features. However, its local receptive field and long-distance weak dependence still hinder the accuracy of complex structure segmentation. In addition to the advantages of Unet, an excellent segmentation network should also introduce MLP, which can increase the global receptive field of the network. Based on this idea, Encoder Conv-MLP (ECM) module, Decoder Conv-MLP (DCM) module, Bottleneck Conv-MLP (BCM) module and Output Conv-MLP (OCM) module are designed to build the U-shaped network. The MLP Cross Gating (MCG) in the ECM, DCM, BCM and OCM is used to fuse the information of the two convolutional blocks, and the extracted fine local features are then used to obtain the global receptive field through MLP. The DCM, BCM and OCM performs skip connection with the ECM. Each DCM module will connect with two different ECM modules, and the OCM module and BCM module will connect with one ECM module. MCNMF-Unet architecture adopts an end-to-end training strategy that relies on a hybrid model design for each block, adaptively acquiring local and global sensory fields at each layer, where Conv is used locally and MLP is used for long-distance interaction, which can make full use of all feature information.

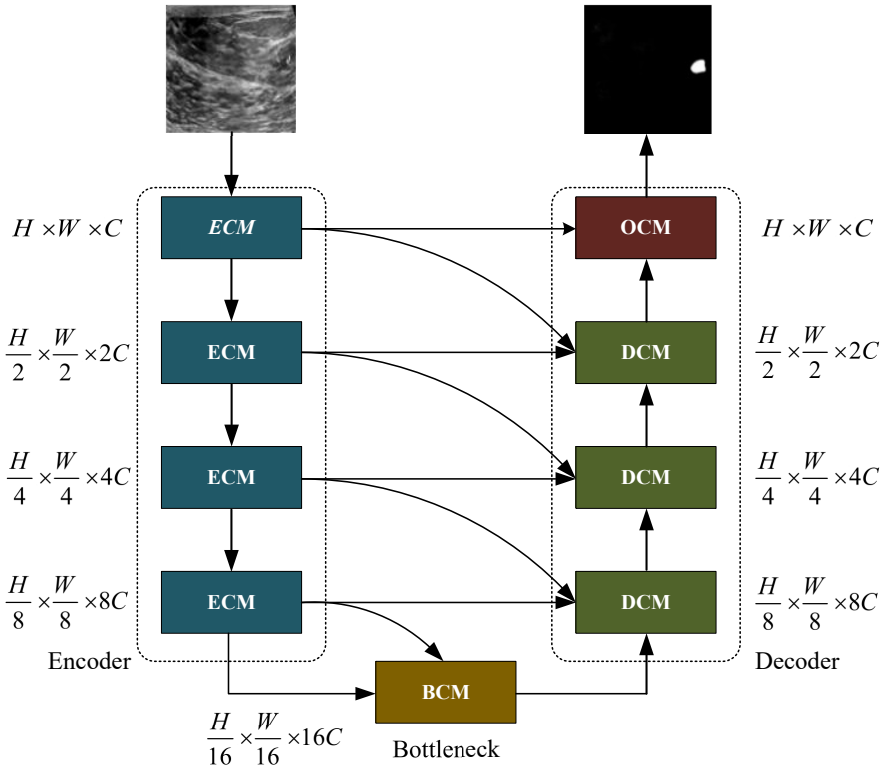

Figure 1. The MCNMF-Unet architecture

The crucial sub-modules

The core sub-modules of MCNMF-Unet are ECM, DCM, BCM and OCM, which form the encoder and decoder of the U-shaped network. The Conv 3×3 block consists of Conv 3×3, BN and ReLU blocks. The ECM block references the encoder design concept of a standard U-shaped network, and considering the problem of information loss during encoder downsampling, the convolutional block downsampling and then upsampling operation is used, so that the information of the next layer can be obtained and at the same time as little information can be lost as possible, as shown in Fig. 2. In addition, the output of two convolutional blocks is obtained at the same layer and enters the MLP Cross Gating module (MCG), which cross-fuses the information from the convolutional blocks. The output of MCG block is used for skip connection with the decoder, and it is also downsampled to double the size and is spliced with the output of the previously downsampled convolutional block for the operation of the latter module.

