Rolling-Unet: Revitalizing MLP's Ability to Efficiently Extract Long-Distance Dependencies for Medical Image Segmentation

Yutong Liu, Haijiang Zhu*, Mengting Liu, Huaiyuan Yu, Zihan Chen, Jie Gao

Beijing University of Chemical Technology, China
2021400211@buct.edu.cn, zhuhj@mail.buct.edu.cn, 2023400221@buct.edu.cn, joneyu1@163.com, chenzihan0484@163.com, 2021210535@mail.buct.edu.cn

Abstract

Medical image segmentation methods based on deep learning network are mainly divided into CNN and Transformer. However, CNN struggles to capture long-distance dependencies, while Transformer suffers from high computational complexity and poor local feature learning. To efficiently extract and fuse local features and long-range dependencies, this paper proposes Rolling-Unet, which is a CNN model combined with MLP. Specifically, we propose the core R-MLP module, which is responsible for learning the long-distance dependency in a single direction of the whole image. By controlling and combining R-MLP modules in different directions, OR-MLP and DOR-MLP modules are formed to capture long-distance dependencies in multiple directions. Further, Lo2 block is proposed to encode both local context information and long-distance dependencies without excessive computational burden. Lo2 block has the same parameter size and computational complexity as a 3×3 convolution. The experimental results on four public datasets show that Rolling-Unet achieves superior performance compared to the state-of-the-art methods.

Introduction

With the rapid development of computer technology and artificial intelligence, the powerful modeling ability of Convolutional Neural Network (CNN) has been widely studied. Deep learning-based segmentation algorithms have also been introduced into medical image. U-Net (Ronneberger, Fischer, and Brox 2015) is one of the most famous network architectures in the field of medical image segmentation, and it is a fully convolutional segmentation network. U-Net's encoder and decoder are symmetrical, forming a U-shaped segment, and fusing feature maps from different stages through skip connections. U-Net can adapt to small training sets and output more accurate segmentation results. This advantage makes U-Net a huge success and widely used. Following this technical route, such as UNet++ (Zhou et al. 2018), Att-UNet (Oktay et al. 2018), 3D U-Net (Çiçek et al. 2016) and V-Net (Milletari, Navab, and Ahmadi 2016) have been developed for image and volume segmentation of various medical imaging modalities. Although these methods perform well, due to the inherent locality of convolution operations, pure CNN architectures are difficult to learn clear global and remote semantic information (Chen et al. 2021).

To overcome the limitations of CNN, inspired by the great success of Transformer in the natural language processing (NLP) domain, researchers have tried to introduce Transformer into the vision domain (Carion et al. 2020). Vision Transformer (ViT) (Dosovitskiy et al. 2020) is completely based on multi-head self-attention mechanism, which enables the network to capture remote dependencies and encode shape representations. However, it requires a large amount of training data to achieve good performance. Moreover, it has high computational complexity, which prevents the network from supporting high-resolution input (Azad et al. 2022). Swin Transformer (Liu et al. 2021) reduces the computation, but at the cost of no information interaction between its windows, resulting in a smaller receptive field. Compared with CNN models, pure Transformer models also perform poorly in capturing local representations (Chen et al. 2021). In view of the characteristics of CNN and Transformer, some methods attempt to combine CNN and Transformer (Chen et al. 2021; Valanarasu et al. 2021; Wang et al. 2022) to further enhance the network's ability. But these methods still cannot balance the performance and computational cost well.

Multilayer perceptron (MLP) or fully connected (FC) is the earliest type of neural network, which consists of multiple linear layers and nonlinear activations stacked together (Rosenblatt 1957). Theoretically, MLP is a universal approximator (Pinkus 1999). However, MLP has large computation and is prone to overfitting when data is insufficient. Moreover, input flattening limits the input resolution. Due to the limitations of hardware and available datasets at that time, the development of MLP was not smooth. In 2021, MLP-Mixer (Tolstikhin et al. 2021) revived the vitality of MLP. It mainly consists of two modules: Token-Mixing MLP and Channel-Mixing MLP, which achieve competitive performance without convolution and attention. MLP has a small inductive bias, and on large datasets, pure MLP architectures can better extract global semantic information. But this also makes it perform poorly on small datasets. To achieve better performance, local bias was introduced (Hou et al. 2022; Tang et al. 2022; Yu et al. 2022; Lian et al. 2021). But they lost sight of the global aspect.

*Corresponding author.

