Dual-MambaNet: A Lightweight Dual-Branch Brain Image Segmentation Network Based on Local Attention and Mamba[⋆]

Feifei Zhang[1,2], Fei Shi[1,2,✉], Dayong Ren[3,✉], Zhenhong Jia[1,2], and Jianyi Wang[1,2]

[1] Xinjiang University, School of Computer Science and Technology,
Xinjiang Urumqi 830046, China

[2] Xinjiang University, Key Laboratory of Signal Detection and Processing,
Xinjiang Urumqi 830046, China
sigofei@xju.edu.cn

[3] National Key Laboratory for Novel Software Technology, Nanjing University,
Nanjing 210023, China
rdyedu@gmail.com

Abstract. Brain tissue segmentation is critical for diagnosing and treating brain diseases. While Mamba-based models excel in the medical field, they face performance bottlenecks with high-resolution MRI images, often losing local feature information in complex texture structures. To address these challenges and enable deployment in resource-limited settings, we propose Dual-MambaNet, a lightweight segmentation model based on Mamba. In Dual-MambaNet, we introduce the Outlook attention module to capture local complex textures and structures in brain MRI images. Subsequently, we combined it with the Mamba block to construct a feature extractor (FE) encoder layer to couple local and global features. Additionally, we integrate dual decoder branches and a multi-level pixel contrastive loss function(MPCL) to better integrate local and global features. This method optimizes global feature representation by refining local complex textures and structural details, effectively capturing multi-level features in MRI images. Experimental results on public brain MRI datasets OASIS1 and MRBrainS13 demonstrate that Dual-MambaNet achieves high segmentation accuracy with minimal parameters and computational complexity, making it suitable for deployment in resource-limited medical environments.

Keywords: MRI · Brain tissue segmentation · Mamba · MPCL.

1 Introduction

The rapid diagnosis of brain and nervous system disorders depends on healthcare professionals' expertise and professional skills. The analysis of brain mag-

[⋆] This work was supported by National Natural Science Foundation of China (No.62261053) and Tianshan Talent Training Project - Xinjiang Science and Technology Innovation Team Program (2023TSYCTD0012).

[✉] Corresponding author

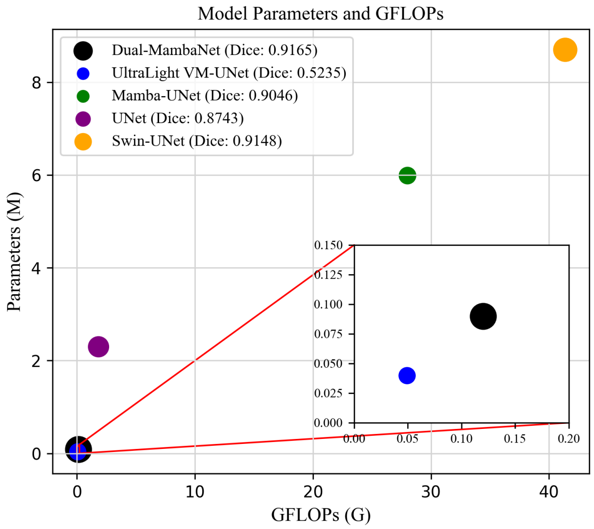


Fig. 1: Comparison of model parameters (Million) and GFLOPs (G) (The size of the circles represents the average Dice score on the OASIS1 dataset (↑)).

netic resonance imaging (MRI) by healthcare professionals requires a significant amount of time and effort. In recent years, with the rapid advancement of computer technology, the use of computer-assisted techniques has improved the speed of segmentation and diagnosis of magnetic resonance imaging (MRI), enhancing the efficiency of medical diagnosis [10].

Convolutional neural networks(CNN), represented by UNet [14], are widely applied in medical image segmentation. CNN excel at capturing local features but may struggle to utilize global contextual information, which can lead to lower segmentation accuracy [2]. Inspired by self-attention mechanisms in natural language processing, Vision Transformer (ViT) was the first to apply multi-head attention mechanisms to visual tasks [3]. Due to its excellent capability in extracting global context, ViT is widely applied in medical image segmentation. However, its quadratic complexity can lead to high computational costs, especially in high-resolution medical image segmentation. In resource-constrained medical environments, these high computational costs pose challenges for model deployment. Therefore, there is an urgent need in medical image segmentation for lightweight algorithms that can achieve high accuracy.

Recently, advancements in state space models (SSM) [12] have provided new insights into lightweight medical image segmentation algorithms. The linear complexity of state space models and their excellent capability to model long-range relationships have led to their widespread application in medical image segmentation, with Mamba as a representative example [4]. To facilitate model deployment in medical environments and improve segmentation accuracy, a series of lightweight Mamba models have been proposed to facilitate deployment in medical environments. LightM-UNet [8] significantly reduces model parameters while maintaining high segmentation accuracy to ensure feasibility for deployment in medical environments. UltraLight VM-UNet [18] introduces the Parallel Vision Mamba (PVM) module, resulting in a more lightweight model that ensures accuracy in skin lesion segmentation tasks.