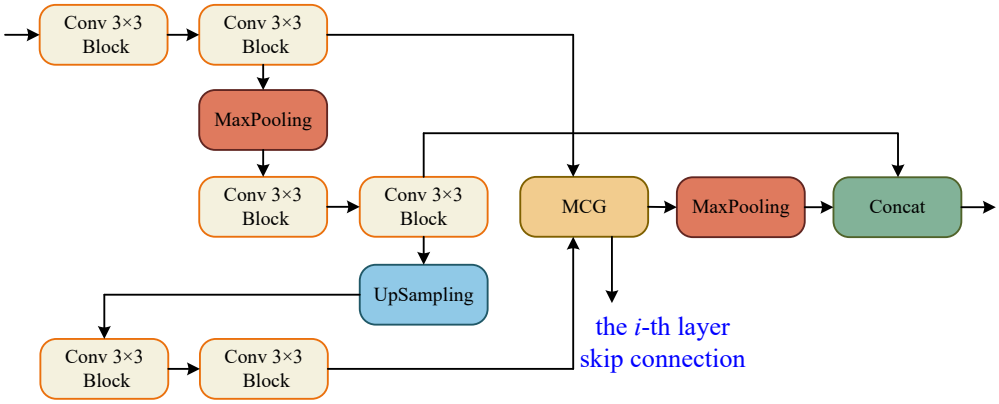

Figure 2. Encoder Conv-MLP (ECM) module

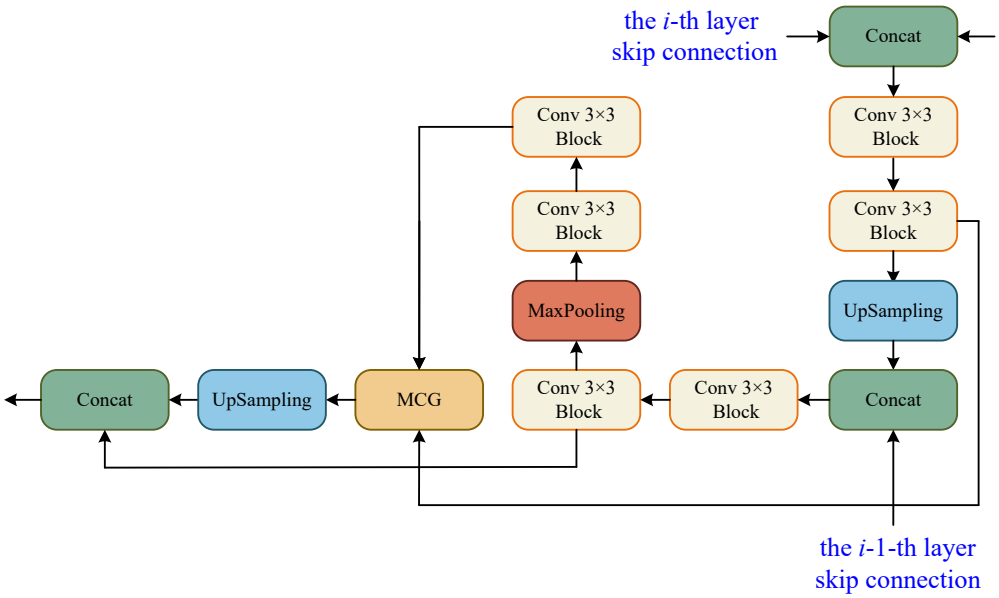

Figure 3. Decoder Conv-MLP (DCM) module

In the DCM, to maximize the restoration of image resolution, we use a convolutional block upsampling and downsampling operation to obtain more information at a larger resolution from the previous layer. This design makes the current layer more forward-looking and can perceive more detailed information. Meanwhile, we have a convolution block in each layer that is spliced with the MCG block of the decoder, and each DCM module will be fused with two decoders of different layers to recover more information. As in the coder design, two convolutional blocks from the same layer are sent to the MCG module for

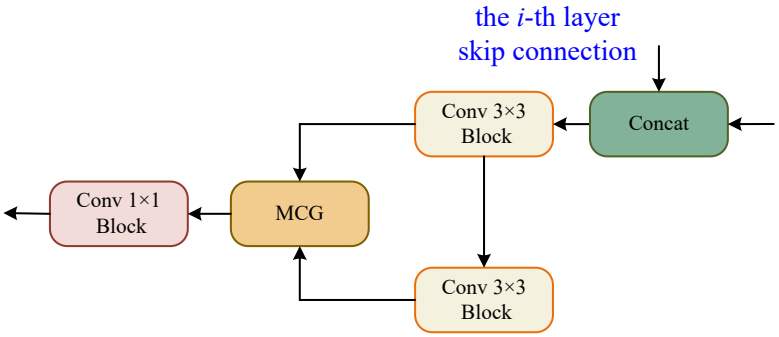

Figure 4. Output Conv-MLP (OCM) module

further cross-fusion. The output of this module is upsampled and then spliced with the previous upsampled convolutional block output in the channel dimension for the next operation. In the BCM block, we follow the same design as DCM. The only difference is that the skip connection of the i-th layer is missing, but the skip connection of the i-1-th layer still retains, that is, the information has been gradually restored at the bottleneck layer, as shown in Fig. 3.