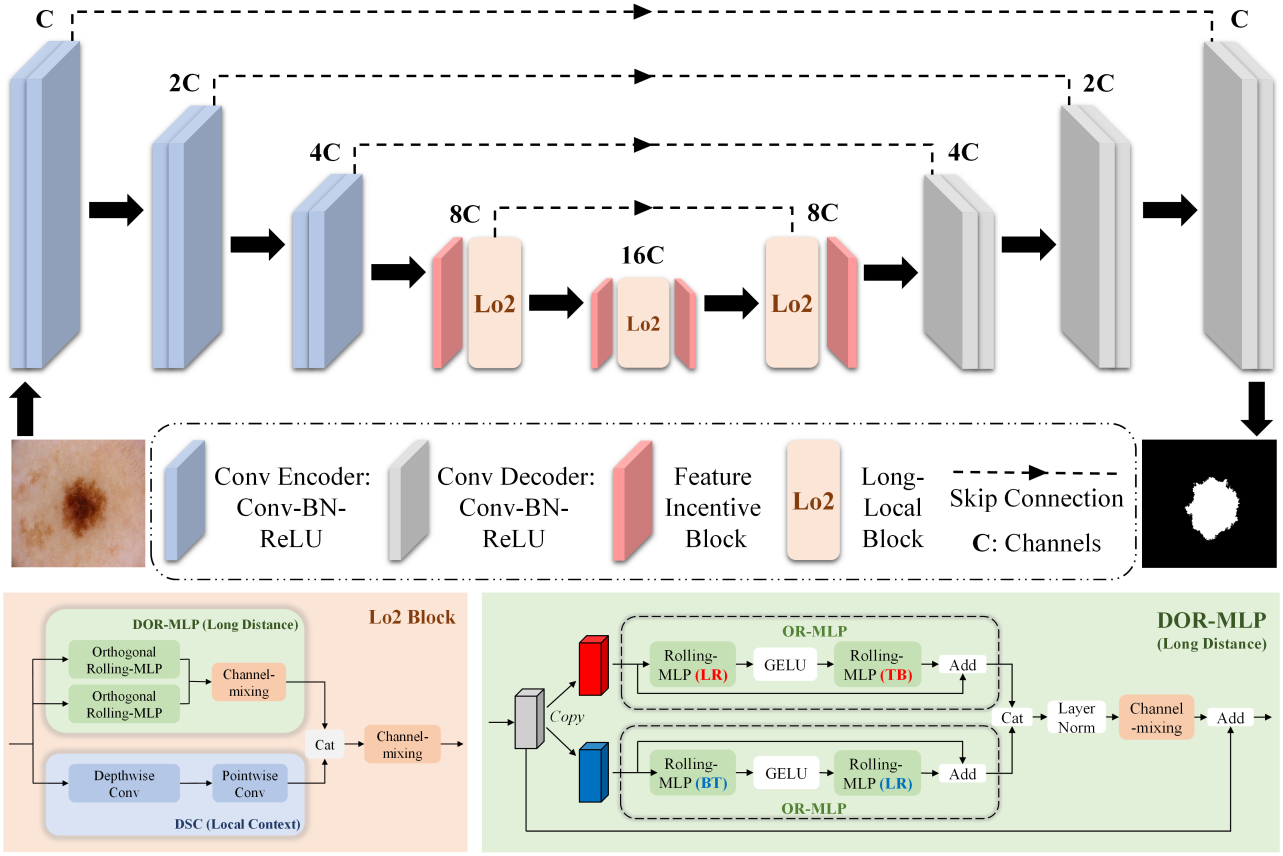


Figure 1: The overview of the proposed Rolling-Unet.

How to capture and fuse local features and long-distance dependencies more effectively is the key to achieve accurate medical image segmentation. In this paper, we rethink this topic: besides combining CNN and Transformer, are there any other methods that can have both local information and long-distance dependencies? The answer is yes. By combining CNN and MLP, this paper proposes a medical image segmentation network named Rolling-Unet. Its core is the flexible Rolling-MLP (R-MLP) module, which can capture linear long-distance dependency in a single direction of the whole image. By concatenating two vertical R-MLP modules, we form the Orthogonal Rolling-MLP (OR-MLP) module, which can capture remote dependencies in multiple directions. We adopt the U-shaped framework of U-Net, including the encoder-decoder structure, bottleneck layer and skip connections, to preserve the fine spatial details. In the 4th layer of the encoder-decoder and the bottleneck layer, we replace the original convolution block with Feature Incentive block and Long-Local (Lo2) block. The Feature Incentive block encodes features and controls the dimension and shape of feature output. Lo2 block consists of Double Orthogonal Rolling-MLP (DOR-MLP contains two complementary OR-MLP) module and Depthwise Separable Convolution (DSC) module, which capture both local context information and long-distance dependencies relationship of the image. Extensive experiments show that our method outperforms the existing best methods. The main contributions of this work are:

- 1) We proposed a new approach to capture long-distance dependency, and constructed the R-MLP module.
- 2) Based on 1, we constructed the OR-MLP and DOR-MLP modules, which can obtain remote dependencies in more directions.
- 3) Based on 2, we proposed the Lo2 block. It simultaneously extracts the local context information and long-distance dependencies, without increasing the computational burden. The Lo2 block has the same level of parameters and computation as a 3×3 convolution.
- 4) Based on 3, we constructed Rolling-Unet networks with different parameter scales. On four datasets, all scales of Rolling-Unet surpassed the existing methods, fully verifying the efficiency of our method.