Despite these studies alleviating the issues of complexity and computational cost to some extent, existing models still face performance bottlenecks when processing high-resolution MRI images, often losing local feature information in complex texture structures. To address these challenges and enable deployment in resource-limited environments, we propose Dual-MambaNet, a lightweight segmentation model based on Mamba. In Dual-MambaNet, we introduce the Outlook attention module to capture local complex textures and structures in brain MRI images. Subsequently, we combine it with the Mamba block to construct a feature extractor (FE) encoder layer, coupling local and global features. Additionally, we propose dual decoder branches and a multi-level pixel contrastive loss function (MPCL) to integrate local and global features better. This approach optimizes global feature representation by refining local complex textures and structural details, effectively capturing multi-level features in MRI images. Fig. 1 compares the parameters and GFLOPs of Dual-MambaNet and other models(UNet, Swin-UNet, Mamba-UNet and UltraLight VM-UNet). As shown, Dual-MambaNet maintains high accuracy while having lower parameters and GFLOPs, facilitating its deployment in resource-limited medical environments.

In this paper, our contributions are as follows:

1. This paper designs a feature extractor (FE) as the encoder part, which extracts structural features through spatial transformation operations achieved by adaptive long-range and short-range computations. Specifically, Mamba is used for extracting global contextual information, while the local attention mechanism (Outlook attention) captures local features.
2. This paper employs a dual-branch decoder to strengthen the coupling of information at different levels and enhance the model's ability to couple global and local features.
3. A multi-level pixel contrastive loss function(MPCL) is proposed to optimize the coupling of the model's low-level and high-level features.
4. This paper proposes a lightweight model for brain MRI image segmentation. The model maintains high segmentation accuracy with a minimal increase in the number of parameters and GFLOPs.

2 Related Work

With the development of artificial intelligence, deep learning has been widely applied to medical image segmentation. Convolutional neural networks (CNN) have been extensively used for image segmentation tasks [9]. UNet [14] has been widely applied in medical image segmentation due to its symmetric encoder-decoder architecture and skip connections [7]. The encoder and decoder of UNet can extract features at different levels, and the skip connections facilitate efficient transformation between these levels. However, this simple fusion method can only partially exploit these features, inevitably creating a semantic gap between features at different levels. To bridge this semantic gap, UNet++ [22] enhances the fusion of high-level and low-level information by adding convolutional layers within the skip connections. Building on this, UNet3+ [6] achieves more accurate segmentation by integrating multi-scale high-level and low-level features. However, methods based on CNNs can only extract local information and need moreapture global contextual information.

To enhance the extraction of global contextual information, TransUNet [2] combines Transformer with UNet, achieving higher accuracy in medical image segmentation. Building on this, UTNet [5] employs multi-scale Transformers to fuse high-level and low-level features better. Given the suitability of ViT [3] for visual tasks and its robust feature extraction capabilities, many studies have integrated ViT and its variants with UNet, yielding improved results.

The linear complexity and long-range relationship modelling capability of state space models have recently led to the widespread application of Mamba models in medical image segmentation. Studies have shown that Mamba is effective in image segmentation [20]. VMamba [23] introduced a hierarchical visual backbone network and Cross-Scan Module (CSM) based on Mamba, making Mamba more suitable for 2D image tasks. Mamba-UNet [17], based on the Swin-UNet architecture, applies pure visual Mamba modules (VSS) to medical image segmentation, outperforming CNN- and Transformer-based models. LightM-UNet [8] was proposed to explore more lightweight models, significantly reducing the number of model parameters. UltraLight VM-UNet [18] verified the significant impact of channel count on model parameters and introduced the Parallel Vision Mamba module (PVM), achieving a more lightweight model while ensuring accuracy in skin lesion segmentation.

Although CNN-based methods can accomplish complex medical image segmentation tasks, they often fail to fully utilize global contextual information, resulting in poor feature extraction capability and, consequently, lower segmentation performance. On the other hand, transformer-based algorithms can effectively extract global contextual information, but they tend to be complex with high computational complexity. This paper proposes a lightweight medical image segmentation model based on Mamba (Dual-MambaNet) to address information loss issues in brain MRI image segmentation and the difficulty of deploying complex models in resource-constrained medical environments.

