BiUNet: Towards More Effective U-Net with Bi-Level Routing Attention

Kun Dong
dongkun22@mails.ucas.ac.cn
Jian Xue
xuejian@ucas.ac.cn
Xing Lan
lanxing19@mails.ucas.ac.cn
Ke Lu*
luk@ucas.ac.cn

School of Engineering Science
University of Chinese Academy of Sciences
Beijing 100049, China

Abstract

The U-Net-like structure has indeed emerged as the paradigm for medical image segmentation due to its excellent performance. However, many variants of U-Net tend to be parameter-heavy or computationally complex, limiting their utilization of fast image segmentation in practical applications. In this paper, we propose BiUNet, a powerful and efficient model which well incorporates a lightweight attention module, Bi-Level Routing Attention (BRA). Besides, to compensate for the information loss caused by downsampling and further enhance the network's performance, we introduce two innovative techniques termed pixel merging and pixel expanding, which are seamlessly integrated into BiUNet. Extensive experiments demonstrate that our model can achieve better performance than the latest networks with fewer parameters and lower FLOPs.

1 Introduction

Medical image segmentation is one of the most critical tasks in medical image analysis. U-Net [15], as a relatively lightweight model, is widely used in the field of medical image analysis due to its excellent generalization and remarkable performance. Derived from it, dozens of networks like UNet++ [27], Att U-Net [13], SwinUnet [1], and TransUNet [2] have been proposed and achieved impressive performance on many difficult datasets, proving the effectiveness of U-shaped architecture with skip connection. However, a common limitation of these models is their tendency to be parameter-heavy or computationally complex, which significantly restricts their application. Therefore, the research on further strengthening U-Net by efficiency is valuable and necessary.

Different from the aforementioned networks, our BiUNet only introduces limited extra parameters but can achieve better performance, as shown in Figure 1. We achieve this by resolving two key limitations of the vanilla U-Net. First, U-Net lacks the capability to model long-range dependencies present in an image. Previous works have focused on addressing

*Corresponding author.

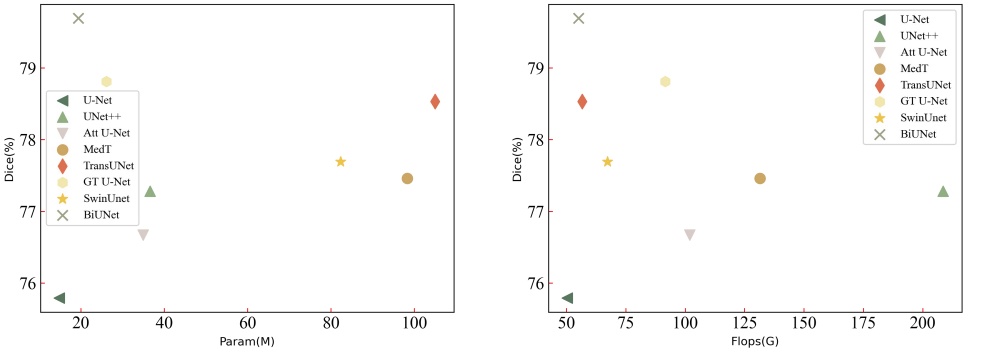

Figure 1: Performance of U-Net-like models on MoNuSeg dataset. Compared to other variants of U-Net, BiUNet (close to the top left corner) achieves better performance with fewer parameters and lower flops.

this issue through the incorporation of attention mechanisms [1, 2, 16, 21]. However, to alleviate the computational burden, these models typically confine the attention mechanism to a fixed window [21, 25] or hand-pick the locations that require attention [5, 16], thus failing to achieve global attention. To maintain the network's computational efficiency without compromising the ability to adequately model global information, we incorporate a powerful and efficient attention mechanism, Bi-Level Routing Attention (BRA)[28], which is lightweight enough and has the ability to achieve global attention. More details can be seen in Section 3.

Secondly, the downsampling operation inevitably causes serious information loss in U-Net network. Max pooling is one of the most widely used downsampling operations, as it preserves crucial information from the previous feature map and prevents overfitting to a certain extent. However, it also has a risk of losing information, as only part of the information in the previous feature map is retained. To address this issue, we design a module termed pixel merging to alleviate the loss of information. Moreover, in the upsampling phase, we replace deconv with pixel expanding for better performance.