Related Work

CNN and Transformer for Medical Image Segmentation

Inspired by U-Net, UNet++ (Zhou et al. 2018) incorporated a set of dense skip connections in the model to alleviate the semantic gap of feature fusion. Several subsequent works leveraged techniques such as attention mechanism, image

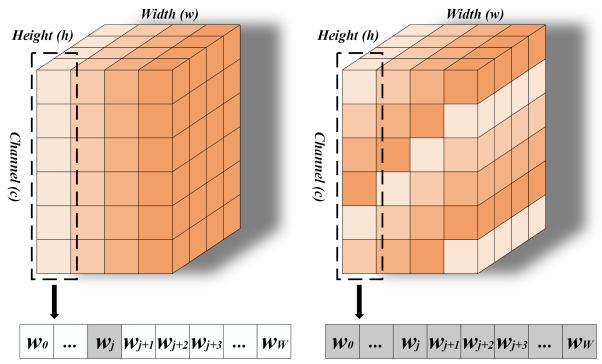


Figure 2: Illustration of the Rolling operation in width direction.

pyramid, and residual architecture (Oktay et al. 2018; Jha et al. 2020)(Jha et al. 2019) to further enhance the performance of CNN-based models. DconnNet (Yang and Farsiu 2023) is a state-of-the-art CNN-based model that exploits directional features extracted from a shared latent space to enrich the overall data representation. In the medical image domain, pure Transformer-based segmentation paradigms have also emerged: such as MISSFormer (Huang et al. 2021), DAE-Former (Azad et al. 2022), Swin–Unet (Cao et al. 2023). Swin–Unet is the first pure Transformer-based U-shaped architecture that adopts Swin Transformer to boost feature representation. Given the respective drawbacks of CNN and Transformer, various works that integrate both paradigms have been proposed. MedT (Valanarasu et al. 2021) devised a gated axial attention model that tackles the issue of limited data samples in medical image. UCTransNet (Wang et al. 2022) introduced a Transformer-based module to substitute the skip connections in U-Net. Despite these works all embrace the strategy of blending global and local features to augment the model capability, they still fall short of satisfying the demand of accurate segmentation of medical images.

MLP Paradigm for Image Tasks

MLP-Mixer (Tolstikhin et al. 2021) is the pioneer of a deep MLP network for vision. Owing to its inferior performance on small datasets, later works endeavored to incorporate local priors in MLP . Vision Permutator (ViP) (Hou et al. 2022) encodes the feature representation with linear projections along both height and width dimensions. Sparse MLP (Tang et al. 2022) follows a similar strategy, except that it directly maps along the image height and width. However, this design lacks flexibility, as its parameter and computation overheads are tied to the image size, which limits the size of the input image. S2MLP (Yu et al. 2022) devised a spatial shift module, which aligns disparate token features to the same channel. AS-MLP (Lian et al. 2021) employs two parallel branches for horizontal and vertical shifts. Nevertheless, these works merely possess local receptive fields, forsaking the original motivation of pure MLP models to capture global features. In the medical image domain, as far as we know, there are few segmentation models based on MLP.

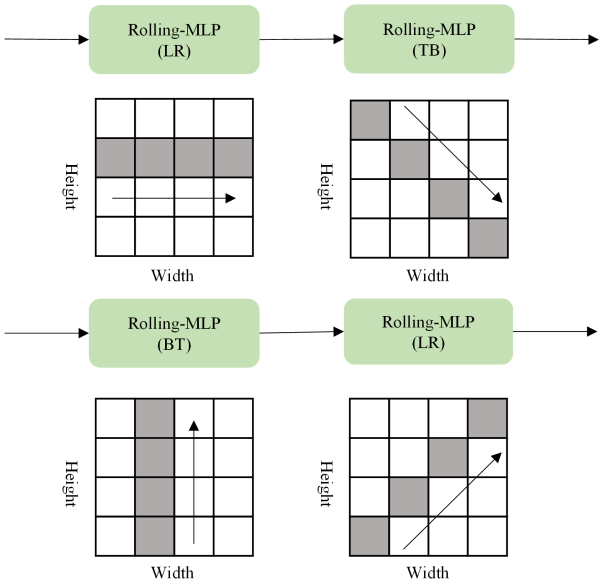


Figure 3: Controlling and combining different R-MLP to obtain long-distance dependencies in multiple directions.

UNext (Valanarasu and Patel 2022) introduced a lightweight model, which adopts an axial shift module, but still can only capture short-range linear receptive fields. PHNet (Lin et al. 2023) is a 3D segmentation network that proposes a multi-layer permutation perceptron module, which augments the primal MLP by preserving positional information.

Method

Architecture Overview

Figure 1 illustrates the overall architecture of the proposed Rolling-Unet, which follows the U-Net design. It consists of an encoder-decoder, a bottleneck layer, and skip connections. The encoder-decoder has four stages of downsampling and upsampling, which are performed by max pooling and bilinear interpolation, respectively. The first three layers of the encoder-decoder contain two standard 3×3 convolution blocks each. The fourth layer and the bottleneck layer employ Feature Incentive blocks to handle feature channel compression and expansion, and Lo2 blocks to capture both local context and long-range dependencies of the image. The skip connections fuse the features of the same scale by addition. Each module is described in detail below.