3 Method

3.1 Architecture Overview

Fig. 2 shows the overall architecture of Dual-MambaNet, which consists of a 6-layer encoder, a 3-layer low-level feature decoder, and a 6-layer high-level feature decoder, forming an asymmetric U-shaped network. In the encoder part, except

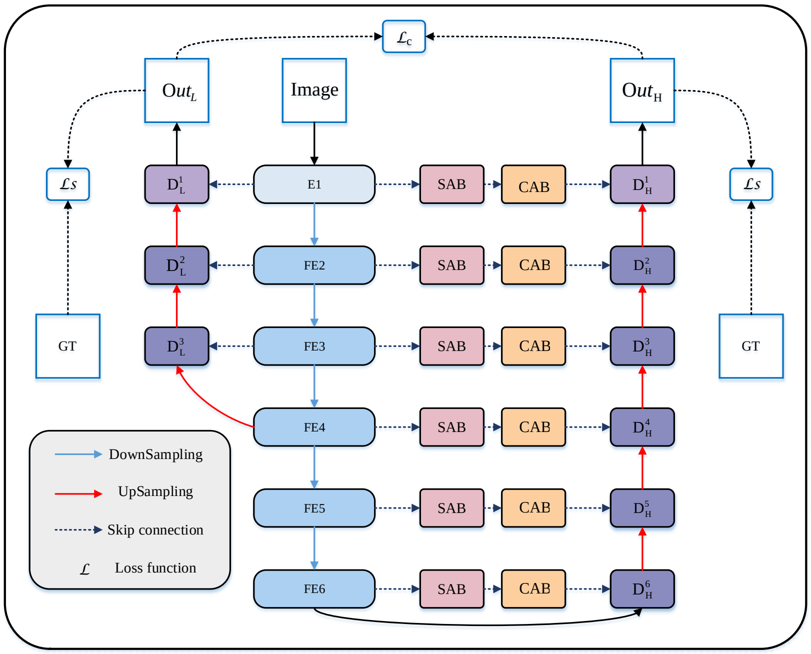


Fig. 2: The overall architecture of Dual-MambaNet.

for the first encoder layer, which uses convolutional layers and local attention mechanisms, all other layers use the feature extractor (FE), as shown in Fig. 3(b). In the decoder part, except for the last decoder layer, all other layers use the parallel Mamba layer (PVM) [18]. The number of channels in each layer of the encoder and decoder structures are [8, 16, 24, 32, 48, 64].

Skip connections use the channel attention bridge (CAB) and spatial attention bridge (SAB) as proposed in [15]. The SAB module includes max pooling, average pooling, and dilated convolutions with shared weights. The CAB module includes global average pooling, concatenation operations, fully connected layers, and a Sigmoid activation function. In the skip connections of the low-level decoder, attention bridges are not used to avoid over-decoding due to the large gap between shallow features (which contain better detail information) and deep features (which contain more semantic information).

The model has two final outputs: the left decoder branch outputs low-level features, and the right decoder outputs high-level features. The low-level output information is also used to finely optimize the high-level output through a pixel-level contrastive loss function, improving the model's segmentation accuracy.

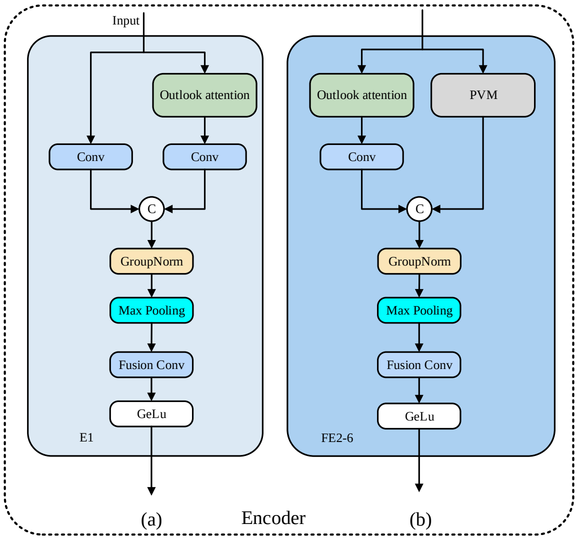


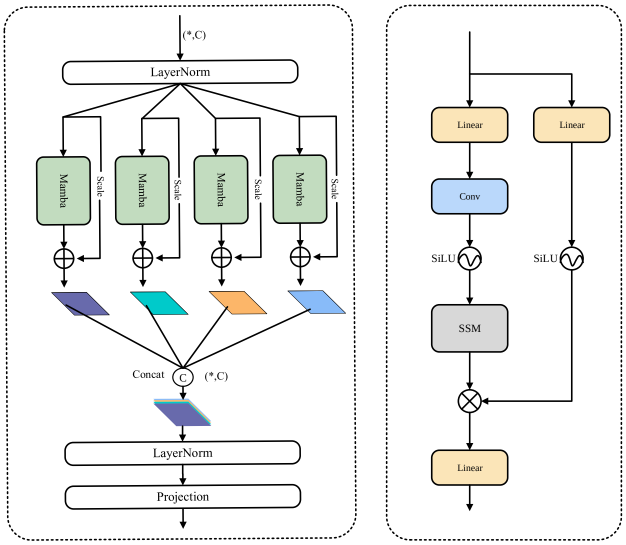
Fig. 3: Encoder Structure Diagram (a) E1: First Layer Encoder Structure. (b) FE2-FE6: Feature Encoder (FE) Constructed.