Our model, with 19.4M parameters and 55.2G flops, demonstrates remarkable effectiveness in extensive experiments. It achieves 79.70% Dice and 70.20% mIoU on QaTa-COV19 benchmark, as well as 79.69% Dice and 66.44% mIoU on MoNuSeg benchmark.

In summary, the key contributions of our approach are:

- We propose BiUNet, a powerful and efficient network. It incorporates the Bi-Level Routing Attention to model global information. To mitigate the information loss caused by downsampling, we design a novel technique, pixel merging. Also, we introduce pixel expanding to further enhance the network's performance.
- BiUNet exhibits remarkable improvement over the vanilla U-Net baseline and achieves excellent performance on QaTa-COV19 [4] and MoNuSeg [9] datasets.

2 Related work

Variants of U-Net in medical image analysis. Semantic segmentation can be viewed as a pixel-level image classification work, and various approaches have been proposed to solve

this task. Among them, U-Net demonstrated superior performance and thus is considered a pioneer in this field. Following the U-shaped architecture with skip connection, dozens of variants have been proposed to address the issues of the original version to improve the performance of the network. For example, UNet++ [27] and UCTransNet [19] improved the skip connection of U-Net. Wang *et al*. [22] proposed a non-local operator, which can be plugged into convolution layers. In contrast to these variants that often come with considerable computational and memory overhead, our proposed BiUNet can achieve impressive performance while utilizing fewer parameters and lower flops.

Transformers with attention mechanism. Transformer [18] was first proposed in the field of natural language processing and achieved state-of-the-art in various tasks. In order to make the transformer applicable to computer vision, images are often first divided into different regions [6]. In comparison with CNNs, transformer can model global content by using attention mechanism. However, the vanilla attention mechanism introduces the quadratic complexity of sequence length, which thus incurs high computation costs and heavy memory footprint. Therefore, it's of vital importance to seek efficient attention modules for various tasks. There have been a variety of approaches proposed to reduce the time and space complexity by using low-rank approximations [23] or sparse connection patterns [3], such as Linformer [20] and Switch Transformer [7]. Recently, in the field of computer vision, there is also a lot of work on efficient transformer. For instance, Deformable Attention Transformer [24] achieves adaptive sparsity by deforming a regular grid. Our approach can be seen as a combination of U-shaped structure and efficient attention mechanism.

3 Preliminary

Bi-Level Routing Attention (BRA) is a dynamic sparse attention mechanism introduced in [28]. It incorporates a bi-level routing strategy that enables a flexible allocation of computations while maintaining content awareness. The key idea is to initially select the most relevant pairs within a coarse window level, after which only a small portion of the window remains. Subsequently, pixel-to-pixel level attention is applied within the remaining windows. The detailed explanation is as follows.

Window-to-window routing with index matrix. Given a 2D input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, we start by deriving the query, key, value tensor, $\mathbf{Q},\mathbf{K},\mathbf{V} \in \mathbb{R}^{H\times W\times C}$, with linear projections:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^q, \mathbf{K} = \mathbf{X}\mathbf{W}^k, \mathbf{V} = \mathbf{X}\mathbf{W}^v, \tag{1}$$

where $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v \in \mathbb{R}^{C\times C}$ are projection weights. Then $\mathbf{Q},\mathbf{K},\mathbf{V}$ are divided into $S\times S$ non-overlapped windows so that each window contains $\frac{HW}{S^2}$ feature vectors. Specifically, we reshape $\mathbf{Q},\mathbf{K},\mathbf{V}$ as $\hat{\mathbf{Q}},\hat{\mathbf{K}},\hat{\mathbf{V}} \in \mathbb{R}^{S^2\times\frac{HW}{S^2}\times C}$.

