

SuperLightNet: Lightweight Parameter Aggregation Network for Multimodal Brain Tumor Segmentation

Feng Yu^{1,2,3} Jiacheng Cao¹ Li Liu¹ Minghua Jiang^{1,3,†}

¹Wuhan Textile University, China ²Nanyang Technological University, Singapore

³Engineering Research Center of Hubei Province for Clothing Information, China

Abstract

Multimodal 3D segmentation involves a significant number of 3D convolution operations, which requires substantial computational resources and high-performance computing devices in MRI multimodal brain tumor segmentation. The key challenge in multimodal 3D segmentation is how to minimize network computational load while maintaining high accuracy. To address the issue, a novel lightweight parameter aggregation network (SuperLightNet) is proposed to realize the efficient encoder and decoder for the high accurate and low computation. A random multiview drop encoder is designed to learn the spatial structure of multimodal images through a random multi-view approach for solving the high computational time complexity that has arisen in recent years with methods relying on transformers and Mamba. A learnable residual skip decoder is designed to incorporate learnable residual and group skip weights for addressing the reduced computational efficiency caused by the use of overly heavy convolution and deconvolution decoders. Experimental results demonstrate that the proposed method achieves a leading reduction in parameter count by 95.59%, the 96.78% improvement in computational efficiency, the 96.86% enhancement in memory access performance, and the average performance gain of 0.21% on the BraTS2019 and BraTS2021 datasets in comparison with the state-of-the-art methods. Code is available at <https://github.com/WTU-MIS-Laboratory/SuperLightNet>.

1. Introduction

Brain tumors are a serious type of neurological disease that can lead to severe neurological dysfunction and even be life-threatening. The choice of treatment varies depending on the type, size, and location of the tumor. Accurate diagnosis and segmentation are crucial for developing an effective treatment plan. Traditional brain tumor segmentation typically relies on manual annotation by radiologists, which is

both time-consuming and labor-intensive. It is also prone to subjective factors, leading to inconsistent segmentation results. Therefore, automated MRI brain tumor segmentation methods have garnered widespread attention.

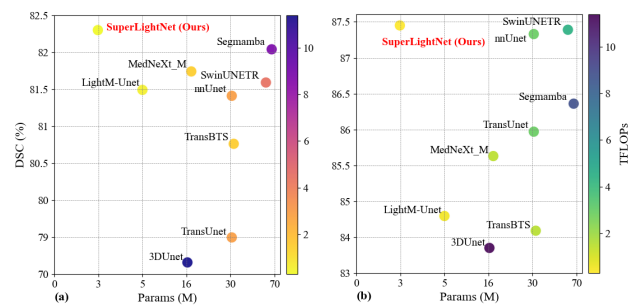


Figure 1. (a) and (b) illustrate the statistical results of lightweight and segmentation performance for BraTS2019 and BraTS2021, respectively. The X-axis represents the number of parameters (smaller is better), and the Y-axis represents the DSC (higher is better). The color depth gradient from light to dark indicates the TFlops(lighter is better).

Deep learning methods have achieved significant success in the field of brain tumor segmentation. The Convolutional Neural Networks (CNNs) are widely used in medical image segmentation, such as UNet [23], UNet++ [33], V-Net [21], DeepMedic [15], HighRes3DNet [16], and nnUnet [14]. These networks are capable of efficiently extracting and processing complex features from medical images. Researchers have progressed from using 2D convolutions to 3D convolutions and from shallow to deep multimodal fusion methods, and there is increasing focus on network architectures that emphasize automated processing and adaptive fusion.

To enhance UNet’s ability to grasp global information, researches have explored incorporating Transformer [27] architectures, utilizing self-attention mechanisms to capture global context by treating images as sequences of adjacent patches. While this approach has proven effective, Transformer-based methods introduce quadratic complexity

[†]Corresponding author

with respect to image size due to the self-attention mechanism. This results in significant computational overhead, especially for tasks requiring dense predictions like medical image segmentation. Consequently, these methods [4, 22] often fail to meet the critical need for computational efficiency in real-world medical applications, where models with low parameter counts and minimal computational demands are essential for mobile healthcare segmentation tasks. In addition, State Space Models (SSMs) [7] have attracted significant attention in the research community recently. Building on the foundations of traditional SSM research, such as Mamba, these models not only establish long-range dependencies but also demonstrate linear complexity with respect to input size. It makes Mamba as a strong contender alongside CNNs and Transformers in the quest for lightweight UNet models. Some recent approaches, like U-Mamba [19], have introduced a hybrid CNN-SSM block that combines the local feature extraction capabilities of convolutional layers with the ability of SSMs to capture long-range dependencies. However, U-Mamba comes with a considerable increase in parameters and computational demands, making it difficult to deploy in lightweight environments for medical segmentation tasks.

