

MEW-UNET: MULTI-AXIS REPRESENTATION LEARNING IN FREQUENCY DOMAIN FOR MEDICAL IMAGE SEGMENTATION

Jiacheng Ruan, Mingye Xie, Suncheng Xiang*, Ting Liu, Yuzhuo Fu*

Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Recently, Visual Transformer (ViT) has been widely used in various fields of computer vision due to applying self-attention mechanism in the spatial domain to modeling global knowledge. Especially in medical image segmentation (MIS), many works are devoted to combining ViT and CNN, and even some works directly utilize pure ViT-based models. However, recent works improved models in the aspect of spatial domain while ignoring the importance of frequency domain information. Therefore, we propose **Multi-axis External Weights UNet (MEW-UNet)** for MIS based on the U-shape architecture by replacing self-attention in ViT with our Multi-axis External Weights block. Specifically, our block performs a Fourier transform on the three axes of the input feature and assigns the external weight in the frequency domain, which is generated by our Weights Generator. Then, an inverse Fourier transform is performed to change the features back to the spatial domain. We evaluate our model on four datasets and achieve state-of-the-art performances. In particular, on the Synapse dataset, our method outperforms MT-UNet by 10.15mm in terms of HD95. Code is available at <https://github.com/JCruan519/MEW-UNet>.

Index Terms— Medical image segmentation, Deep learning, Multi-axis External Weights

1. INTRODUCTION

Medical image segmentation (MIS) can assist relevant medical staff in locating the lesion area and improve the efficiency of clinical treatment, which has great practical value. In recent years, UNet [1], an encoder-decoder model based on U-shape architecture, has been widely utilized for MIS. Due to its strong scalability, many works are carried out based on the U-shape architecture. For example, UNet++ [2] reduces the semantic difference between the encoder and decoder by introducing the dense connection. Att-UNet [3] introduces a gating mechanism to make the model focus on the targets.

The above improvements are all based on CNNs, and the natural locality of the convolution operation makes net-

works obtain global information poorly. ViT [4], due to its self-attention mechanism (SA), improves the long-range dependency modeling ability, and focuses on holistic image semantic information, which benefits intensive prediction tasks, such as image segmentation. Therefore, recent improvements can be divided into two types. On the one hand, hybrid structures based on CNNs and ViTs are widely used. For example, UCTransNet [5] replaces the skip connection in UNet with the CTrans module, alleviating the problem that the skip connection between the encoder and decoder may lead to incompatible features. MT-UNet [6] uses CNN at a shallow level and Local-Global SA at a deep level, combined with the external attention mechanism, to obtain richer representation information. On the other hand, pure ViTs are conducted in MIS. For instance, Swin-UNet [7] performs better by replacing the convolution operation in UNet with the Swin Transformer Block.

Although the above models have achieved good results, they are all based on the spatial domain, and few works explore MIS in the frequency domain. The frequency domain information can help the model distinguish the lesion area and background more clearly, which is indispensable for MIS. In the general vision, GFNet [8] utilizes 2D discrete Fourier transform (2D DFT) to change features from the spatial domain to the frequency domain, and filters in the frequency domain are used for learning representation, which is enlightening. However, the frequency domain information is only extracted in a single axis, resulting in incomplete global information. In addition, it ignores the importance of local information in image feature extraction.

To address the problem, we propose the Multi-axis External Weights mechanism (MEW) that can simultaneously obtain more comprehensive global and local information. To be specific, the feature map is divided into four parts along the channel dimension. For the first three branches, we transform the features into the frequency domain using 2D DFT along the three different axes. Then, we multiply the frequency domain maps by the corresponding learnable weights to obtain the frequency domain information and global knowledge. In addition, depthwise separable (DW) convolution operation [9] is conducted for the remaining branch to obtain local information. After that, via replacing SA in ViT with our MEW, Multi-axis External Weights Block (MEWB) is obtained. Fi-

* Suncheng Xiang and Yuzhuo Fu are the co-corresponding authors. This work was partially supported by the National Natural Science Foundation of China (Grant No. 61977045).

nally, based on MEWB and U-shape architecture, we present MEW-UNet, a powerful network for MIS.