By applying per-window average on $\hat{\mathbf{Q}},\hat{\mathbf{K}}$, we get $\mathbf{Q}^w,\mathbf{K}^w \in \mathbb{R}^{S^2\times C}$. Then we can derive window-to-window score matrix:

$$\mathbf{A}^w = \mathbf{Q}^w(\mathbf{K}^w)^T, \tag{2}$$

where $\mathbf{A}^w \in \mathbb{R}^{S^2\times S^2}$. Each entry in $\mathbf{A}^w$ measures how relevant two windows are. For each window, only the indexes of the top-k windows are stored. In particular, with row-wise top-k operator, we can get an index matrix, $\mathbf{I}^w \in \mathbb{N}^{S^2\times k}$. Each row of this matrix contains the indexes of the k windows that are most relevant to the window corresponding to that row.

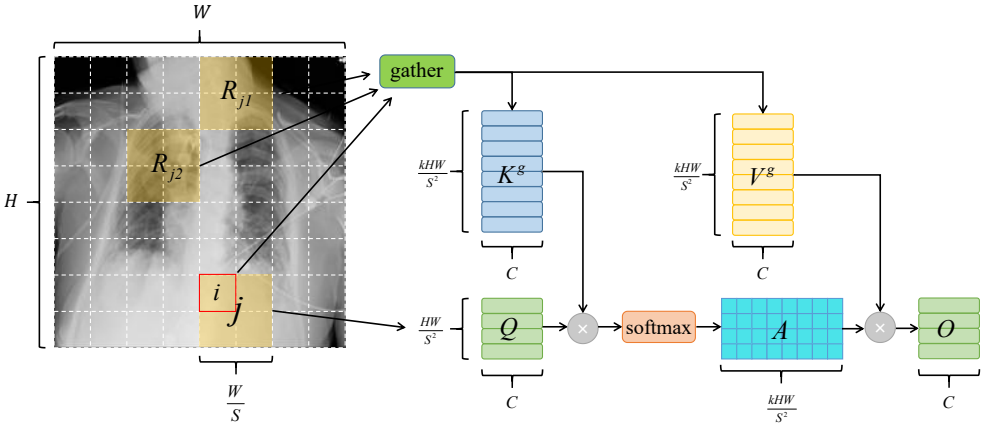

Figure 2: BRA utilizes the sparsity to compute only with pixels in the most relevant k windows ($k = 3$ in this example). The gather operation is used to make the calculation process only include GPU-friendly matrix multiplication.

Pixel-to-pixel level attention. Although we know the indexes of the k windows most relevant to a window, it's non-trivial to implement pixel-to-pixel level attention efficiently, because these windows are scattered over the whole feature map, which is unfriendly to the GPU. Therefore, we need to gather key and value tensor first, i.e.

$$\mathbf{K}^g = \text{gather}(K, \mathbf{I}^w), \mathbf{V}^g = \text{gather}(V, \mathbf{I}^w), \tag{3}$$

where $\mathbf{K}^g, \mathbf{V}^g \in \mathbb{R}^{S^2 \times \frac{kHW}{S^2} \times C}$. Finally, we apply pixel-to-pixel level attention. As illustrated in Figure 2, if pixel i is in window j and $R_{j1}, R_{j2}, ..., R_{jk}$ are the k most relevant windows to window j, pixel i will pay attention to all pixels in these k windows.

4 Method

This section introduces BiUNet, our efficient and powerful segmentation framework. As depicted in Figure 4, compared to the vanilla U-Net, BiUNet has made two main improvements: (1) Introduce Bi-Level Routing Attention in the decoding phase to help the network learn long-range dependencies; and (2) Utilize pixel merging and pixel expanding to replace the max pooling and deconv operation respectively, which can reduce information loss derived from sampling to enhance network performance. In the rest of this section, the proposed improvements are introduced in detail.