Inspired by certain 2D lightweight networks [25], we aim to achieve low latency and precise segmentation in multimodal tasks through efficient lightweight strategies, which is crucial for enhancing medical efficiency and ensuring patient safety. To achieve efficient medical image segmentation, we propose a lightweight multimodal segmentation algorithm. By designing a lightweight and robust random multiview drop encoder and a learnable residual skip decoder, the algorithm achieves precise segmentation while significantly reducing the complexity of the network model and minimizing computational load. The comparison of SuperLightNet’s lightweight performance with the state-of-the-art methods is illustrated in Fig. 1

This paper makes the following contributions:

- We design a random multiview drop encoder with minimal parameters (2.71M), which learns the spatial structure of multimodal images through a random multiview approach. Combined with proposed drop residual (dropRes) model, it enhances robustness, offering a lightweight solution with improved performance compared to other methods.
- We propose a learnable residual skip decoder with the goal of extreme lightweight network. It incorporates learnable residual and group skip weights, replacing fully convolutional feature extraction. Additionally, the non-linear layers are split using a specific channel separation approach to further reduce the number of parameters. This design achieves extreme lightweighting of the multimodal decoder (0.26M) while maintaining commendable performance.
- We design a lightweight parameter aggregation network for multimodal brain tumor segmentation, which is an encoder-decoder architecture that combines robustness with extreme lightweight design. It enables highly efficient 3D multimodal segmentation. Compared to the state-of-the-art methods, our network demonstrates a leading reduction in parameter count by 95.59% (2.97M), the 96.78% improvement in computational efficiency (0.282 TFlops), the 96.86% enhancement in memory access performance (45.8G), and an average performance gain of 0.21% on the BraTS2019 and BraTS2021 datasets.

2. Related Work

2.1. Brain Tumor Segmentation

Brain tumor segmentation is a particularly challenging task in medical imaging. Traditionally, many segmentation frameworks are based on the UNet architecture, which features a symmetrical design with a contracting path to capture contextual information and an expanding path for accurate localization. nnUNet [14] is a highly effective method for brain tumor segmentation, introducing improvements such as postprocessing, region-based training, and robust data augmentation. Transformer-based approaches have also shown great potential in this area. For example, TransBTS [29] combines Transformers with 3D CNNs, taking advantage of the 3D CNN’s ability to model local context while using Transformers to capture global semantic relationships. Furthermore, Swin Transformer blocks are integrated into the UNETR [9] model, where features from the Swin Transformer [8] encoder are connected to the FCNN-based decoder at various resolutions via skip connections.

The above methods improve segmentation by using postprocessing, data augmentation, introducing Transformers to capture global information, and employing shift-windows for multi-scale [28] feature extraction. However, these methods typically come with high computational or memory overhead. Our study focuses on enhancing computational efficiency while maintaining segmentation accuracy, addressing computational efficiency limitations in previous methods.

2.2. Lightweight 3D Segmentation Network

Lightweight 3D segmentation networks are pivotal in computer vision and deep learning, with broad applications in medical imaging, autonomous driving, and augmented reality. In medical imaging, they enable real-time processing by accelerating inference speed without compromising segmentation accuracy.