R-MLP Module

Given a feature matrix $X \in {}^{H\times W\times C}$ with spatial resolution $H \times W$ and channel number C, where $h_i (i \in [1, H])$ denotes the height index, $w_j (j \in [1, W])$ denotes the width index, and $c_k (k \in [1, C])$ denotes the channel index, we perform a Rolling operation on the feature maps of each channel layer in the feature matrix along the same direction, as shown in Figure 2 (taking the width direction as an example). The Rolling operation consists of two steps: shifting and cropping. First, the feature map with channel index c_k

Method	Parmas(M)	BUSI			GlaS		
		IoU ↑	F1 ↑	HD95 ↓	IoU ↑	F1 ↑	HD95 ↓
U-Net(2015)	31.04	64.25±1.63	77.55±1.23	7.57±2.44	87.62±0.29	93.35±0.16	0.83±0.18
UNet++(2018)	36.63	65.68±1.66	78.56±1.26	7.72±2.16	87.99±0.52	93.58±0.29	0.81±0.16
Att-UNet(2018)	34.88	65.97±1.91	78.79±1.29	8.36±2.11	87.90±0.47	93.40±0.26	0.82±0.29
MedT(2021)	1.37	52.15±3.47	67.68±3.18	10.23±1.17	————	————	————
UCTransNet(2022)	66.24	67.27±1.04	79.62±0.74	6.19±0.45	87.80±0.16	93.46±0.12	0.78±0.26
UNeXt(2022)	1.47	61.78±1.46	75.52±0.91	8.33±0.42	83.95±1.09	91.22±0.67	1.04±0.10
DconnNet(2023)	25.49	67.16±0.61	79.63±0.61	6.97±2.81	87.22±0.59	93.12±0.36	0.93±0.15
Rollling-Unet(S)	1.78	65.52±2.82	78.43±2.10	6.19±0.62	86.19±0.35	92.51±0.27	1.00±0.08
Rollling-Unet(M)	7.10	66.99±0.61	79.50±0.35	**5.76±0.95**	86.60±0.82	92.75±0.53	0.90±0.15
Rollling-Unet(L)	28.32	**67.81±1.80**	**80.17±1.19**	7.29±2.50	**88.02±0.28**	**93.59±0.17**	**0.64±0.27**

Table 1: Results on the BUSI and GlaS dataset. The IoU, F1 and HD95 are in 'mean±std' format. The best results are bold.

has a shifting step of k. Then, taking the feature map with channel index c_0 as the reference, we crop the excess parts of the other feature maps to the missing parts. Finally, we perform a channel projection with weight sharing at each spatial location index (h_i, w_j) to encode long-distance dependency. In Figure 2, the original feature matrix has only one width w_j feature at a fixed spatial index (h_i, w_j) for all channels. After applying the Rolling operation in width direction, different channels have different width features. When $C \geq W$, we can encode the width features of the entire image, which can be understood as global, unidirectional, linear receptive fields. When $C < W$, this linear receptive field is non-global. Similarly, R-MLP can also capture long-distance dependency in height direction.

It is well known that MLP is sensitive to the positional information of the input. R-MLP performs cyclic operations of shifting and cropping the feature maps, making the positional index order on each channel non-fixed. This preliminarily reduces the sensitivity of R-MLP to position. Secondly, by using weight sharing, all channel projections share a set of parameters, which further reduces the sensitivity.

OR-MLP and DOR-MLP

R-MLP can encode the long-range dependency along either the width or height direction. How can we capture the long-distance dependency along other direction? By applying R-MLP first along the width direction and then along the height direction, it is equivalent to the synchronous shifting operation of the feature map in two orthogonal directions, resulting in a diagonal receptive field. As shown in equation (1), for an input X, we first apply R-MLP along one direction MLP_R^1, and then concatenate another R-MLP along the perpendicular direction MLP_R^2. We use the GELU activation function in between, and then add a residual connection with the input X. This forms the Orthogonal Rolling-MLP (OR-MLP) module, as illustrated in Figure 1.

$$MLP_{OR}(X) = (MLP_R^2(GELU(MLP_R^1(X)))) + X \quad (1)$$

R-MLP is a highly flexible module with great potential. The sign of the shifting step k determines the encoding order. When using R-MLP alone, reversing the order does not affect the linear receptive field extraction. However, when using OR-MLP, the sign of k is crucial. For the width direction, given a positive k value, it represents moving from left to right (LR), and a negative k value represents moving from right to left (RL). For the height direction, given a positive k value, it represents moving from top to bottom (TB), and a negative k value represents moving from bottom to top (BT). As shown in Figure 3, we consider two complementary OR-MLP modules. The first one applies R-MLP along the LR direction first and then sequentially along the TB direction. The second one applies R-MLP along the BT direction first and then sequentially along the LR direction. By parallelizing these two OR-MLPs, we capture the long-range dependencies along four directions: width, height, positive diagonal, and negative diagonal! As shown in equation (2), for an input X, we first apply an OR-MLP MLP_{OR}^1, and then parallelize another OR-MLP MLP_{OR}^2. We concatenate their outputs along the channel dimension and apply LayerNorm. Then we use Channel-mixing (CM) (Tolstikhin et al. 2021) to fuse the features and reduce the channels back to C. Finally, we add a residual connection with the input X. This forms the Double Orthogonal Rolling-MLP (DOR-MLP) module, as depicted in Figure 1.

$$\begin{aligned} MLP_{DOR}(X) =& CM(LN(Concat[MLP_{OR}^1(X), \\ & MLP_{OR}^2(X)])) + X \end{aligned} \quad (2)$$

Lo2 Block and Feature Incentive Block

The DOR-MLP module captures the global, linear long-range dependencies along four directions in two-dimensional space, but it lacks the local context information. We argue that better integrating local information and global dependencies is crucial for performance improvement. Depthwise Separable Convolution (DSC) is a natural choice (Chollet 2017). Because it has very few parameters and computational costs, which is compatible with DOR-MLP. It is a well-established fact that the Channel-mixing in MLP-Mixer, the MLP in ViT, and the R-MLP in this paper are all equivalent to the standard 1×1 convolution in CNN, which allows feature interaction between different channels.