In the OCM module, the two convolutional blocks are connected serially and their output is fed to the MCG module for information fusion, followed by a 1×1 convolutional block yields the final segmentation result, as shown in Fig. 4.

MLP Cross Gating

In order to effectively integrate the information from convolution and MLP, we used the MLP Cross Gating (MCG) module, as shown in Fig. 5. This module can well preserve the ability of convolution operation to extract local and fine features, and also extend the feature map to the global receptive field. MCG receives two inputs, which are the outputs of two different convolution blocks. Let the two inputs be U_1 and V_1, $U_1, V_1 \in \mathbb{R}^{H\times W\times C}$, then U_2 and V_1 can be obtained as:

$$U_2 = \sigma(Ds(LN(U_1))) \tag{1}$$

$$V_2 = \sigma(Ds(LN(V_1))) \tag{2}$$

where LN is the Layer Normalization, Ds is the Dense layers, and σ is the GELU activation. Next, U_2 and V_2 enter the Multi-axis and Multi-windows MLP modules, respectively, and the data is divided into four channels for processing, as shown in Eq. (3).

$$L_1, L_2, G_1, G_2 = S(\sigma(Ds(LN(X)))) \tag{3}$$

where S is the Split 4 heads operation for the different axes. The output end is concatenated according to Eq. (4).

$$M(X) = Do(Ds([Block_1(L_1), Block_2(L_2), Global_1(G_1), Global_2(G_2)])) \tag{4}$$

where $[\cdot,\cdot,\cdot,\cdot]$ is the concatenation. $M(X)$ denotes the output of the Multi-axis and Multi-windows MLP modules and which is used for the fusion of the two paths:

$$U_3 = U_2 \odot M(V_2) \tag{5}$$

$$V_3 = V_2 \odot M(U_2) \tag{6}$$

where $\odot$ represents corresponding multiplication of corresponding elements, and obtains the output of each path through the residual connection with the input:

$$U_4 = U_1 + (\rho(Ds(U_3))) \tag{7}$$

$$V_4 = V_1 + (\rho(Ds(V_3))) \tag{8}$$

where ρ is the Dropout activation. To meet the system requirements and obtain an output of the system, we add the two paths together.

$$M = U_4 + V_4 \tag{9}$$

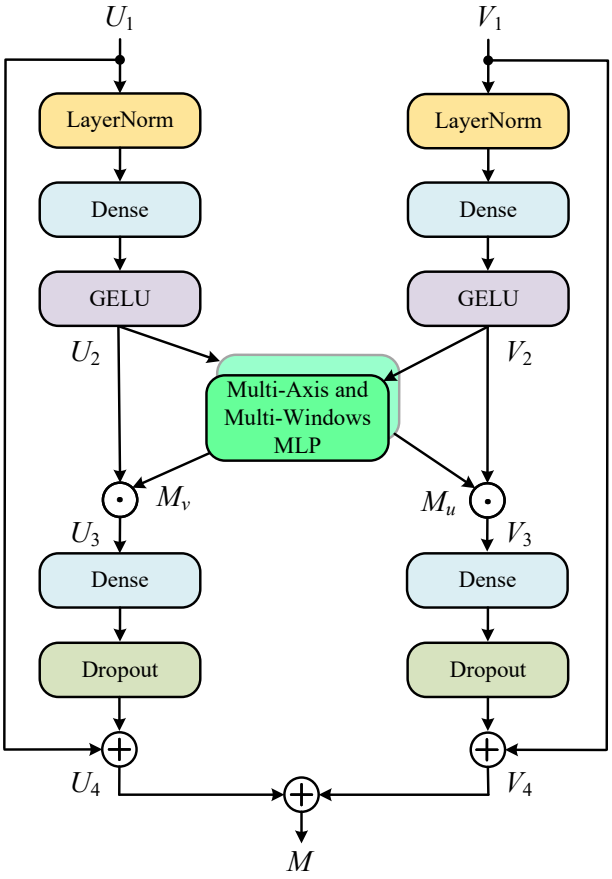

Figure 5. MLP Cross Gating (MCG)