3.2 Parallel Vision Mamba Layer (PVM)

The Mamba module used in this paper is the Parallel Vision Mamba layer (PVM) proposed in [18], which is based on the Mamba module [4]. Its structure is shown in Fig. 4(b). In [18], the impact of the number of channels on the Mamba parameter count was extensively discussed, demonstrating that the number of channels has an exponential effect on the Mamba parameter count. Based on this conclusion, the PVM Layer was proposed. The structure of PVM is shown in Fig. 4(a). PVM mainly combines Mamba with residual connections and adjustment factors. The feature token X (with C channels) first passes through a LayerNorm layer and is then divided into four sub-features (each with C/4 channels) along the channel dimension. Each sub-feature is then fed into Mamba, and the outputs are subjected to residual and adjustment operations to optimize the ability to capture long-range spatial information. Finally, the four features are concatenated along the channel dimension to form X_{out}, which has the exact dimensions as the original input X. X_{out}, then undergoes LayerNorm and linear projection operations to transform it to the exact dimensions as the original image. This allows Mamba to enhance the capture of long-range spatial relationships without introducing additional parameters and computational complexity.

3.3 Outlook attention

The Outlook attention in this paper is based on [21], and its specific structure is shown in Fig. 5. Specifically, for each spatial position (i, j), Outlook Atten-



（a）PVM

（b）Mamba

Fig. 4: (a) Parallel Vision Mamba (PVM) structure diagram. (b) Mamba Block.

tion calculates its similarity with all neighbours within a local window of size K×K centred at (i, j). Unlike self-attention, which requires a Query-Key matrix multiplication to compute attention, Outlook Attention simplifies this process through a reshaping operation.

Formally, given the input X, each C-dimensional token is first projected using two linear weight layers $W_A \in \mathbb{R}^{C\times K^4}$and $W_V \in \mathbb{R}^{C\times C}$ into the outlook weights $A \in \mathbb{R}^{H\times W\times K^4}$ and value representations $V \in \mathbb{R}^{H\times W\times C}$, respectively. Let $V_{\Delta_{i,j}} \in \mathbb{R}^{C\times K^2}$ denote the values within the local window centred at (i, j), This process is represented by Eq. 1.

$$V_{\Delta_{i,j}} = \left\{ V_{i+p-\left\lfloor \frac{K}{2} \right\rfloor, j+q-\left\lfloor \frac{K}{2} \right\rfloor} \right\}, \cdots 0 \leq p, q < K, \tag{1}$$

where, $\left\lfloor \frac{K}{2} \right\rfloor$ represents the floor function of $\frac{K}{2}$.

In Outlook attention, the outlook weights at position (i, j) can be directly used as the aggregated attention weights by reshaping them into $\widehat{\mathrm{A}}_{\mathrm{i,j}} \in \mathbb{R}^{\mathrm{K}^2\times \mathrm{K}^2}$, followed by the softmax activation function. Consequently, the value projection process can be written as Eq. 2:

$$Y_{\Delta_{i,j}} = MatMul\left(Softmax(\hat{A}_{i,j}), V_{\Delta_{i,j}} \right), \tag{2}$$

where $\hat{A}_{i,j}$ is the reshaped outlook weights, and $V_{\Delta_{i,j}}$ represents the values within the local window centred at (i, j).

Outlook attention densely aggregates the projected value representations. By summing the differently weighted values from the same position across different local windows, the output result is obtained as shown in Eq. 3:

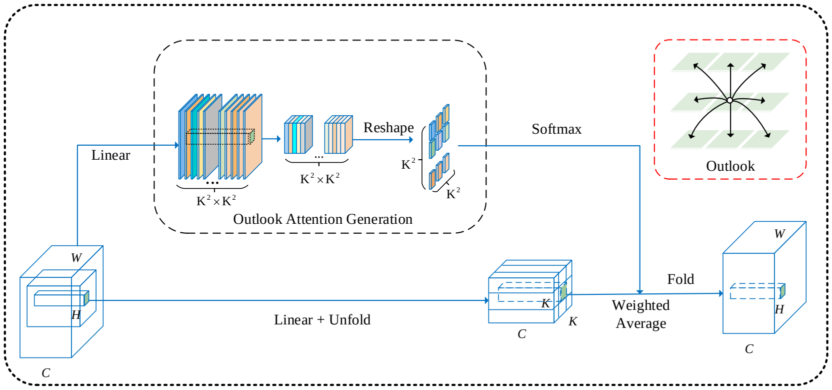


Fig. 5: Outlook attention structure diagram.

$$\tilde{Y}_{i,j} = \sum_{0\leq m,n<K} Y^{i,j}_{\Delta_{i+m-\left[\frac{K}{2}\right],\, j+n-\left[\frac{K}{2}\right]}}. \tag{3}$$

3.4 Loss Function

Pixel Level Contrastive Learning Contrastive learning is widely used in self-supervised learning. Its main idea is discrimination between positive and negative samples, primarily achieved by using a metric function to encourage the network to bring positive samples closer while pushing negative samples apart. In medical image segmentation, contrastive learning addresses the critical issue of sparse annotated samples in datasets while enhancing the model's generalization ability and augmenting the model's capacity for feature extraction [19].