Method	IoU ↑	F1 ↑	HD95 ↓
U-Net(2015)	82.97	90.39	1.79
UNet++(2018)	83.34	90.66	1.56
Att-UNet(2018)	83.31	90.61	1.69
MedT(2021)	81.48	89.49	1.89
UCTransNet(2022)	83.96	91.02	1.66
UNeXt(2022)	82.90	90.38	2.04
DconnNet(2023)	83.86	90.93	2.04
Rollling-Unet(S)	84.15	**91.13**	**1.51**
Rollling-Unet(M)	**84.16**	91.09	1.69
Rollling-Unet(L)	83.74	90.90	1.99

Table 2: Results on the ISIC 2018 dataset (Image size = 256).

Method	IoU ↑	F1 ↑	HD95 ↓
U-Net(2015)	80.98	89.14	3.37
UNet++(2018)	81.40	89.44	3.59
Att-UNet(2018)	81.45	89.44	2.78
MedT(2021)	——	——	——
UCTransNet(2022)	83.14	90.47	2.89
UNeXt(2022)	83.12	90.51	2.32
DconnNet(2023)	83.60	90.78	2.78
Rollling-Unet(S)	**84.14**	**91.11**	2.17
Rollling-Unet(M)	83.96	90.94	2.42
Rollling-Unet(L)	83.94	90.99	**1.90**

Table 3: Results on the ISIC 2018 dataset (Image size = 512).

The Rolling operation in R-MLP does not involve any parameters or FLOPs, so the parameters of R-MLP is $\mathrm{O}(C^2)$, and the FLOPs is $\mathrm{O}(HWC^2)$. It can be further derived that the parameters and FLOPs of OR-MLP are $\mathrm{O}(2C^2)$ and $\mathrm{O}(2HWC^2)$ respectively, and the parameters and FLOPs of DOR-MLP are $\mathrm{O}(6C^2)$ and $\mathrm{O}(6HWC^2)$ respectively. As depicted in Figure 1, we parallelize DOR-MLP with DSC, and then concatenate their outputs along the channel dimension, and finally use Channel-mixing to fuse the features and restore the channels to C. This forms the Long-Local (Lo2) block, see equation (3). In DSC, we use a 3×3 convolution kernel. Hence, we can derive that the parameters of Lo2 block is $\mathrm{O}(9C^2)$, and the FLOPs is $\mathrm{O}(9HWC^2)$. This is of the same level as a standard 3×3 convolution.

$$Lo2(X) = CM(Concat[MLP_{DOR}(X), DSC(X)]) \quad (3)$$

We employ the Feature Incentive block in the 4th layer of the encoder and the bottleneck layer. It is essentially a convolution block that mainly used to encode the feature and channel number changes. Since subsequent Lo2 block mainly conducts MLP, we adopt GELU activation function and LayerNorm, following a series of prior MLP works. In the 4th layer of the decoder, the Feature Incentive block is composed of a convolution block, RELU activation function and BatchNorm, as subsequent networks conduct convolution operations, following a series of CNN habits.

Method	IoU ↑	F1 ↑	HD95 ↓
U-Net(15)	69.81±0.34	82.22±0.24	1.86±0.15
Att-UNet(18)	69.90±0.41	82.28±0.29	1.80±0.17
UCTransNet(22)	69.13±0.31	81.74±0.22	2.00±0.00
UNeXt(22)	66.81±0.04	80.10±0.03	2.04±0.07
DconnNet(23)	69.90±0.44	82.28±0.31	2.00±0.00
Rollling-Unet(S)	69.40±0.28	81.94±0.19	1.90±0.17
Rollling-Unet(M)	69.55±0.38	82.03±0.27	1.96±0.08
Rollling-Unet(L)	**70.40±0.43**	**82.63±0.30**	**1.71±0.00**

Table 4: Results on the CHASEDB1 dataset. The metrics are in 'mean±std' format.