Multi-axis and Multi-windows MLP

The Multi-axis and Multi-windows MLP (MsM) module is inspired by Zhao et al. (2021), which proposes a sparse self-attention operation. The different forms of sparse self-attention on the two axes of the block image are performed, which can capture local and global information in parallel. However, the test images input by this module need a fixed size, which is easy to cause boundaries blur or artifacts, and is not friendly to model training for large-sized images. Based on Tu et al. (2022), we proposed the Multi-axis and Multi-windows MLP framework, which established a basic framework for 4-axis feature map information fusion, as shown in Fig. 6.

MsM module divides the input feature map into four heads by channel, and uses MLP in each head to fuse the corresponding information, two of which are sent to the local branch and the other two to the global branch. In the local branch, two heads are fed into two axes, each with an input tensor of $(H, W, \frac{C}{4})$. Meanwhile one of the axes is divided into $(\frac{H}{b}) \times (\frac{W}{b})$ non-overlapping patches by window size $[b, b]$, and the tensor of $(\frac{H}{b} \times \frac{W}{b}, b \times b, \frac{C}{4})$ is obtained by blocking; the other axis is divided into $(fracH2b \times fracW2b)$ non-overlapping patches by window size $[2b, 2b]$, and the tensor of $(\frac{H}{2b} \times \frac{W}{2b}, 2b \times 2b, \frac{C}{4})$ is obtained by blocking. In the global branch, one of the axes uses a $[b, b]$ grid division to get window of $(\frac{H}{b}, \frac{W}{b})$ and the tensor of $(b \times b, \frac{H}{b} \times \frac{W}{b}, \frac{C}{4})$ obtained by griding; similarly the other axis uses a $[2b, 2b]$ grid division to get window of $(\frac{H}{2b}) \times (\frac{W}{2b})$ and the tensor of $(2b \times 2b, \frac{H}{2b} \times \frac{W}{2b}, \frac{C}{4})$ obtained by griding.

In Fig. 6, we set $b = 2$ as an example, with the top half representing the local branch and the bottom half representing the global branch. As shown in Fig. 6, the input feature map is divided into 4-axis evenly by channel, where the one half goes into the local header and another half goes into the global header. In both local and global branches, different size block is used to divide and MLP is used for different levels of information mixing. Local branches correspond to local blending, global branches correspond to global blending, and the different colored squares in the diagram represent the degree of information blending