4.1 Pixel Merging and Expanding

In U-shaped networks, downsampling and upsampling are essential operations. Specifically, in the encoding phase, features are learned through layer-by-layer downsampling. In the decoding phase, features at different scales are fused by upsampling and skip connection [15]. However, max pooling commonly used for downsampling results in some loss of information. Therefore, as shown in Figure 3, we propose to use pixel merging to replace the role

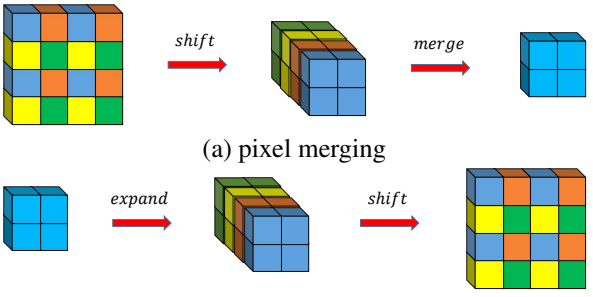

Figure 3: Improved upsampling and downsampling operation.

of max pooling, and correspondingly, in the upsampling phase, we use pixel expanding instead of deconvolution. To simplify the notations, we discuss the case that the stride for downsampling and upsampling are both 2×2. A detailed explanation is given as follows.

Pixel Merging. As shown in Figure 3(a), consider a local region of size 2×2, denoted as $\mathbf{R} \in \mathbb{R}^{2\times 2\times C_{\text{in}}}$, which owns four feature vectors. We first shift these four feature vectors to form one with more channels. Specifically, we reshape $\mathbf{R}$ as $\hat{\mathbf{R}} \in \mathbb{R}^{4C_{\text{in}}}$. Then conv 1×1, layer norm, and ReLU are used in turn to get the output $\mathbf{O} \in \mathbb{R}^{C_{\text{out}}}$, i.e., a feature vector. Downsampling by this operation can better preserve the information in $\mathbf{R}$ and thus enhance the performance of the network.

Pixel Expanding. As shown in Figure 3(b), given an input feature map $\mathbf{X} \in \mathbb{R}^{\frac{H}{2^r}\times\frac{W}{2^r}\times C_{\text{in}}}$, we first use conv 1×1, layer norm and ReLU in turn to get the output $\hat{\mathbf{O}} \in \mathbb{R}^{\frac{H}{2^r}\times\frac{W}{2^r}\times 4C_{\text{out}}}$. Then we reshape $\hat{\mathbf{O}}$ as $\mathbf{O} \in \mathbb{R}^{\frac{H}{2^{r-1}}\times\frac{W}{2^{r-1}}\times C_{\text{out}}}$. The influence of using pixel merging and expanding will be discussed in Section 5.4.

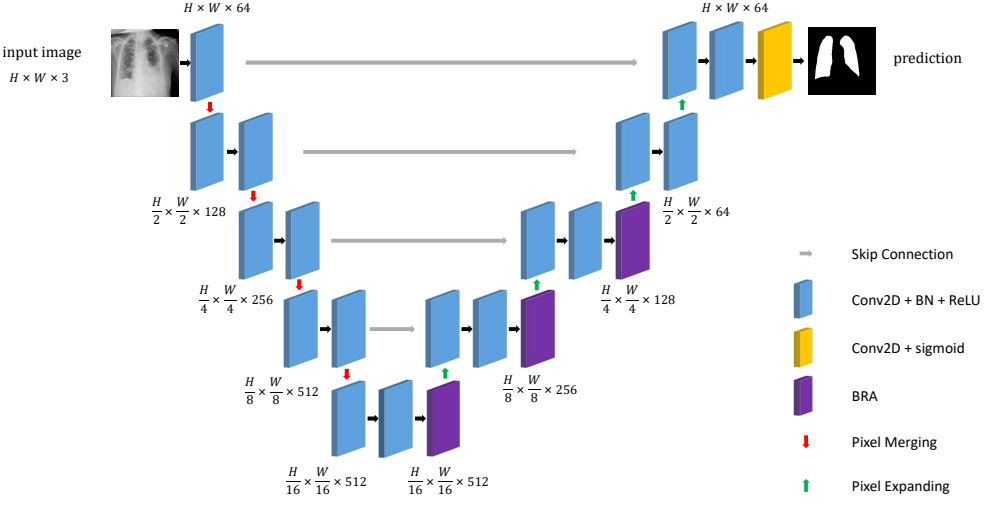

Figure 4: Overview of the proposed BiUNet architecture.