The classic UNet architecture has been widely used in 3D image segmentation. The lightweight version of 3D UNet [6] reduces computational cost by decreasing the

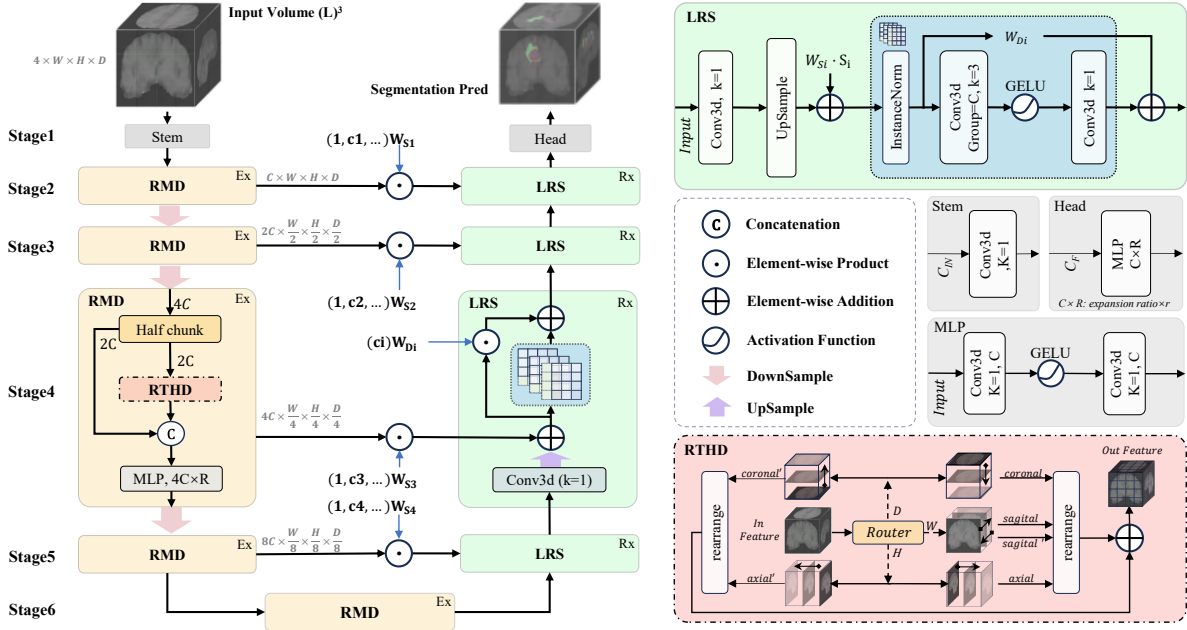


Figure 2. The workflow of our proposed SuperLightNet. The lightweight RMD encoder stacks based on Ex , learning to align with the latent space of $1.0 \times 1.0 \times 1.0 mm^3$ multimodal data while being highly computationally efficient. The ultra-lightweight LRS decoder, with only $0.27M$ parameters, reconstructs the latent space information, utilizing group-weighted skip connections to determine the appropriate fusion ratio. The entire network without any unnecessary operations.

number of channels and employing more efficient convolutional layers. VoxResNet [5] incorporates the concept of residual networks and has been further designed for lightweight efficiency. Mobile3DNet is optimized specifically for 3D data based on the lightweight principles of MobileNet [12]. ShuffleNet3D leverages grouped convolutions and channel shuffling mechanisms from ShuffleNet [32] to achieve a lightweight design. EGE-Unet [25] achieves advanced lightweight performance in 2D medical segmentation tasks by using a quarter-channel dot-product attention mechanism. TransNuSeg [11] reduces model complexity by replacing the Transformer bottleneck with a token Multilayer Perceptron (MLP) bottleneck, enabling lightweight segmentation for 2D nuclei tasks. LETNet [26]’s streamlined encoder-decoder with a lightweight dilated bottleneck achieves efficient 2D segmentation with fewer parameters. LightM-Unet [17] utilizes the Mamba structure and lightweight upsampling to achieve lightweight segmentation in a single modality.

The above methods optimize 3D data by reducing the number of channels, splitting channels, improving convolutional layers, and using group convolutions to enhance computational efficiency. In contrast, this paper further explores the overall architecture and feature extractors, re-designing structures such as skip connections and drop residual (dropRes), with a focus on how lightweight mul-

timodal medical segmentation can achieve enhanced efficiency through optimized feature extractors and overall architecture.

3. Method

3.1. Overview

To enhance the overall lightweight efficiency of the network, we propose an extremely lightweight Random Multiview Drop (RMD) encoder, combined with an efficient Learnable Residual Skip (LRS) decoder featuring adaptive skip residual weights. The bottleneck layer utilizes the same lightweight encoder. Our overall framework, as shown in Fig.2, follows the encoder-decoder architecture [10, 13, 24], which is highly effective in graphical processing.