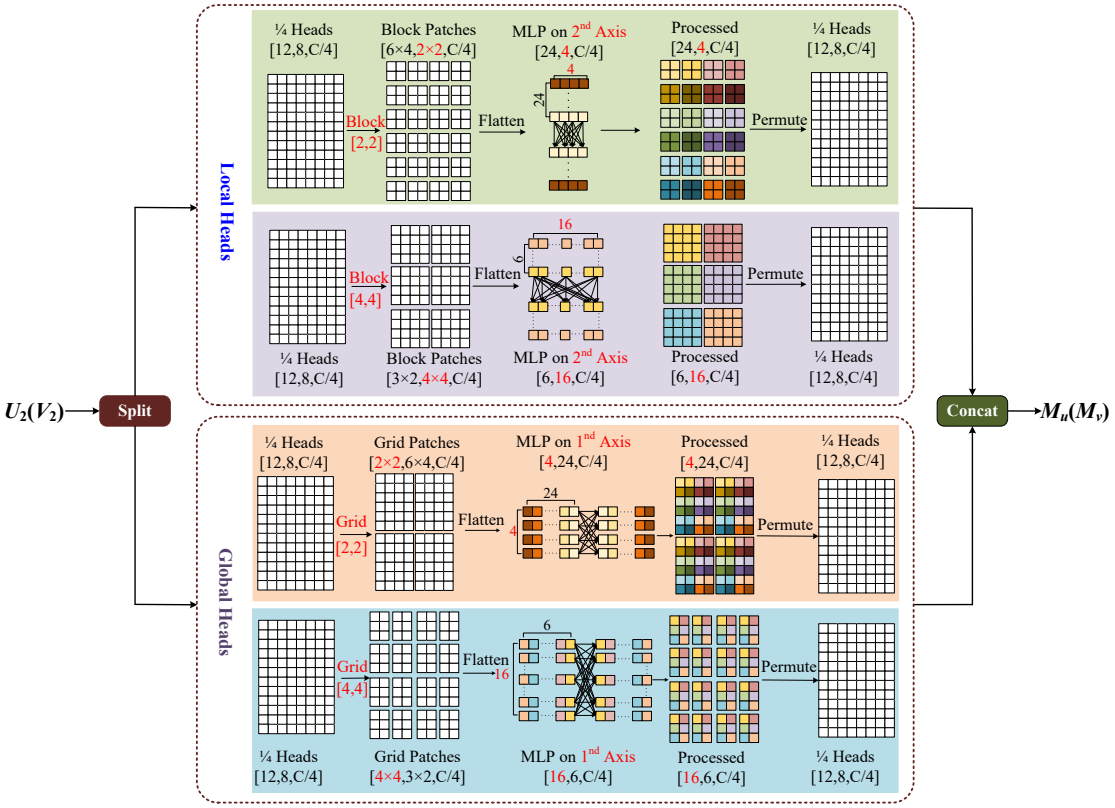

Figure 6. Multi-axis and Multi-windows MLP (MsM)

within the windows. The final processed 4-axis is concated and output. The module is designed to be completely end-to-end from input to output, without the need for specific size and cropping of the images, which will not have any negative effects due to cropping, and the complexity of the model is linear.

EXPERIMENTS

Datasets

The Breast Ultra Sound Images (BUSI) (Al-Dhabyani et al., 2020) is a medical images dataset of breast cancer by ultrasound scans. The BUSI dataset collected included breast ultrasound images of women aged 25 to 75 collected in 2018. The number of patients is 600 women. The dataset consists of 780 images, all of which are cropped to different sizes to remove unused and unimportant borders from the images. The images are in PNG format and divided into three categories: normal, benign and malignant. Each image has its ground truth (masked image). We use the benign and malignant images, a total of 647 images are adjusted to 256×256 RGB images.

The International Skin Imaging Collaboration (ISIC 2018) (Codella et al., 2019) is the world's largest skin image analysis challenge and has organized the world's largest public dermoscope image library. There are 2594 images in the dataset, which contain three different categories, including 20.0% melanoma, 72.0% nevus, and 8.0% seborrheic keratosis. The dataset consists of images of various resolutions, and we adapt all images to 512×512 RGB images.

The CVC-ClinicDB (Tajbakhsh et al., 2015) is a data set of colonoscopy images composed of endoscope images. These images are extracted from the video sequence of the colonoscopy. The Ground Truth image consists of a mask corresponding to the area covered by the polyp in the image. So the segmentation task only considers the images of polyps, a total of 612 images from 31 colonoscopy sequences were obtained. We resize all images to 256×256 RGB images.

Implementation details

All experiments were conducted using an RTX3090 (24GB) graphics card and implemented with the Python 3.8 framework. The batch size was set to 8, and a total of 100 epochs were run. The dataset was randomly combined into training and validation sets with a ratio of 7:3, in order to increase the diversity of training samples and improve the generalization ability of the model. Data augmentation was also applied, including random image rotation, hue and brightness adjustments, and cropping.

MCNMF-Unet employs a loss function that combines Binary Cross-Entropy and Dice coefficients. The loss function is described as:

$$BCELoss(Y,\hat{Y}) = -\sum_{i=1}^{N} Y(x_i) \cdot log\hat{Y}(x_i) \tag{10}$$

$$DiceLoss(Y,\hat{Y}) = 1 - \frac{2 \cdot |Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \tag{11}$$

$$L = 0.5BCELoss(Y,\hat{Y}) + DiceLoss(Y,\hat{Y}) \tag{12}$$

In order to evaluate the performance of our proposed framework relative to the baseline approach, we use the F1_score coefficient and the IOU coefficient as evaluation metrics. The F1-score is a measure of classification problems, and it is often used as the final measure in binary classification or multi-classification problems. F1_score is harmonic average of accuracy rate and recall rate, with a maximum value of 1 and a minimum value of 0. F1_score is defined as shown in equation (13):

$$F_1 = 2 \cdot \frac{Y \cdot \hat{Y}}{Y + \hat{Y}} \tag{13}$$

The Intersection over Union (IoU) score is a standard metric for the performance of object segmentation problems, and its definition is shown in equation (14). Given a set of images, the IoU measure provides the similarity between the predicted and ground truth regions of the objects in the set of images.

$$IoU = \frac{|Y \cap \hat{Y}|}{|Y| \cup |\hat{Y}|} \tag{14}$$

Verification of model performance

The MCNMF-Unet is compared with existing models in this section to verify the effectiveness of the method. We chose classical medical image segmentation algorithms and the more popular models nowadays for comparisons, such as Unet (Ronneberger et al., 2015), Unet++ (Zhou et al., 2019), Res-Unet (Zhang et al., 2018), Attention-Unet (Oktay et al., 2018), MultiResUnet (Ibtehaz and Rahman, 2020), and Trans-Unet (Chen et al., 2021b).