The contributions of this paper can be summarized as follows: 1) the Multi-Axis External Weights block is proposed to obtain global and local information at the same time, and frequency domain feature signals are introduced to help the model fully understand the context; 2) Based on U-shape and ViTs structure, the SA is replaced by our block, and a powerful MIS model, MEW-UNet is obtained; 3) We have conducted sufficient experiments on four datasets and achieved state-of-the-art results.

2. METHODS

2.1. Multi-axis External Weights Block

In MIS, recent methods focus more on obtaining more abundant information in the spatial domain while ignoring the importance of the frequency domain. For the spatial domain, the boundary between the segmentation object and background is often unclear, while in the frequency domain, the objects are at different frequencies, which is accessible to distinguish [10]. Although [10] and [8] have introduced the idea of the frequency domain, the global information they acquired is not sufficient. Therefore, we propose MEW based on 2D DFT for different axes to obtain more global information in the frequency domain, as shown in Figure 1 (a).

Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$ and C, H, W is the channel, height and width of the map. Firstly, X is divided into four equal parts along the channel dimension. Furthermore, we sent parts into four different branches, respectively. MEW can be expressed by formulas (1) to (4).

$$x_1, x_2, x_3, x_4 = \text{Split}(X) \quad (1)$$

$$x_{i(I,J)} = W_{(I,J)} \odot \mathcal{F}_{(I,J)}[x_i] \quad (2)$$

$$x'_i = \mathcal{F}_{(I,J)}^{-1}[x_{i(I,J)}], x'_4 = \text{DW}(x_4) \quad (3)$$

$$Y = \text{Concat}(x'_1, x'_2, x'_3, x'_4) + X \quad (4)$$

where $i = 1, 2, 3$ corresponds to the first three branches. $W_{(I,J)}$ and $\mathcal{F}_{(I,J)}$ denote the learnable external weight and the 2D DFT for the corresponding axes. When $i = 1$, I and J represent Height-Width axes. When $i = 2$, I and J present Channel-Width axes. When $i = 3$, I and J represent Channel-Height axes. \odot is the element-wise product. $\mathcal{F}_{(I,J)}^{-1}$ refers to the 2D inverse DFT. Split and Concat represent the split and concatenation operation along the channel dimension.

In the first branch, we perform 2D DFT on the Height-Width axes to obtain the feature map in the frequency domain. Afterward, the element-wise product operation is performed between the feature map and the corresponding learnable external weight. Finally, we apply the 2D inverse DFT to change the map to the spatial domain. Similarly, for the second and third branches, we perform the above operations

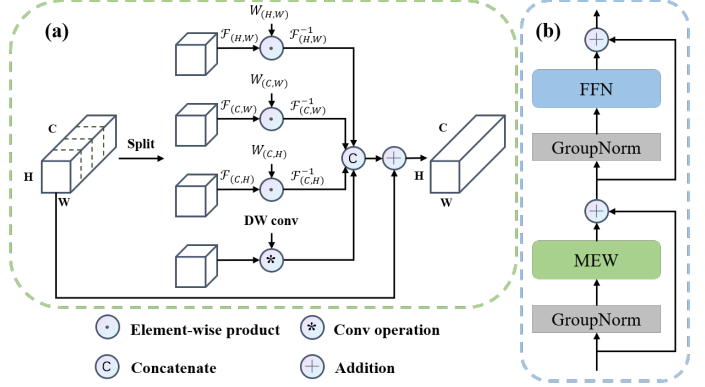


Fig. 1. (a) Multi-axis External Weights mechanism. (b) Multi-axis External Weights block. $\mathcal{F}_{(H,W)}$ and $\mathcal{F}_{(H,W)}^{-1}$ refer to conducting 2D DFT and the 2D inverse DFT along the height and width axes of the feature map. $\mathcal{F}_{(C,W)}$, $\mathcal{F}_{(C,W)}^{-1}$, $\mathcal{F}_{(C,H)}$, and $\mathcal{F}_{(C,H)}^{-1}$ could be illustrated as the same.

on the Channel-Width and Channel-Height axes. Thus, more abundant global information can be learned through the multi-axis operation. Note that the external weight is generated by our Weight Generator (see section 2.2). In addition, local information is equally vital in the MIS task. Therefore, for the fourth branch, we utilize the DW convolution to obtain the local information while reducing the amount of computation and parameters. Then, the feature maps of the four branches are concatenated along the channel dimension to restore the same size as the input. Finally, the residual connection of the input feature map is performed to obtain the output.