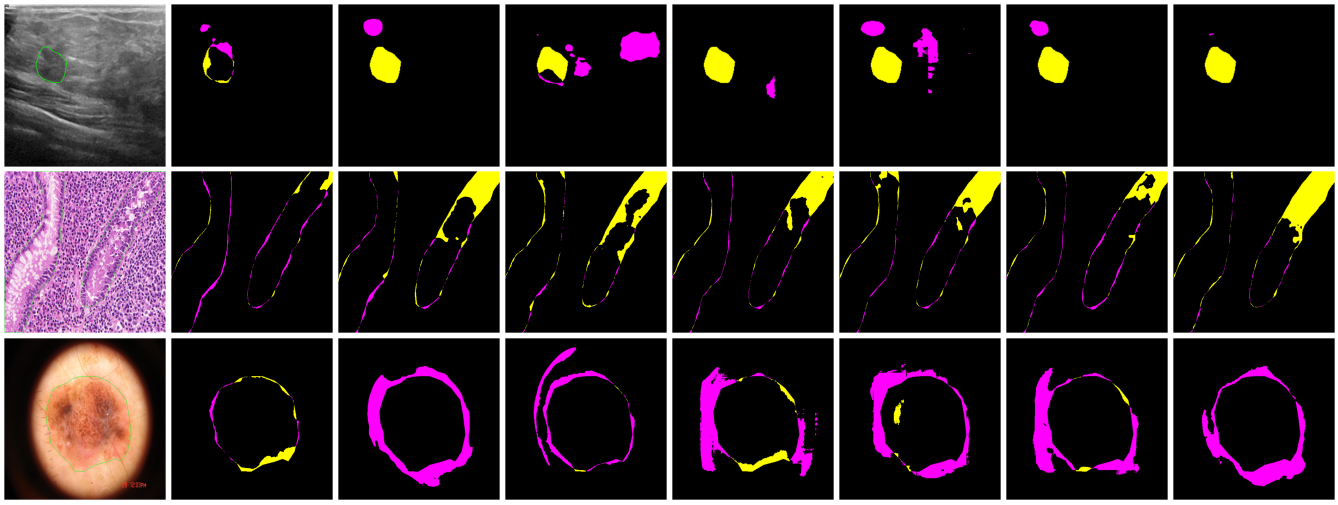
Experiments

Datasets

We evaluated our method on four datasets with different characteristics, data sizes and image resolutions: the International Skin Imaging Collaboration (ISIC 2018), the Breast UltraSound Images (BUSI), the Gland Segmentation dataset (GlaS) and the CHASEDB1. The ISIC 2018 dataset contains skin images acquired by cameras and the corresponding skin lesion segmentation maps. We only used the training set of the ISIC 2018 dataset, which contains 2594 images. The difficulty of this dataset lies in the fact that the segmentation targets often have blurry boundaries, which is exacerbated by the increasing of image size. Therefore, we resized the images to two resolutions of 256×256 and 512×512 and conducted experiments separately. The BUSI dataset consists of ultrasound images of normal, benign, and malignant breast cancer and the corresponding segmentation maps. It has similar problems with the ISIC 2018 dataset, but they have different lesion types and imaging methods. We used 647 ultrasound images of benign and malignant breast tumors, resized to 256×256. The GlaS dataset contains 165 images, which we resized to 512×512. The CHASEDB1 dataset is a vessel segmentation dataset with 28 images of 999×960 resolution. To preserve the details of the thin vessels, we resized the images to 960×960.

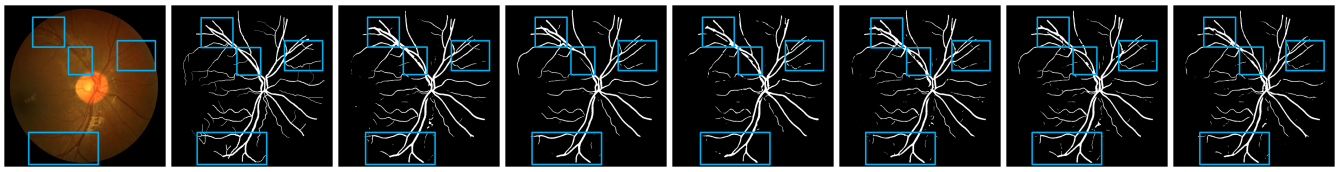
Implementation Details

We implemented Rolling-Unet using Pytorch on a NVIDIA A6000 GPU. For the ISIC 2018, BUSI and GlaS datasets, the batch size was set to 8 and the learning rate was 0.0001 (Valanarasu and Patel 2022). For the CHASEDB1 dataset, the batch size was set to 4 and the learning rate was 0.001 (Tomar et al. 2022). We used the Adam optimizer to train the model, and used a cosine annealing learning rate scheduler with a minimum learning rate of 0.00001. The loss function was a combination of binary cross entropy (BCE) and dice loss. We randomly split each dataset into 80% training and 20% validation subsets. To account for the limited data size of the BUSI, GlaS and CHASEDB1 datasets, we repeated this process three times and reported the average and standard deviation of the results. To evaluate the network's ability fairly, all experiments did not use any pre-trained weights and post-processing methods, and only applied two simple



(a) Image+GT (b) Ours (c) DconnNet (d) UNeXt (e) U-Net (f) Att-UNet (g) UCTransNet (h) UNet++

Figure 4: Qualitative comparison of Rolling-Unet with other state-of-the-art methods. From top to bottom are the BUSI, GlaS and ISIC2018 datasets. The first column is the original image, with the green contour indicating the Ground Truth. In the visualized segmentation results, purple indicates over-segmentation, and yellow indicates under-segmentation



(a) Image (b) GT (c) Ours (d) DconnNet (e) UNeXt (f) U-Net (g) Att-UNet (h) UCTransNet

Figure 5: Qualitative comparison on the CHASEDB1 dataset.

online data augmentations: random rotation and flipping. We trained for 400 epochs in total.