The evaluation of the MCNMF-Unet model is mainly compared in terms of the number of parameters, computational complexity and segmentation accuracy. Table 1 shows the comparison results. Compared with the base line model (Unet), Params only increased by 2.12M, and GFLOPs decreased by 2.04. It is worth noting that compared with the most popular TransUnet, Params is more than three times smaller with a reduction of 72.07M, which greatly reduces the memory usage during the training process and reduces the computational complexity. In terms of validating segmentation performance, we compare IoU and F1_scores in three public datasets: BUSI, ISIC2018 and CVC-ClinicDB.

MCNMF-Unet shows the best performance on all three datasets. In the BUSI dataset, the IoU and F1_scores of MCNMF-Unet reached 70.59% and 81.60%, respectively. Compared with the results of the other six models, the IoU coefficients increased by 3.67% $\sim$ 8.68%. the F1_scores increased by 2.30% $\sim$ 6.58%. In the ISIC 2018 dataset, MCNMF-Unet achieved IoU and F1_scores of 80.66% and 89.06% respectively, and improved the IoU coefficients by 0.15% $\sim$ 7.97% compared to the other six models. the F1_score improved by 0.15% $\sim$ 5.34%. In the CVC-ClinicDB dataset, the IoU and F1_scores of 84.72% and 91.39%, respectively. The IoU coefficients improved by 1.02% $\sim$ 7.02%. The F1_score improved by 0.62% $\sim$ 4.64%. The comparative results indicate the better segmentation performance of MCNMF-Unet.

As the results of the evaluation metrics in Table 1, MCNMF-Unet achieves the best results, but visual observations are needed to determine whether the proposed model works as expected. To this end,

Table 1. Performance comparison of different networks in BUSI, ISIC2018 and CVC-ClinicDB dataset.

Networks	Params (in MB)	GFLOPs	BUSI		ISIC 2018		CVC-ClinicDB	
			IoU(%)	F1_score(%)	IoU(%)	F1_score(%)	IoU(%)	F1_score(%)
Unet	31.13	55.84	62.05	76.02	72.69	83.72	83.32	90.59
Unet++	9.16	34.75	62.72	76.43	74.24	84.76	83.27	90.55
Res-Unet	62.74	94.56	64.82	77.47	73.16	84.07	82.57	90.12
Attention-Unet	51.99	56.95	65.25	77.50	75.10	85.38	83.70	90.77
MultiResUnet	**7.25**	**18.60**	61.91	75.02	74.71	85.07	77.70	86.75
Trans-Unet	105.32	38.52	66.92	79.30	80.51	88.91	82.63	90.19
MCNMF-Unet	33.25	53.80	**70.59**	**81.60**	**80.66**	**89.06**	**84.72**	**91.39**

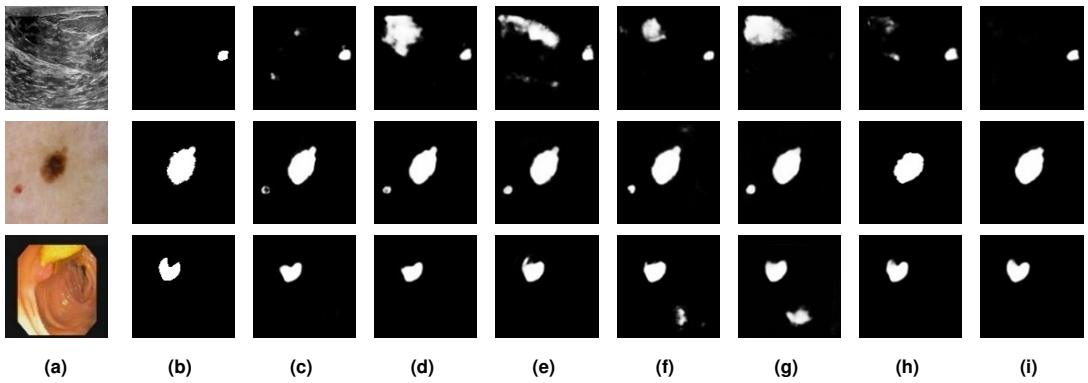

Figure 7. Qualitative comparisons. Row 1: ISIC dataset, Row 2: BUSI dataset, Row 3: CVC-ClinicDB. (a)Input, (b)Ground Truth, (c)Unet, (d)Unet++, (e)Res-Unet, (f)Attention-Unet,(g)MultiResUnet, (h)Trans-Unet, (i)MCNMF-Unet.