4.2 Architecture Design of BiUNet

The vanilla U-Net architecture lacks the capability to effectively capture long-range dependencies within an image. Previous approaches have attempted to address this issue by incorporating attention mechanisms [1, 2, 16, 21]. However, to alleviate the computational burden, these methods either confine the attention mechanism to a fixed window or manually select specific locations for attention, resulting in a lack of global attention. In contrast to these approaches, Bi-Level Routing Attention (BRA) strikes a balance between computational efficiency and the ability to achieve global attention, allowing the network to effectively learn long-range dependencies. Hence, to capture global content information effectively, we seamlessly integrate it into our network.

With BRA, pixel merging, and pixel expanding, we propose a new network, BiUNet. As shown in Figure 4, we follow the U-shaped structure with five stages, while the height and width of the feature map for stage i are $\frac{H}{2^{i-1}}, \frac{W}{2^{i-1}}$, respectively. In the encoding stage, we use pixel merging to mitigate the information loss caused by max pooling in downsampling and utilize convolution to progressively expand the receptive field. In the decoding stage, pixel expanding is correspondingly used for upsampling and the influence can be seen in Section 5.4. In order to make the network have the ability to understand long-distance dependencies present in an image, we insert BRA after convolution to learn correlations between different windows that may be located far away from each other. This is important because the convolution kernel alone may not effectively capture the relationships between such distant regions. It's worth noting that we only add BRA in stage 3, 4, 5 to avoid too much extra computational overhead or memory footprint. For simplification, we only discuss the case of single-head self-attention in Section 3, while multi-head self-attention [18] is used in practice. The configurations of BRA in different stages are listed in Table 1.

stage	S	k	pixels in window	heads
3	7	1	64	2
4	7	4	16	4
5	7	16	4	8

Table 1: Configurations of BRA in different stages.

The ablation study on the location where the BRA module is inserted can be found in Section 5.4. Note that skip connection and concatenate operation are performed between each layer.

The loss function we use consists of two parts, one is the cross-entropy loss (L_{CE}) formulated as:

$$L_{\text{CE}} = -\sum_{i=1}^{N}\sum_{j=1}^{K}\frac{1}{N}\cdot y_{ij}\log(p_{ij}), \tag{4}$$

where N and K represent the number of pixels and classes respectively. The other is dice loss (L_{Dice}), which is described as:

$$L_{\text{Dice}} = 1 - \sum_{i=1}^{N}\sum_{j=1}^{K}\frac{1}{NK}\cdot\frac{2|p_{ij}\cap y_{ij}|}{|p_{ij}|+|y_{ij}|}, \tag{5}$$

where y_{ij} means whether pixel i belongs to class j, which takes 1 if pixel i belongs to class j, otherwise 0. p_{ij} is the output of the model, which is the probability that pixel i belongs to

class j. Therefore, the total loss function [11] is defined as follows:

$$L = \alpha L_{\text{CE}} + \beta L_{\text{Dice}}, \tag{6}$$

where hyperparameter α and β are loss weights corresponding to L_{CE} and L_{Dice}, respectively.

5 Experiments

5.1 Datasets

In the experiments, two datasets are used to evaluate the effectiveness of our proposed BiUNet. One is the QaTa-COV19 dataset [4]. Researchers from Qatar University and Tampere University compiled this dataset which consists of 9258 COVID-19 chest X-ray images, including the ground-truth segmentation masks for the COVID-19 pneumonia segmentation task for the first time. The other is the MoNuSeg dataset [9], which is created using H&E stained tissue images captured at 40× magnification. The training data consists of 30 images with around 22,000 single kernel annotations, while the test set contains over 7000 single kernel annotations in 14 images. Images in our experiments are all resized to 224×224. The specific division of the two datasets is listed in Table 2.

	QaTa-COV19	MoNuSeg
Train set	5716	24
Val set	1429	6
Test set	2113	14

Table 2: The division of datasets.

5.2 Implementation Details

All experiments were conducted using the PyTorch platform [14] and trained/tested on NVIDIA A100 GPU. Different batch sizes are applied for the two datasets as their data sizes are different. Specifically, the default batch sizes are 2 and 24 for the MoNuSeg and QaTa-COV19 datasets, respectively. We use Cosine Decay and AdamW [12] optimizer with weight decay set to 0.0001. The initial learning rates are 0.001 for the MoNuSeg dataset and 0.0003 for the QaTa-COV19 dataset. An early stop mechanism until the performance of the network does not improve is applied during the training process. In this work, loss weights α and β are both set to 0.5.