Given an input voxel shape of $M \times W \times H \times D$, where M , W , H , and D represent the dimensions, the input first passes through a fully convolutional stem layer, generating shallow feature map voxels of $C \times W \times H \times D$, where C is 24, representing a fixed number of filters. Next, we introduce a lightweight encoder with a layer depth of E_x , where $E_x \in (1, 1, 2, 2)$, for random multi-view feature extraction. After each group of feature extraction layers, a downsampling layer is applied, doubling the number of channels while halving the voxel resolution. Finally, we set the layer

depth of the bottleneck layer to 2, making a total of 5 feature fusion blocks, including the bottleneck layer.

Instead of the traditional fully convolutional decoder in UNETR [9], we introduce a super lightweight upsampling decoder with layers of depth $R_x \in (1, 1, 1, 1)$, matching the architecture depth 4. This decoder utilizes learnable group skip weights, learnable scale residual weights, and employs linear interpolation combined with fixed group MLP instead of transposed convolutions, maintaining accuracy while ensuring a lightweight design. Unlike the encoder, in the decoder, the operations of halving the channels and doubling the voxel resolution are both completed within each block.

3.2. Random Multiview Drop Encoder

To achieve lightweight multimodal 3D segmentation, Random Tri-view Hadamard Drop (RTHD) employs a random axis drop residual mechanism, which focuses on extracting the most relevant segmentation features with minimal parameters, while maintaining a lightweight design, as shown in Fig.3. Given the input data $\mathcal{X}_{in} \in \mathbb{R}^{M \times W \times H \times D}$, it first passes through the fully convolutional stem, followed by the RTHD and DownSample modules, where RTHD consists of R_x layers, and DownSample is always 1 layer.

Specifically, given a voxel volume of \mathcal{X}_{in} , where the resolutions of the input channels C are $W \times H \times D$, the encoder first splits the input into $1/2C$, with half of it \mathcal{X} undergoing InstanceNorm3d for normalization, followed by RTHD. The other half $\hat{\mathcal{X}}$ does not participate in the gradient regression RTHD learning and is directly merged into the subsequent non-linear layer. The MLP uses a 3D-convolution-GELU-3D-convolution structure, with the MLP's channel expansion ratio set to 2.

The RTHD encoder is the core operator, which uses a random factor $P(\cdot) \sim \text{Rand}\{0, 1, 2\}$ to generate an integer between 0-2. It then contains three branches corresponding to the anatomical coronal, axial, and sagittal axes, each serving as a lightweight interpreter (lightweight feature extractor) branch. These three branches transform the $\mathcal{X} \in \mathbb{R}^{(B,C,W,H,D)}$ data into

$$\begin{cases} \mathcal{X}_{coronal} \in \mathbb{R}^{(H \times B, C, W, D)} \\ \mathcal{X}_{axial} \in \mathbb{R}^{(W \times B, C, H, D)} \\ \mathcal{X}_{sagittal} \in \mathbb{R}^{(D \times B, C, W, H)} \end{cases} \quad (1)$$

Along with their inverted forms $\mathcal{X}_{coronal}^r$, \mathcal{X}_{axial}^r , $\mathcal{X}_{sagittal}^r$, extracting features from three randomly assigned axial and inverse axial views. Each randomly assigned axial view is fed into a 2D encoder Group Hadamard Product Attention (G_{hpa}). The G_{hpa} splits the input data into four groups, the first three performing Hadamard Product Attention (HPA) operations on the three axes, the other using only Depthwise Separable Convolution (DW) on the feature map, and finally connecting them along the channel. The process of RTHD can be expressed as

$$RTHD(\mathcal{X}) = P(\cdot) \sim \{G_{hpa}(\mathcal{X}_i + \mathcal{X}_i^r)\}_{i=1}^3 \quad (2)$$

where $P(\cdot)$ represents feature extraction in the three directions of $i \in \{Coronal, Axial, Sagittal\}$, \mathcal{X}_i and \mathcal{X}_i^r represent axial and inverse axial, respectively. G_{hpa} denotes performs a group multi-axis hadamard product attention operation on both axes and adds the results. The whole encoder process can be expressed as

$$RMD(X) = MLP((RTHD(Norm(\mathcal{X})), \hat{\mathcal{X}})_{Cat}) \quad (3)$$

where \mathcal{X} and $\hat{\mathcal{X}}$ indicate $\frac{1}{2}C$ partitioning of input data by channel, and $Norm$ indicates instance normalization. Cat represents splicing the split channels into the dimensions of the input data, and MLP represents the nonlinear layer containing the shrink process, using a full convolution 3D and GELU structure with a channel expansion rate of 2.