We replace our MEW with SA in ViT, as shown in Figure 1 (b), and obtain our MEWB, which can be expressed as formulas (5) and (6).

$$X' = \text{MEW}(\text{GroupNorm}(X)) + X \quad (5)$$

$$Y = \text{FFN}(\text{GroupNorm}(X')) + X' \quad (6)$$

Note that we apply GroupNorm [11] instead of LayerNorm in vanilla ViT. Since we directly utilize $X \in \mathbb{R}^{C \times H \times W}$ as the input instead of performing the patch embedding operation. Besides, medical image dataset samples are small in size, and the batch size is often set to be small when conducting experiments. GroupNorm with four groups is conducted in our network to reduce the impact of batch size.

2.2. External Weights Generator

MIS belongs to the layout-specific task. Namely, the variance between samples is slight, but within samples is significant in a certain medical dataset. Only using the learnable weights initialized randomly can not learn the semantic relationship between different regions well [12]. Therefore, this paper introduces an External Weights Generator to transform the initialized weights.

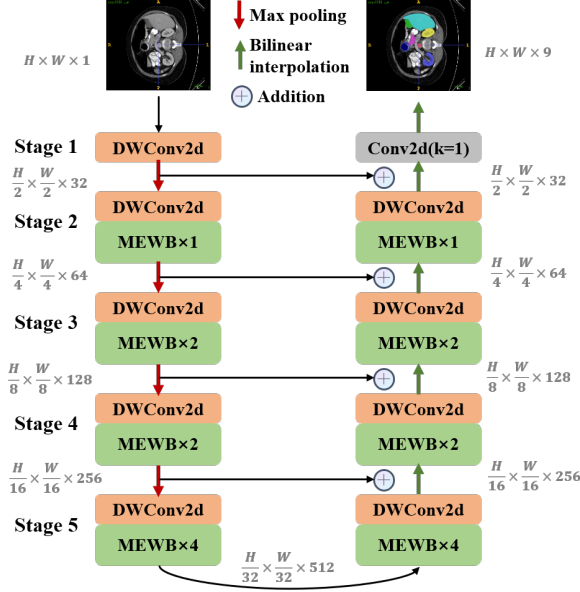


Fig. 2. The illustration of our MEW-UNet.

Specifically, the External Weights Generator is composed of several DW convolutions. For the weight corresponding to the Height-Width axes, we initialize a learnable tensor followed by a bilinear interpolation to make its dimensions the same as the input. Then, three 2D Inverted residual blocks [9] are performed on the learnable weight. We apply the same operations as before for the weights of the Channel-Width and Channel-Height axes, except for replacing 2D Inverted residual blocks with 1D Inverted residual blocks.

2.3. MEW-UNet

The structure of our MEW-UNet is shown in Figure 2. This model is based on the five-stage U-shape architecture and consists of symmetric encoders and decoders. In the first stage, only the DW convolution is used to aggregate the local features. In the following four stages, DW convolution operation is first used to change the number of channels, and then the MEWB is applied to obtain the global and local information. In addition, the number of channels is $\{32, 64, 128, 256, 512\}$, and the number of MEWB used in the last four stages is $\{1, 2, 2, 4\}$. It is worth noting that our model does not need to be pre-trained, unlike previous models, such as TransUNet and TransFuse.

3. EXPERIMENTS

3.1. Datasets and Evaluation

ISIC17 and **ISIC18**, two public datasets of skin lesion segmentation, have 2150 and 2694 dermoscopy images with segmentation mask labels, respectively. We randomly divide

Table 1. Comparative experimental results on the ISIC17 and ISIC18 dataset. (**Bold** indicates the best.)