In Fig. 7, we also give examples of visual comparisons of the segmentation in BUSI, ISIC 2018 and CVC-ClinicDB.

MCNMF-Unet obtains better results compared to other more popular networks. These visual results show that MCNMF-Unet can successfully recover finer segmentation details. For complex scenes, it is not prone to unexpected segmentation results.

Ablation experiments

We conduct extensive ablation experiments to verify the performance of MCNMF-Unet. All experimental evaluations are performed on the BUSI dataset, and the experimental parameter settings are described in Section 4.2. The experiment verifies the superiority of our proposed module while controlling other variables the same.

Importance of MCG

In the experiment, the MCG module is replaced with convolution and attention modules(replacing the corresponding modules in Fig. 2), and the details are shown, including the Params and GFLOPs after replacing these modules, as well as the IoU and F1_scores on the BUSI dataset. We mainly used the convolution module (Conv), Position Attention Module (PAM), Channel Attention Module (CAM), Convolutional Block Attention Module (CBAM) and Coordinate Attention Module (CoordAM). The data in Table 2 shows that the MCG module still achieves the best segmentation performance while keeping the learnable parameters of the model and the lowest computational complexity.

Effects of Multi-axis and Multi-windows MLP approach

The experiments also further explored the impact of the proposed Multi-axis and Multi-windows MLP approach. As shown in Table 3, we mainly made changes to the Multi-axis and Multi-windows MLP module (Fig. 3) and conducted comparative experiments between the single-axis Block branch (Block), the single-axis Grid branch (Grid), the single-axis Block followed by Gride serial branch (Block-Grid),

Table 2. The impact of MCG modules on network performance.

Variant	Params	GFLOPs	IoU(%)	F1_score(%)
Conv	38.43	57.79	68.47	79.0
CAM	38.43	58.86	69.11	79.32
PAM	38.98	257.21	69.28	79.51
CBAM	34.62	52.94	69.12	79.95
Coord.AM	34.96	52.94	68.98	79.78
MCG(Ours)	**33.25**	**51.80**	**70.54**	**81.53**

the single-axis Grid followed by Block branch (Grid-Block), the two-axis Block-Grid branch(2-Axis) and Multi-axis and Multi-windows MLP (Ours). Through these experiments, it can be seen that the Multi-axis and Multi-windows MLP approach achieves the optimal segmentation results, with the IoU and F1_score improvement reaching 1.07%-1.8% and 1.04%-1.75%, respectively.

Table 3. Effect of Multi-axis and Multi-windows MLP on network performance

Variant	IOU(%)	F1_score(%)
Block	68.86	79.96
Grid	69.28	80.42
Block-Grid	69.14	80.05
Grid-Block	68.72	79.78
2-Axis	69.47	80.49
MsM(Ours)	**70.54**	**81.53**

CONCLUSIONS

In this paper, we design a novel deep neural network architecture based on U-shaped structure for medical image segmentation. The core idea is to fuse the advantages of CNN and MLP into each encoder and decoder layer, while capturing feature map information from multiple windows and multiple dimensions, and always having a global receptive field. Our proposed MCNMF-Unet is essentially an improved network based on Unet, which can be effectively modeled with a small amount of parameters and computation. The experimental results show that MCNMF-Unet outperforms the state-of-the-art baseline on various benchmarks for the segmentation of biomedical images from different types. Moreover, our network achieves very excellent results for image segmentation with complex backgrounds. For future work, we plan to explore more scientific and effective ways of information fusion between convolution and MLP, and extend our method to 3D medical image segmentation.

REFERENCES

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

Azad, R., Arimond, R., Aghdam, E. K., Kazerouni, A., and Merhof, D. (2022). Dae-former: Dual attention-guided efficient transformer for medical image segmentation. *arXiv preprint arXiv:2212.13504*.

Azad, R., Jia, Y., Aghdam, E. K., Cohen-Adad, J., and Merhof, D. (2023). Enhancing medical image segmentation with transception: A multi-scale feature fusion approach. *arXiv preprint arXiv:2301.10847*.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2023). Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., and Gao, W. (2021a). Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021b).

Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, S., Xie, E., Ge, C., Liang, D., and Luo, P. (2021c). Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*.

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

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

Dalmaz, O., Yurt, M., and Çukur, T. (2022). Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614.

Ding, X., Chen, H., Zhang, X., Han, J., and Ding, G. (2022a). Repmlpnet: Hierarchical vision mlp with re-parameterized locality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587.

Ding, Y., Zhang, Z., Zhao, X., Hong, D., Cai, W., Yu, C., Yang, N., and Cai, W. (2022b). Multi-feature fusion: graph neural network and cnn combining for hyperspectral image classification. *Neurocomputing*, 501:246–257.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., and Sagun, L. (2021). Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. (2021). Transformer in transformer. *arXiv e-prints*, pages arXiv–2103.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E. K., Cohen-Adad, J., and Merhof, D. (2023). Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6202–6212.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Huang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X., Chen, Y.-W., and Tong, R. (2022). Scaleformer: Revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. *arXiv preprint arXiv:2207.14552*.

Huang, X., Deng, Z., Li, D., and Yuan, X. (2021). Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*.

Ibtehaz, N. and Rahman, M. S. (2020). Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87.

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

Jin, Q., Meng, Z., Sun, C., Cui, H., and Su, R. (2020). Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. *Frontiers in Bioengineering and Biotechnology*, 8:1471.

Kadry, S., Rajinikanth, V., Taniar, D., Damaševičius, R., and Valencia, X. P. B. (2022). Automated segmentation of leukocyte from hematological images—a study using various cnn schemes. *The Journal of Supercomputing*, 78:6974–6994.

Kalake, L., Dong, Y., Wan, W., and Hou, L. (2022). Enhancing detection quality rate with a combined

hog and cnn for real-time multiple object tracking across non-overlapping multiple cameras. *Sensors*, 22(6):2123.

Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., and Goh, R. (2021a). Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv:2105.09511*.

Li, Y., Jin, P., Yang, F., Liu, C., Yang, M.-H., and Milanfar, P. (2021b). Comisr: Compression-informed video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2543–2552.

Liu, H., Dai, Z., So, D. R., and Le, Q. V. (2021a). Pay attention to mlps. *arXiv e-prints*, pages arXiv–2105.

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

Ni, J., Wu, J., Elazab, A., Tong, J., and Chen, Z. (2022). Dnl-net: deformed non-local neural network for blood vessel segmentation. *BMC Medical Imaging*, 22(1):1–14.

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

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*, pages 234–241. Springer.

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

Sun, Y., Su, L., Luo, Y., Meng, H., Li, W., Zhang, Z., Wang, P., and Zhang, W. (2022). Global mask r-cnn for marine ship instance segmentation. *Neurocomputing*, 480:257–270.

Tajbakhsh, N., Gurudu, S. R., and Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *arXiv e-prints*, pages arXiv–2105.

Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W. E., and Willsky, A. (2003). A shape-based approach to the segmentation of medical imagery using level sets. *IEEE transactions on medical imaging*, 22(2):137–154.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. (2022). Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780.

Valanarasu, J. M. J. and Patel, V. M. (2022). Unext: Mlp-based rapid medical image segmentation network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 23–33. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv e-prints*, pages arXiv–1706.

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

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. (2021a). Transbts: Multimodal brain tumor segmentation using transformer. 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 109–119. Springer.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021b). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.

Wu, J., Fu, R., Fang, H., Zhang, Y., and Xu, Y. (2023). Medsegdiff-v2: Diffusion based medical image segmentation with transformer. *arXiv preprint arXiv:2301.11798*.

Xie, X., Pan, X., Zhang, W., and An, J. (2022). A context hierarchical integrated network for medical image segmentation. *Computers and Electrical Engineering*, 101:108029.

Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). Cotr: Efficiently bridging cnn and transformer for 3d 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 III 24*, pages 171–180. Springer.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. (2021). Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14816–14826.

Zhang, F., Liu, H., Cao, C., Cai, Q., and Zhang, D. (2022). Rvlsm: Robust variational level set method for image segmentation with intensity inhomogeneity and high noise. *Information sciences*, 596:439–459.

Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

Zhao, L., Zhang, Z., Chen, T., Metaxas, D. N., and Zhang, H. (2021). Improved transformer for high-resolution gans. *arXiv e-prints*, pages arXiv–2106.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890.

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