Dual-MambaNet uses two decoders: a low-level decoder and a high-level decoder. The output of the low-level decoder is considered the segmentation result for generating pseudo-labels, while the output of the high-level decoder is regarded as the segmentation result for the accurate labels. Inspired by [16], we propose using our novel improved multi-level pixel contrastive loss function(MPCL) between the outputs of the two decoders. This approach optimizes the output of the high-level decoder based on the output of the low-level decoder, thereby improving the final output of the model.

Considering that in brain images, each tissue has a relatively small size and many pixels belong to the background, these background pixels do not provide sufficient features for the network. Therefore, we propose using adaptive average pooling to filter out unimportant background pixels, enhancing the model's feature extraction capability. Additionally, we apply L2 regularization on the channel dimension to sparsify the features, thereby improving the model's generalization and robustness. Specifically, our proposed multi-level pixel contrastive loss function(MPCL) can be expressed as Eq. 4:

$$\mathcal{L}_{\mathrm{MPCL}} = \frac{\sum \|(G(D_\theta(D_L \cup D_H)), G(D_\theta(D_H))\|_2^2}{N}, \tag{4}$$

where D_θ is the decoder using AdaptiveAvgPool, G is the L2 regularization operation along the channel axis, and N is the number of input data. D_L and D_H represent the outputs of the low-level and high-level decoders, respectively, and $\cup$ denotes the union operation. To effectively utilize the low-level decoder output to optimize the high-level decoder output, we consider the high-level decoder output as the low-level decoder output to maximize the distance between different level outputs, thereby improving the model's performance.

Toatal Loss In our model, decoders and labels use standard cross-entropy loss($\mathcal{L}_{CE}$) and Dice loss functions($\mathcal{L}_{Dice}$). A multi-level pixel contrastive loss function(MPCL) is also used between the outputs of the two decoders.

The total loss is defined as Eq. 5:

$$\mathcal{L}_{\mathrm{total}} = \lambda \left(\frac{\frac{\mathcal{L}^{\mathrm{L}}_{\mathrm{Dice}} + \mathcal{L}^{\mathrm{L}}_{\mathrm{CE}}}{2} + \frac{\mathcal{L}^{\mathrm{H}}_{\mathrm{Dice}} + \mathcal{L}^{\mathrm{H}}_{\mathrm{CE}}}{2}}{2} \right) + (1 - \lambda)\mathcal{L}_{\mathrm{MPCL}}, \tag{5}$$

where $\mathcal{L}^L_{Dice}$and $\mathcal{L}^H_{Dice}$represent the Dice losses for the low-level and high-level feature outputs, respectively. Similarly, $\mathcal{L}^L_{CE}$ and $\mathcal{L}^H_{CE}$ represent the cross-entropy losses for the low-level and high-level feature outputs, respectively. $\mathcal{L}_{\mathrm{MPCL}}$ represents the multi-level pixel contrastive loss function between the high-level and low-level feature outputs. The weighting factor *lambda*, set empirically to 0.9, balances the contributions between the contrastive loss and the other loss functions. As shown in Fig. 6, we illustrate the process of the proposed dual-branch decoder and multi-level pixel contrastive loss function collaboratively optimizing the final output. In Fig. 6, D_H represents the high-level decoder branch, and D_L represents the low-level decoder branch.

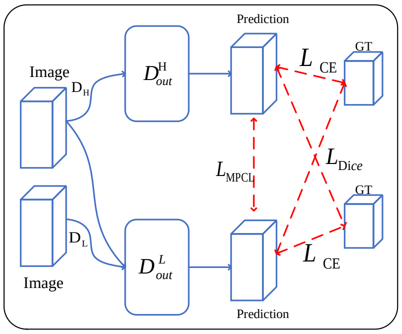


Fig. 6: A diagram of the dual-branch decoder framework based on multi-level pixel contrastive learning.

4 Experiments and Results

4.1 Datasets

OASIS1: The OASIS-1 dataset [11] used in this experiment is from the Open Access Series of Imaging Studies (OASIS). It comprises 421 subjects aged between 18 and 96 years. Each subject has a T1-weighted magnetic resonance imaging (MRI) scan. The dataset labels classify brain tissue into the cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM).
MICCAI 2013 MR BRAIN IMAGE SEGMENTATION: The MRBrainS13 challenge dataset consists of 20 subjects acquired using a 3.0T Philips Achieva MR scanner at the University Medical Center Utrecht, Netherlands [13]. The dataset includes multi-sequence MRI brain scans, such as T1, T1-IR and T2-FLAIR used for the challenge. The dataset labels classify brain tissue into the cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM).