The purpose of this design is that network overfitting tends to increase with the number of convergence iterations, often due to overly homogeneous feature extraction. Therefore, this module is designed to prevent ineffective learning while significantly reducing the number of parameters. The parameter count for the feature extraction part including the bottleneck layer is less than 2 million.

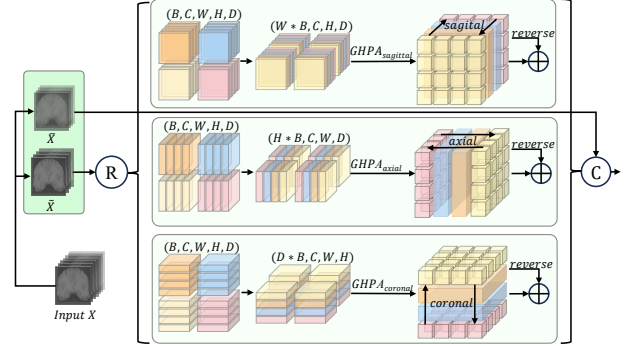


Figure 3. Random multiview drop encoder. By applying a random function to assign different lightweight perspectives to voxel features without parallel computation, it enhances robustness while reducing computational and parameter overhead. During inference, the computational load of this module is further halved.

3.3. Learnable Residual Skip Decoder

Both the lightweight 3D segmentation encoder and decoder are essential. Fully convolutional decoders (referencing UNETR and other similar structures like SwinUNETR) have been proven to be effective, but the computational cost and parameter count present significant challenges for lightweight designs. In this paper, we propose a lightweight interpreter for learnable residual skip upsampling.

Specifically, it accepts input from the bottleneck with dimensions $\mathcal{F}_{bn} \in \mathbb{R}^{16C \times \frac{1}{16}W \times \frac{1}{16}H \times \frac{1}{16}D}$ and is designed according to a combination of upsampling and interpretation. Delving deeper, the input data \mathcal{F}_{in} first passes through a fully convolutional layer for channel (hidden feature layer) feature fusion, followed by a non-trainable linear upsampling. We have discarded transposed convolutions due to their high parameter and computational costs. The upsampled data $Up(\cdot)$ is combined with a new feature data \mathcal{F}_{upr} through group-weighted W_{S_i} shape in $(1, C, \dots)$, skip \mathcal{S}_i addition with learnable parameters. This process can be expressed as

$$\mathcal{F}_{upr} = Up(Conv^{k=1}(\mathcal{F}_{in})) + W_{S_i} \odot \mathcal{S}_i \quad (4)$$

And a residual structure is introduced, followed by non-linear mapping via an InstanceNorm and fixed-value group convolutional MLP. Here, the group is set to 12, and GELU is used. This process can be expressed as

$$Repr(\mathcal{F}) = Norm(MLDW(\mathcal{F}_{upr})) \quad (5)$$

Finally, the result is merged with the scale residual through learnable weighted scaling to complete the upsampling, producing the output interpreter result.

$$Dec(\mathcal{F}) = MLDW(\mathcal{F}_{upr}) + W_{Di} \odot \mathcal{F}_{upr} \quad (6)$$

4. Experiments

4.1. Datasets

We evaluate our proposed lightweight parameter aggregation network on the BraTS2019 [1, 2, 20] and BraTS2021 [3, 30] datasets, which are widely used in current medical segmentation tasks. The BraTS datasets are multimodal (T1, T1ce, T2, FLAIR) MRI datasets. For each patient, the BraTS data is registered to a common spatial resolution and undergoes skull stripping. The four raw modality volumes for each dataset are uniformly sized at $240 \times 240 \times 155$. Within each volume, the brain tumor region is segmented into three primary sub-regions: the enhancing tumor (ET) region, the tumor core (TC) region (including the enhancing tumor and necrotic areas), and the whole tumor (WT) region (including the tumor core and edema areas).