5.3 Comparison with SOTA

To verify the segmentation performance of our proposed BiUNet, we compare it with other recent and widely used medical image segmentation networks. Specifically, We compare BiUNet with two types of frameworks for a comprehensive evaluation, including three convolutional baselines like U-Net [15], Att U-Net [13], UNet++ [27], and four recent transformer baselines like MedT [17], TransUNet [2], GT U-Net [10], SwinUnet [1]. Note that we use Dice score and mIoU as criteria to evaluate the performance of different models.

Network	Param (M)	Flops (G)	QaTa-COV19		MoNuSeg	
			Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
U-Net [15]	14.8	50.4	**78.96**	**69.33**	**75.79**	**62.96**
UNet++ [27]	36.6	208.6	79.12	69.72	77.28	64.44
Att U-Net [13]	34.9	101.9	79.31	70.04	76.67	63.47
MedT [17]	98.3	131.5	77.47	67.51	77.46	63.37
TransUNet [2]	105.0	56.7	78.63	69.13	78.53	65.05
GT U-Net [10]	26.1	91.6	79.19	69.72	78.81	65.23
SwinUnet [1]	82.3	67.3	78.07	68.34	77.69	63.77
BiUNet	19.4	55.2	**79.70**	**70.20**	**79.69**	**66.44**

Table 3: Performance comparison between our proposed BiUNet and other convolutional and transformer baselines on QaTa-COV19 and MoNuSeg datasets.

The quantitative experimental results are tabulated in Table 3, where the best results are boldfaced. Through Table 3, we can observe that our proposed BiUNet achieves better segmentation performance than all the convolutional and transformer baselines. It is worth noting that with our improvements, BiUNet has a significant segmentation performance improvement over the vanilla U-Net. In particular, it improves the Dice score by 0.74% and the mIoU by 0.87% on the QaTa-COV19 dataset. On the MoNuSeg dataset, BiUNet improves the Dice score by 3.90% and the mIoU score by 3.48%.

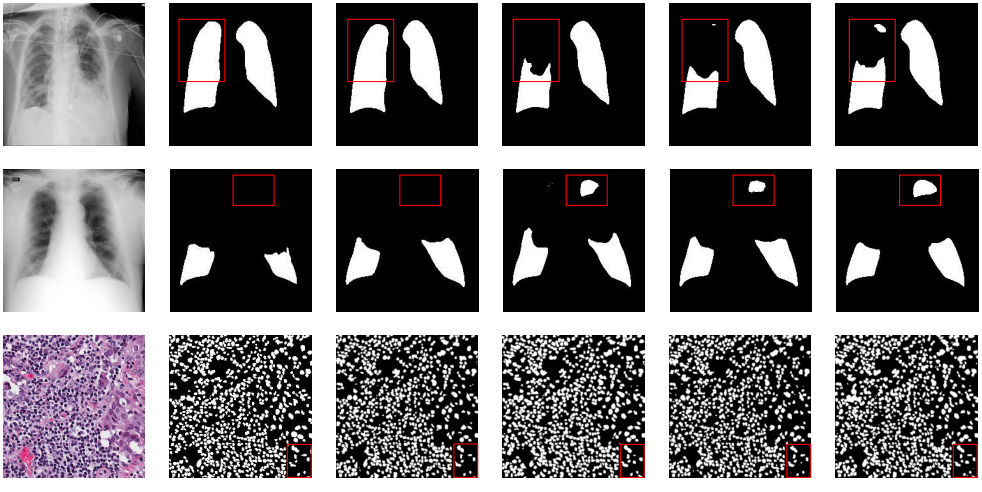

Figure 5: Qualitative results on QaTa-COV19 and MoNuSeg datasets.

The qualitative results of BiUNet along with other baselines on MoNuSeg and QaTa-COV19 datasets are demonstrated in Figure 5, where three baseline networks are selected for comparison. It can be observed that our model BiUNet has the ability to produce excellent semantic segmentation predictions. In Figure 5, we use red boxes to highlight regions where BiUNet generates better segmentation results than other models.