4.2 Implementation Details

All experiments were conducted on a GeForce RTX 3090Ti GPU system with 24GB memory and Ubuntu 22.04, Python 3.8.19, PyTorch 2.2.0, and CUDA 11.8. The model is used for 2D medical image segmentation. We randomly split the two datasets into training, testing, and validation sets in an 8:1:1 ratio. All images were normalized and resized to 224×224, and data augmentation techniques, including vertical flip, horizontal flip, and random rotation, were applied. The Dual-MambaNet model was trained for 40,000 iterations with a batch size 24. The AdamW optimizer was used with a learning rate of 1e-4 and a weight decay set to 1e-4. Network performance was evaluated on the validation set every 200 iterations, and model weights were saved only when the new best performance was achieved on the validation set.

4.3 Comparison Methods

To ensure a fair comparison, the baseline methods (UltraLight VM-UNet) [18], Mamba-UNet [17], UNet [14], and Swin-Unet [1] were also trained under the same hyperparameter configurations without loading pre-trained models. We directly compared Dual-MambaNet with the baseline method (UltraLight VM-UNet) and other methods based on CNN, Transformer and Mamba.

4.4 Evaluation Metrics

This study also employed three objective evaluation metrics for quantitative comparison of our proposed method: (1) Similarity Measurement: Dice coefficient (denoted by an upward arrow ↑), where values closer to 1 indicate better performance. (2) Difference Measurements: Hausdorff Distance (HD) 95% and Average Surface Distance (ASD) (both denoted by a downward arrow ↓), where lower values are better, indicating higher similarity between the predicted segmentation and the ground truth.

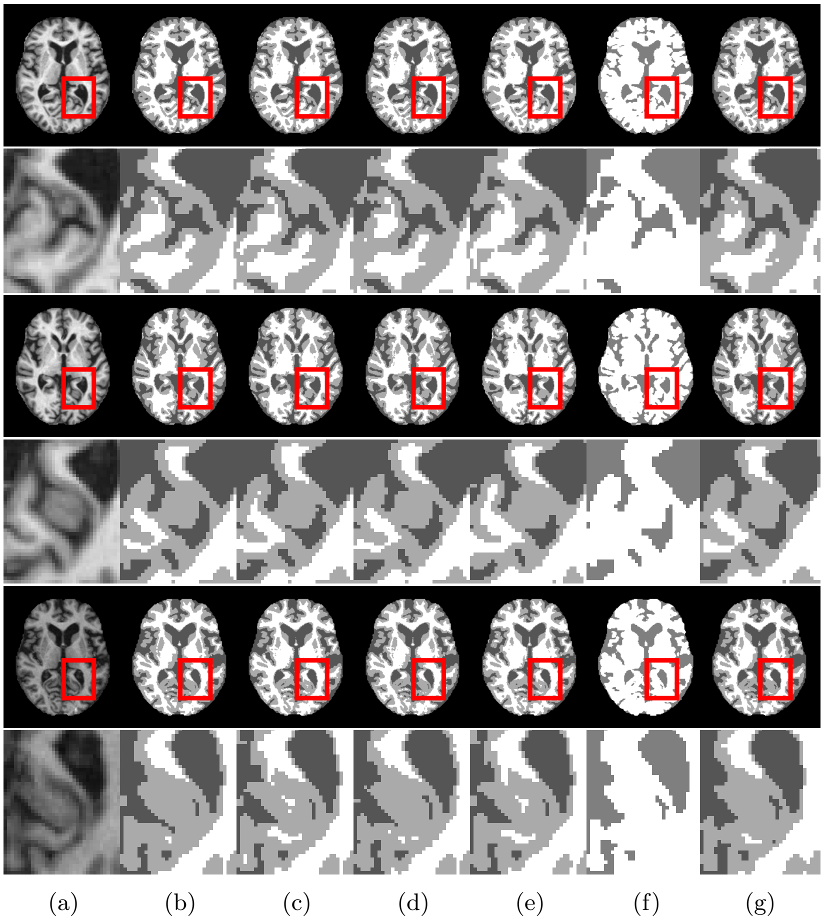


Fig. 7: Segmentation results comparison among different models on the OASIS1 dataset, with localized zoom-in comparison. (a)Image. (b) GT. (c) Mamba-UNet. (d) Swin-UNet. (e) UNet. (f) UltraLight VM-UNet. (g) Dual-MambaNet.

4.5 Qualitative Results

Fig. 7 and Fig. 8 present three randomly selected original image samples from the OASIS1 and MRBrainS13 datasets. They compare the segmentation results of all baseline methods, including Dual-MambaNet, on the OASIS1 and MRBrainS13 datasets, along with zoomed-in views of local details.