Comparison with State-of-the-Art Method

We evaluated Rolling-Unet against other state-of-the-art methods, including CNN-based methods: U-Net (Ronneberger, Fischer, and Brox 2015), UNet++ (Zhou et al. 2018), Att-Unet (Oktay et al. 2018), DconnNet (Yang and Farsiu 2023); Transformer-based methods: UCTransNet (Wang et al. 2022), MedT (Valanarasu et al. 2021); and MLP-based method: UNeXt (Valanarasu and Patel 2022). MedT failed to produce results on the GlaS, ISIC 2018 (Image size = 512) and CHASEDB1 datasets due to memory constraints. Similarly, UNet++ did not yield results on the CHASEDB1 dataset. To fully demonstrate the efficiency of Rolling-Unet, we trained different sizes of Rolling-Unet. when the channel number C = 16 / 32 / 64 in Figure 1, they are named as Rolling-Unet (S) / Rolling-Unet (M) / Rolling-Unet (L) respectively. We adopted Intersection over Union (IoU), F1 score and 95% Hausdorff Distance (HD95) as evaluation metrics.

The evaluation results on BUSI and GlaS are presented in Table 1. The results on ISIC 2018 are shown in Table 2 and Table 3. The result on CHASEDB1 is shown in Table 4. Our method outperformed all the other methods on all datasets. Especially on BUSI and ISIC 2018, Rolling-Unet obtained a significant advantage. In these two datasets, many targets have blurry boundaries, which make them difficult to distinguish from the background. Rolling-Unet more effectively extracted remote dependencies to enhance the segmentation performance. The experiment of changing the image size on ISIC 2018 further verified this conclusion. Only Rolling-Unet and UNeXt maintained similar performance when the image size increased, while other methods showed different degrees of decline. For the phenomenon that the metrics of Rolling-Unet (X) are lower than those of Rolling-Unet (S) in ISIC 2018, we have two hypotheses. One is the fluctuation of training, which requires taking the average of multiple results to reduce the impact. Another is that the semantic information of this dataset is relatively simple, and more network parameters are prone to overfitting, thereby reducing performance. Recent lightweight models (Valanarasu and Patel 2022; Ruan et al. 2023; Cheng et al. 2023) also reflect this point from the side. In the follow-up work, we will explain this phenomenon through more experiments.

On GlaS and CHASEDB1, no method achieved a significant advantage, but Rolling-Unet was still the best with a small standard deviation. The images in GlaS have dense,

DSC	R-MLP	OR-MLP	DOR-MLP	IoU ↑	F1 ↑	HD95 ↓
				79.48	88.19	4.05
*				80.62	88.94	2.87
	*			81.62	89.50	3.73
*	*			82.11	89.85	3.16
		*		83.39	90.67	2.43
*		*		83.84	90.92	**2.16**
			*	83.46	90.63	2.27
*			*	**84.14**	**91.11**	2.17

Table 5: Ablation experiments on the ISIC 2018 dataset.

Method	IoU ↑	F1 ↑	HD95 ↓
MLP	81.10	89.22	2.70
R-MLP	**84.14**	**91.11**	**2.17**

Table 6: Ablation experiments on the ISIC 2018 dataset.

Method	IoU ↑	F1 ↑	HD95 ↓
Series 1	83.65	90.82	2.72
Series 2	83.62	90.84	**2.05**
Parallel	**84.14**	**91.11**	2.17

Table 7: Ablation experiments on the ISIC 2018 dataset.

tiny cells and tissues; the segmentation targets and the background often have similar textures, colors, as well as shapes. In the CHASEDB1 dataset, the thicker vessels are not difficult for all methods, and the difficulty of segmentation lies in those thin vessels, as shown in Figure 5. These require more powerful methods to solve.

The parameter amounts of the models are provided in Table 1. We define models with parameter amount less than 2M as primary models, and models with parameter amount greater than 20M as secondary models (only Rolling-Unet (L) is between 2-20M). On the four datasets, our method is the best in both primary and secondary models, proving the efficiency of the method.

In Figure 4, we visualized the difference map between the segmentation results and the Ground Truth to highlight the differences. Purple indicates over-segmentation, and yellow indicates under-segmentation. Due to space limitations, we omitted the results of MedT. In the images of BUSI and ISIC 2018, we can see that the segmentation target lacks a clear boundary. In the segmentation results, other methods than Rolling-Unet have generated a lot of under-segmentation or over-segmentation regions. This demonstrates that Rolling-Unet is good at extraction of the target contours. The targets in GlaS have complex boundaries, and only Rolling-Unet achieved segmentation results close to Ground Truth. The visualization results of the CHASEDB1 dataset are shown in Figure 5. Almost all methods can correctly segment the thick vessels, and the subtle difference lies in the thin vessels inside the blue box. Rolling-Unet considered the long-distance dependencies features of the image, so it improved the segmentation effect of the thin vessels.

Ablation Studies

To investigate the impact of various factors on the model performance, we performed ablation experiments on the ISIC 2018 dataset (Image size = 512). The details are described as follows.