Dataset	Model	mIoU	DSC	Acc	Spe	Sen
ISIC17	UNet [1]	76.98	86.99	95.65	97.43	86.82
	UTNetV2 [13]	77.35	87.23	95.84	98.05	84.85
	TransFuse [14]	79.21	88.40	96.17	97.98	87.14
	MEW-UNet (Ours)	81.38	89.73	96.57	97.99	89.52
ISIC18	UNet [1]	77.86	87.55	94.05	96.69	85.86
	UNet++ [2]	78.31	87.83	94.02	95.75	88.65
	Att-UNet [3]	78.43	87.91	94.13	96.23	87.60
	UTNetV2 [13]	78.97	88.25	94.32	96.48	87.60
	SANet [15]	79.52	88.59	94.39	95.97	89.46
	TransFuse [14]	80.63	89.27	94.66	95.74	91.28
	MEW-UNet (Ours)	81.90	90.05	95.18	96.99	89.57

datasets in a ratio of 7:3 for training and testing. We report five evaluation metrics, including Mean Intersection over Union (mIoU), Dice Similarity Coefficient (DSC), Accuracy (Acc), Sensitivity (Sen), and Specificity (Spe).

Synapse, a public multi-organ segmentation dataset, has 30 abdominal CT cases, including abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, stomach). Following [17], we use 18 cases for training and 12 cases for testing. Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD95) are used to evaluate our method on this dataset.

ACDC, a public cardiac MRI dataset, is composed of 100 MRI scans with three organs, including left ventricle (LV), right ventricle (RV) and myocardium (MYO). Following [6], we use the same split of 70 training cases, 10 validation cases and 20 testing cases. Dice Similarity Coefficient (DSC) is reported as the metric.

3.2. Implementation details

Empirically, for the ISIC17 and ISIC18 datasets, all images are resized to 256×256 . For the Synapse dataset, all images are resized to 224×224 . To avoid overfitting, data augmentation is performed, including random flip and random rotation. The loss function is BceDice loss. We set batch size equal to 8, and utilize CosineAnnealingLR as the scheduler. Here, we itemize the initial learning rate (lr), maximum epochs (ep), and optimizer (opt) for three datasets:

- ISIC17 and ISIC18: lr=1e-3; ep=300; opt=AdamW.
- Synapse: lr=3e-3; ep=600; opt=SGD.

Besides, the same settings as [6] are conducted for the ACDC dataset, except that we set batch size equal to 4.

Table 2. Comparative experimental results on the Synapse dataset. DSC of every single class is also reported.

Model	DSC(%) \uparrow	HD95(mm) \downarrow	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net [16]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
UNet [1]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet [3]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUnet [17]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet [5]	78.23	26.75	-	-	-	-	-	-	-	-
MT-UNet [6]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MEW-UNet (Ours)	78.92	16.44	86.68	65.32	82.87	80.02	93.63	58.36	90.19	74.26

Table 3. Comparative experimental results on the ACDC dataset. DSC of every single class is also reported.

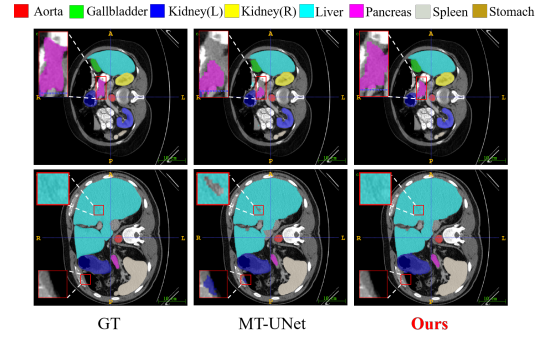
Model	DSC(%)	RV	Myo	LV
R50 UNet [17]	87.60	84.62	84.52	93.68
R50 AttnUNet [17]	86.90	83.27	84.33	93.53
R50 ViT [17]	86.19	82.51	83.01	93.05
TransUnet [17]	89.71	86.67	87.27	95.18
Swin-Unet [7]	88.07	85.77	84.42	94.03
MT-UNet [6]	90.43	86.64	89.04	95.62
MEW-UNet (Ours)	91.00	88.82	88.61	95.56

Table 4. Ablation Studies for our MEWB on the ISIC18 dataset. \circ denotes replacing GroupNorm with BatchNorm.