As shown in the results of Fig. 7 and Fig. 8, as well as the enlarged views of local details, Dual-MambaNet can segment all categories completely compared to the Baseline (UltraLight VM-UNet). It can extract local features better while also capturing high-level semantic information. Compared to other classic models based on CNN, Transformer, and Mamba, Dual-MambaNet can also fully extract local features. As seen in the enlarged local view in Fig. 7, Dual-MambaNet can

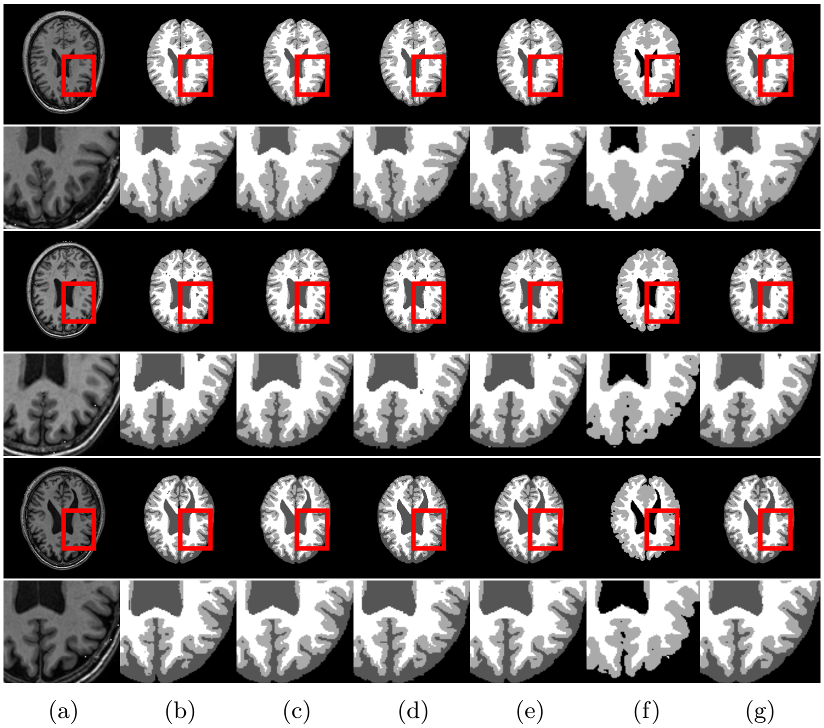


Fig. 8: Segmentation results comparison among different models on the MRBrainS13 dataset, with localized zoom-in comparison. (a)Image. (b) GT. (c) Mamba-UNet. (d) Swin-UNet. (e) UNet. (f) UltraLight VM-UNet. (g) Dual-MambaNet.

recognize more complex local features while maintaining the integrity of global information. As shown in the enlarged local view in Fig. 8, Dual-MambaNet can better recognize edge information and maintain the integrity of global features. In the above analysis, Dual-MambaNet can fully extract global features while better capturing complex textures and structural features such as edges.

4.6 Quantitative Results

Table. 1 and Table. 2 directly compare Dual-MambaNet with other segmentation networks on the OASIS1 and MRBrainS13 datasets, respectively, including similarity and difference metrics. The best-performing results are highlighted in bold, and '—' indicates that the model did not segment that category.

Quantitative results indicate that on large-scale datasets, Dual-MambaNet performs comparably to CNN-based and Transformer-based models on some metrics while surpassing classical models and the latest Vision Mamba models on others. Additionally, Dual-MambaNet has lower parameter counts and GFLOPs.

Table 1: Comparison of objective evaluation metrics of models on the OASIS1 dataset.

Model	Dice(↑)			HD95(↓)			ASD(↓)			Para(M)	GFLOPs(G)
	CSF	GM	WM	CSF	GM	WM	CSF	GM	WM		
Mamba-UNet	0.9118	0.9114	0.8905	1.2102	1.7833	3.4910	0.3183	0.5163	1.0448	28.00	5.99
UNet	0.8719	0.8717	0.8793	1.4852	1.9535	4.0379	0.3059	0.6240	1.4087	1.81	2.3
Swin-UNet	0.9160	0.9187	0.9098	1.2913	**1.2347**	**1.8963**	**0.2536**	0.4269	0.4949	41.34	8.71
UltraLight VM-UNet	0.8528	0.7177	—	1.6120	6.5613	—	0.5413	0.4943	—	**0.049**	**0.04**
Dual-MambaNet	**0.9161**	**0.9198**	**0.9136**	**1.1439**	1.3626	2.5324	0.3222	**0.3771**	**0.2542**	0.10	0.08

Table 2: Comparison of objective evaluation metrics of models on the MRBrainS13 dataset.

Model	Dice(↑)			HD95(↓)			ASD(↓)			Para(M)	GFLOPs(G)
	CSF	GM	WM	CSF	GM	WM	CSF	GM	WM		
Mamba-UNet	0.6655	0.6918	0.7150	2.3588	3.4587	4.2645	0.5810	1.0214	1.3920	28.00	5.99
UNet	0.6614	0.7003	0.7327	2.1166	2.5988	5.0152	**0.4488**	0.7471	1.7367	1.81	2.3
Swin-UNet	0.6683	0.7045	**0.7417**	**1.7946**	**1.5695**	4.5107	0.4843	**0.4716**	**1.3532**	41.34	8.71
UltraLight VM-UNet	—	0.6219	0.6895	—	4.8375	4.7374	—	1.4408	1.6855	**0.049**	**0.04**
Dual-MambaNet	**0.6697**	**0.7077**	0.7199	2.3193	2.5511	**4.2528**	0.5565	0.7432	1.5689	0.10	0.08

Compared to UltraLight VM-UNet, Dual-MambaNet significantly improves segmentation accuracy with a slight increase in complexity. Dual-MambaNet also demonstrates good generalization ability and robustness on small datasets, accurately predicting segmentation masks. Despite a slight increase in parameters and GFLOPs compared to the baseline model (UltraLight VM-UNet), Dual-MambaNet significantly enhances segmentation performance. Dual-MambaNet achieves higher segmentation accuracy than other methods while maintaining lower parameter counts and GFLOPs.