In the experiments on BraTS2019, we employ five-fold cross-validation to ensure the stability and reliability of the results. In each fold, 80% of the data is used for training, while 20% is used for validation. In the experiments on BraTS2021, the data is randomly divided into 7:1:2 ratio for the training, validation, and test sets, with the validation set not participating in gradient training.

4.2. Evaluation Metrics

We utilize the Dice Similarity Coefficient (DSC) and Surface Dice Coefficient (SDC) as evaluation metrics, which

are commonly used to assess the accuracy of segmentation. The DSC measures the overall segmentation quality, while the SDC evaluates the accuracy of the segmented surface.

4.3. Implementation Details

All models are implemented using Python 3.8 and PyTorch 2.2, in combination with the MONAI 1.3 framework, and parameter evaluation is conducted using callops [31]. In the experiments, the models are trained for about 40,000 iterations on four A100 GPUs, with a batch size of 2 per GPU. We employ a cosine annealing learning rate scheduler to dynamically adjust the optimizer’s learning rate, with an initial learning rate set to 1e-3, a weight decay of 1e-2, and use the AdamW optimizer along with the binary cross entropy With DiceLoss function. The input 3D data is cropped to a uniform size of $128 \times 128 \times 128$. To increase data diversity, random flip augmentation is applied, and random intensity shift is used to enhance the data.

4.4. Comparison with Others

Table.1 presents the experimental results of all models on BraTS2019, conducted using the 5-CV mode. Our SuperLightNet achieves state-of-the-art (SOTA) segmentation performance and model parameter efficiency on this dataset, with comprehensive performance scores of 82.30% DSC and 73.23% SDC, outperforming the previous SOTA performance of Segmamba, which has 82.04% DSC and 72.95% SDC, by 0.26% and 0.28%, respectively. Compared to other models, our approach demonstrates even greater advantages. Most notably, our parameter count is reduced by 95.59% compared to Segmamba. Compared to the contemporaneous model LightM-Unet, which features a very lightweight parameter count of 5.02M, our approach surpasses it in the multimodal segmentation task by achieving 0.81% higher DSC and 1.81% higher SDC, all while using fewer parameters (2.97M) on the BraTS2019 task.

Table.2 presents the experimental results of all models on BraTS2021, conducts with the TVE mode. Our network, SuperLightNet, also achieves SOTA-level performance on this relatively abundant dataset. Specifically, SuperLightNet outperforms the current SOTA, SwinUNETR, by 0.06% in DSC and reaches a leading SDC of 82.99%. Notably, our model’s parameter count is 95.22% smaller compared to the SOTA.

Fig.4 shows a qualitative comparison of the results on BraTS2019. In the figure, we present the segmentation results of two samples using the Flair modality as the background. From these results, it is evident that during the 40,000 iterations of training, our SuperLightNet exhibits a more stable and rapid convergence performance. In case A, our network shows minimal regional tissue segmentation loss for the ED region and provides accurate predictions for the NCR. In case B, our network demonstrates better coher-