DW	$W_{(H,W)}$	$W_{(C,W)}$	$W_{(C,H)}$	mIoU	DSC
\circ				80.06	88.92
\checkmark				80.14	88.98
\checkmark	\checkmark			80.82	89.39
\checkmark	\checkmark	\checkmark		81.27	89.46
	\checkmark	\checkmark	\checkmark	81.29	89.68
\checkmark	\checkmark	\checkmark	\checkmark	81.90	90.05

3.3. Comparison with State-of-the-Arts

We compare our model with State-of-the-Arts in recent years, such as MT-UNet (ICASSP22) [6], UCTransNet (AAAI22) [5], TransFuse (MICCAI21) [14], SANet (MICCAI21) [15], and so on. The experimental results are shown in Table 1, Table 2 and Table 3. For the ISIC17 and ISIC18 datasets, our MEW-UNet outperforms all other state-of-the-arts on the mIoU and DSC metrics. For the Synapse dataset, our MEW-UNet is **0.33%** and **0.69%** higher than MT-UNet and UCTransNet in terms of DSC. Besides, it is worth noting that our model surpasses MT-UNet and UCTransNet **10.15mm** and **10.31mm** in the aspect of HD95. For the ACDC dataset,

**Fig. 3.** The visualization comparison on the Synapse dataset.

our MEW-UNet also achieves the best DSC metric. Some segmentation results are visualized by ITK-SNAP [18] in Figure 3.

3.4. Ablation Studies

The core idea of our MEW-UNet is multi-axis operations. Thus, we increase the number of multi-axis operations one by one to verify the effectiveness of our proposed block. The ablation performances are shown in Table 4, which reveals that our multi-axis operations can help in modeling more comprehensive frequency domain information and global knowledge. In addition, we also replace GroupNorm with BatchNorm to prove that the former is a better choice.

4. CONCLUSIONS

In this paper, we propose a Multi-axis External Weights block for acquiring frequency domain information and global knowledge. And our Weight Generator is applied to obtain external weight for medical image segmentation. Furthermore, we replace the self-attention mechanism with our block to construct MEW-UNet. Experiments show that our model achieves state-of-the-art performances. It is believed that our work will lead to new enlightenment for the subsequent model developments in the frequency domain.

5. REFERENCES

- [1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [2] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep learning in medical image analysis and multi-modal learning for clinical decision support*, pp. 3–11. Springer, 2018.
- [3] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [5] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar 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*, 2022, vol. 36, pp. 2441–2449.
- [6] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong, “Mixed transformer u-net for medical image segmentation,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2390–2394.
- [7] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, “Swin-unet: Unet-like pure transformer for medical image segmentation,” *arXiv preprint arXiv:2105.05537*, 2021.
- [8] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou, “Global filter networks for image classification,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 980–993, 2021.
- [9] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [10] Yonghao Huang, Chuan Zhou, Leiting Chen, Junjing Chen, and Shanlin Lan, “Medical frequency domain learning: Consider inter-class and intra-class frequency for medical image segmentation and classification,” in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 897–904.
- [11] Yuxin Wu and Kaiming He, “Group normalization,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [12] Jierun Chen, Tianlang He, Weipeng Zhuo, Li Ma, Sangtae Ha, and S-H Gary Chan, “Tvconv: Efficient translation variant convolution for layout-aware visual processing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12548–12558.
- [13] Yunhe Gao, Mu Zhou, Di Liu, and Dimitris Metaxas, “A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks,” *arXiv preprint arXiv:2203.00131*, 2022.
- [14] Yundong Zhang, Huiye Liu, and Qiang Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 14–24.
- [15] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui, “Shallow attention network for polyp segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 699–708.
- [16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.
- [17] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [18] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig, “User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability,” *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.