4.7 Ablation Study

Dual-MambaNet involves three key components: 1) Outlook Attention; 2) Dual Decoder Branches; 3) Multi-Level Pixel Contrastive Loss(MPCL). We compare the parts proposed in this study through ablation studies. To validate the effectiveness of the proposed model and its improvements, extensive ablation experiments were conducted on the MRBrainS13 dataset, using Dice and HD95, to evaluate the performance of each component quantitatively. The best-performing values are highlighted in bold, and '—' indicates that the model did not segment that category. The results are shown in Table. 3. In this table, 'Atten' represents the improved Outlook attention, 'Double' represents the duale decoder branches structure, and 'MPCL' represents the multi-level pixel contrastive loss function. Additionally, ✓ indicates that the component is used and ✗ indicates that the component is not used.

Table 3: Comparison of Ablation Experiment Results.

Model			Dice(↑)			HD95(↓)		
Double	Atten	MPCL	CSF	GM	WM	CSF	GM	WM
✗	✗	✗	—	0.6219	0.6895	—	4.8375	4.7374
✓	✗	✗	0.6348	0.6627	0.6672	2.9957	3.9822	7.0246
✓	✗	✓	0.6414	0.6629	0.6845	2.9074	3.3953	5.3508
✓	✓	✗	0.6370	0.6727	0.6787	3.0993	3.4120	6.8035
✗	✓	✗	0.6321	0.6542	0.6654	3.0875	3.9764	6.9864
✓	✓	✓	**0.6697**	**0.7077**	**0.7199**	**2.3193**	**2.5511**	**4.2528**

As shown in Table 3, although using local attention alone for feature extraction improves the model's accuracy, the Dual-MambaNet with the dual-branch decoder captures the complex features of brain MRI images more effectively. This indicates that the dual-branch decoder enhances the model's ability to couple multi-level information, thereby improving segmentation accuracy. Furthermore, using the pixel-level contrastive loss function for output optimization further improves segmentation accuracy, demonstrating that this loss function strengthens the coupling ability of the dual-branch decoder. While using any single component alone can improve segmentation performance, the model achieves the best performance when all three components are combined. These results show that Dual-MambaNet can segment brain MRI images with high accuracy.

5 Conclusion

This paper addresses the performance bottlenecks and loss of local feature information in brain tissue segmentation of high-resolution MRI images by proposing the Dual-MambaNet model. This model combines the Outlook attention module with Mamba to construct a feature extractor (FE) encoder layer, effectively connecting local and global features. Additionally, dual decoder branches and a multi-level pixel contrastive loss function (MPCL) are introduced to optimize feature representation. Experimental results on the OASIS1 and MRBrainS13 datasets demonstrate that Dual-MambaNet achieves high segmentation accuracy with lower parameters and GFLOPs, making it suitable for deployment in resource-limited medical environments. This research provides a promising solution for medical image segmentation under constrained computational resources.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
5. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
6. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059. IEEE (2020)
7. Li, Z., Zhang, C., Zhang, Y., Wang, X., Ma, X., Zhang, H., Wu, S.: Can: Context-assisted full attention network for brain tissue segmentation. Medical Image Analysis **85**, 102710 (2023)
8. Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L.: Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. arXiv preprint arXiv:2403.05246 (2024)
9. Liu, C., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei-Fei, L.: Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 82–92 (2019)
10. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1013–1023 (2021)
11. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
12. Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. arXiv preprint arXiv:2206.13947 (2022)
13. Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Biessels, G.J., Viergever, M.A.: MR Brain Segmentation Challenge 2013 Data (2024). https://doi.org/10.34894/645ZIN, https://doi.org/10.34894/645ZIN
14. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)

15. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1150–1156. IEEE (2022)
16. Wang, Z., Ma, C.: Semi-mamba-unet: Pixel-level contrastive cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. arXiv preprint arXiv:2402.07245 (2024)
17. Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079 (2024)
18. Wu, R., Liu, Y., Liang, P., Chang, Q.: Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. arXiv preprint arXiv:2403.20035 (2024)
19. You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S.: Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging **41**(9), 2228–2237 (2022)
20. Yu, W., Wang, X.: Mambaout: Do we really need mamba for vision? arXiv preprint arXiv:2405.07992 (2024)
21. Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S.: Volo: Vision outlooker for visual recognition. IEEE transactions on pattern analysis and machine intelligence **45**(5), 6575–6586 (2022)
22. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: 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. pp. 3–11. Springer (2018)
23. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)