Lo2 Block consists of DOR-MLP and DSC modules in parallel. The former is responsible for capturing long-distance dependencies, and the latter is responsible for extracting local context information. To ensure that the combination of DOR-MLP and DSC is optimal, and to explore their respective contributions, the experimental results are shown in Table 5. Regardless of the presence or absence of the DSC module, the performance of R-MLP, OR-MLP, and DOR-MLP progressively increases. This demonstrates the effectiveness of the proposed module for capturing long-distance dependency, and approves the idea of extracting long-distance dependencies from multiple directions. When combined with the DSC module, the performance can be further enhanced. Therefore, it is essential to fuse the remote dependencies and local context information.

To rule out the performance improvement caused by the increase of parameters and FLOPs, we replaced the R-MLP in Rolling-Unet with a regular MLP. This makes the model lose the ability to capture long-distance dependencies while keeping the parameters and FLOPs consistent. As shown in Table 6, the performance dropped significantly. This result is expected, as the Rolling-Unet without the ability to capture long-range dependencies has a similar network structure to the original U-Net.

Further, we explored the combination of DOR-MLP and DSC. Series 1 means executing DOR-MLP first and then DSC. Series 2 means executing DSC first and then DOR-MLP. Parallel means connecting DSC and DOR-MLP in parallel, the two branches are executed concurrently, and the features are integrated by Channel-mixing in the end. The results are shown in Table 7. There is little difference between Series 1 and Series 2, and the best is Parallel. This proves that: the order of extracting local features and remote dependencies is not important, and it is best to fuse them after extracting them simultaneously.

Conclusion

In this paper, we propose Rolling-Unet model that can capture long-range dependencies without increasing the computational cost, and outperform the existing methods. It is worth noting that the remote dependencies from multiple directions are not global receptive fields, they are still a compromise of MLP in a strict sense. However, R-MLP is a very flexible module. By combining it, it can also capture large-scale regional features and even global features. In future work, we will explore this aspect. We will also investigate its potential in three-dimensional medical image segmentation, as well as other image tasks.

Acknowledgements

This research was supported by the National Key R&D Program of China (2022YFF0607503).

References

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*.

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 Karlinsky, L.; Michaeli, T.; and Nishino, K., eds., *Computer Vision – ECCV 2022 Workshops*, 205–218. Cham: Springer Nature Switzerland. ISBN 978-3-031-25066-8.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

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

Cheng, J.; Gao, C.; Wang, F.; and Zhu, M. 2023. SegNetr: Rethinking the local-global interactions and skip connections in U-shaped networks. arXiv:2307.02953.

Chollet, F. 2017. Xception: Deep Learning With Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

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*.

Hou, Q.; Jiang, Z.; Yuan, L.; Cheng, M.-M.; Yan, S.; and Feng, J. 2022. Vision permutator: A permutable mlp-like architecture for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1328–1334.

Huang, X.; Deng, Z.; Li, D.; and Yuan, X. 2021. MISSFormer: An Effective Medical Image Segmentation Transformer. *CoRR*, abs/2109.07162.

Jha, D.; Riegler, M. A.; Johansen, D.; Halvorsen, P.; and Johansen, H. D. 2020. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, 558–564. IEEE.

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

Lian, D.; Yu, Z.; Sun, X.; and Gao, S. 2021. As-mlp: An axial shifted mlp architecture for vision. *arXiv preprint arXiv:2107.08391*.

Lin, Y.; Fang, X.; Zhang, D.; Cheng, K.-T.; and Chen, H. 2023. A Permutable Hybrid Network for Volumetric Medical Image Segmentation. arXiv:2303.13111.

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

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

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*.

Pinkus, A. 1999. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8: 143–195.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Rosenblatt, F. 1957. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Ruan, J.; Xie, M.; Gao, J.; Liu, T.; and Fu, Y. 2023. EGE-UNet: an Efficient Group Enhanced UNet for skin lesion segmentation. arXiv:2307.08473.

Tang, C.; Zhao, Y.; Wang, G.; Luo, C.; Xie, W.; and Zeng, W. 2022. Sparse MLP for image recognition: Is self-attention really necessary? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2344–2351.

Tolstikhin, I. O.; 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. *Advances in neural information processing systems*, 34: 24261–24272.

Tomar, N. K.; Jha, D.; Riegler, M. A.; Johansen, H. D.; Johansen, D.; Rittscher, J.; Halvorsen, P.; and Ali, S. 2022. FANet: A Feedback Attention Network for Improved Biomedical Image Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

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

Valanarasu, J. M. J.; and Patel, V. M. 2022. UNeXt: MLP-Based Rapid Medical Image Segmentation Network. In

Wang, L.; Dou, Q.; Fletcher, P. T.; Speidel, S.; and Li, S., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, 23–33. Cham: Springer Nature Switzerland. ISBN 978-3-031-16443-9.

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

Yang, Z.; and Farsiu, S. 2023. Directional Connectivity-Based Segmentation of Medical Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11525–11535.

Yu, T.; Li, X.; Cai, Y.; Sun, M.; and Li, P. 2022. S2-mlp: Spatial-shift mlp architecture for vision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 297–306.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. 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*, 3–11. Springer.