5.4 Ablation Studies

The Effectiveness of Proposed Improvements. A series of ablation experiments are conducted to verify the effectiveness of different components. As can be seen in Table 4, Pixel Merging (PM), Pixel Expanding (PE), and BRA are effective. It is worth noting that model I is vanilla U-Net and model VI is our proposed BiUNet.

ID	PM	PE	BRA	Dice (%)	mIoU (%)
I				75.95	62.96
II	✓			78.40	65.06
III		✓		77.38	64.42
IV			✓	76.73	63.62
V	✓	✓		79.33	65.88
VI	✓	✓	✓	**79.69**	**66.44**

Table 4: Ablation study on effectiveness of components on MoNuSeg dataset.

Ablation study on the location where the BRA module is inserted. We conduct ablation experiments on QaTa-COV19 dataset to investigate where it is most suitable to insert BRA. It can be seen in Table 5 that only inserting BRA in the decoding stage allows the model to get the best performance. Note that we only add it in the last three stages of the network either in the encoding or decoding phase in order to reduce the additional computational cost and memory footprint. E represents the encoding phase while D means the decoding phase.

E	D	Dice (%)	mIoU (%)
✓		79.19	69.71
	✓	**79.70**	**70.20**
✓	✓	79.41	69.88

Table 5: Ablation study on the location where the BRA module is inserted.

Ablation study on different attention modules. By incorporating various attention modules into the decoding phase, we investigate their impact on the performance of the network. As shown in Table 6, BRA outperforms other attention modules. Note that None represents the vanilla U-Net without any attention module.

Attention type	QaTa-COV19		MoNuSeg	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
None	78.96	69.33	75.79	62.96
Swin [25]	79.02	69.48	77.18	63.75
NAT [8]	79.11	69.62	77.67	64.15
Pooling [26]	79.10	69.55	78.50	64.86
BRA	**79.70**	**70.20**	**79.69**	**66.44**

Table 6: Ablation study on different attention modules on QaTa-COV19 and MoNuSeg datasets.

6 Conclusion

In this work, we have introduced a new deep network architecture BiUNet for medical image segmentation. In order to leverage the advantage of the attention mechanism in learning long-range semantic information and remain efficient, we incorporate a lightweight enough attention mechanism, BRA. To reduce the loss of information due to downsampling we utilize pixel merging, which corresponds to the use of pixel expanding in the upsampling phase. Extensive experiments have shown the proposed BiUNet can achieve excellent performance. In the future, we will combine the proposed modules with other models and conduct experiments on more data modalities to verify the generalization.

7 Acknowledgments

We would like to thank the anonymous reviewers for their valuable suggestions. This work was supported by the National Natural Science Foundation of China (62027827, 61929104, 61972375), the Beijing Natural Science Foundation (7222167) and the R&D Program of Beijing Municipal Education Commission (2019022).

References

[1] 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. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer, 2023.

[2] 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.

[3] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[4] Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2306–2310. IEEE, 2022.

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022.

[6] 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.

[7] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

[8] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. *arXiv preprint arXiv:2204.07143*, 2022.

[9] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019.

[10] Yunxiang Li, Shuai Wang, Jun Wang, Guodong Zeng, Wenjun Liu, Qianni Zhang, Qun Jin, and Yaqi Wang. Gt u-net: A u-net like group transformer network for tooth root segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 386–395. Springer, 2021.

[11] Zihan Li, Yunxiang Li, Qingde Li, You Zhang, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *arXiv preprint arXiv:2206.14718*, 2022.

[12] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2017.

[13] 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.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[16] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 459–479. Springer, 2022.

[17] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] 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*, volume 36, pages 2441–2449, 2022.

[20] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[21] Wenxiao Wang, Wei Chen, Qibo Qiu, Long Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wei Liu. Crossformer++: A versatile vision transformer hinging on cross-scale attention. *arXiv preprint arXiv:2303.06908*, 2023.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[23] Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148. IEEE, 2020.

[24] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.

[25] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.

[26] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[27] 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 Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[28] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson Lau. Biformer: Vision transformer with bi-level routing attention. *arXiv preprint arXiv:2303.08810*, 2023.