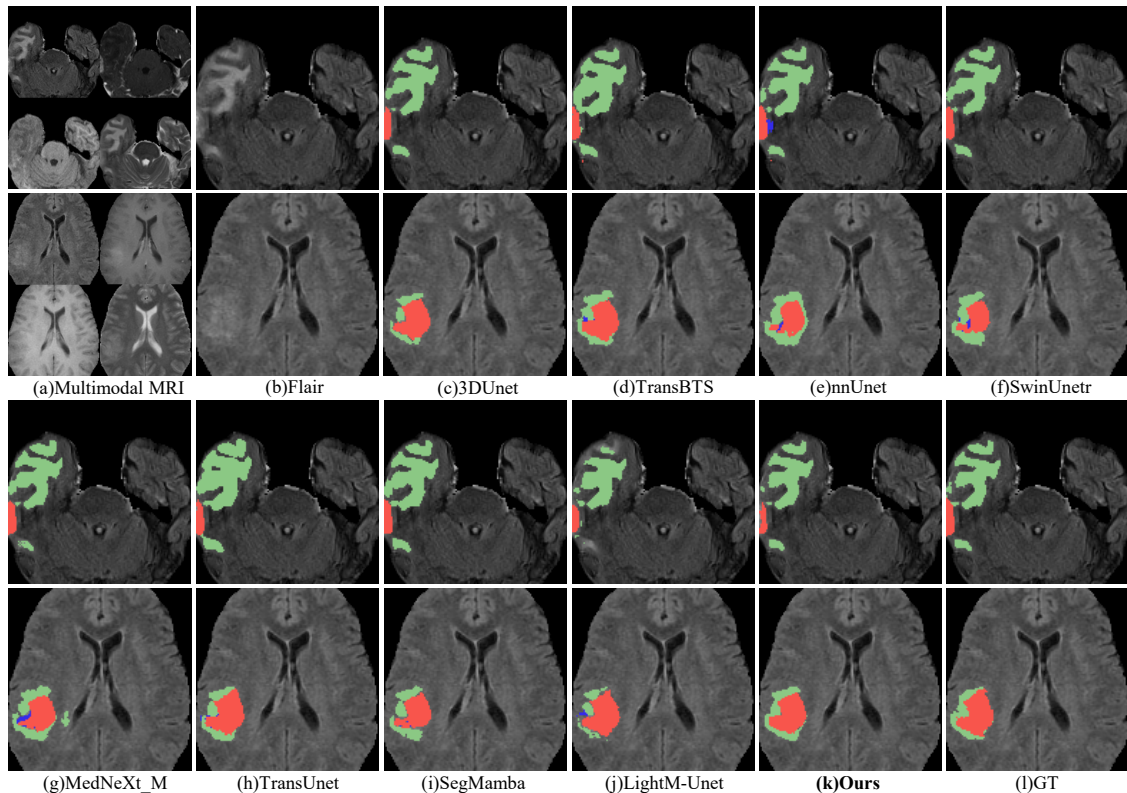


Figure 4. The visualized segmentation of BraTS datasets. The first and third rows correspond to case A, while the second and fourth rows correspond to case B, with the Flair modality used as the background. The distinctions between different cases are clearly observable. The blue and green regions represent the necrotic tumor core (NCR) and peritumoral edema (ED). The enhancing tumor (ET) is represented by the red region. The tumor core (TC) includes the combination of the red and blue regions. The enhancing whole tumor (WT) comprises the combination of the red, blue, and green regions.

Table 1. Quantitative evaluations on BraTS2019 validation set using 5-Cross-Validation(5-CV) Mode. The first 2 results are marked with **bold** and underlined.

| Iteration:39783 | | BraTS2019 | | | | | | | | | | | |
|----------------------------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Model | Params | subset1 | | subset2 | | subset3 | | subset4 | | subset5 | | Avg. | |
| | | DSC | SDC | DSC | SDC | DSC | SDC | DSC | SDC | DSC | SDC | DSC | SDC |
| 3DUnet(MICCAI,2016) | 16.32 | 69.20 | 58.40 | 70.71 | 59.40 | 76.64 | 63.80 | 77.96 | 63.87 | 69.35 | 57.64 | 72.77 | 60.62 |
| TransBTS(MICCAI,2021) | 32.99 | 79.57 | 68.73 | 75.82 | 63.77 | 83.04 | 74.03 | 85.50 | 76.85 | 79.86 | 71.56 | 80.76 | 70.99 |
| nnUnet(Nature Method,2021) | 31.19 | 79.70 | 69.18 | 79.35 | 68.90 | 82.34 | 73.78 | <u>85.68</u> | 76.89 | 79.98 | 71.26 | 81.41 | 72.00 |
| SwinUNETR(MICCAI,2021) | 62.19 | 79.48 | 69.41 | 80.01 | 70.07 | 82.66 | 74.20 | 85.51 | 77.09 | 80.27 | 71.91 | 81.59 | 72.54 |
| TransUnet(MIA,2024) | 31.20 | 77.60 | 65.55 | 77.69 | 67.02 | 78.93 | 69.11 | 84.29 | 74.86 | 75.81 | 65.01 | 78.86 | 68.31 |
| MedNeXt_M(MICCAI,2023) | 17.55 | <u>80.42</u> | 66.39 | 79.45 | 67.44 | 82.24 | 68.89 | 85.44 | 73.95 | 81.17 | 70.64 | <u>81.74</u> | 69.46 |
| SegMamba(MICCAI,2024) | 67.42 | 79.80 | <u>69.41</u> | 80.56 | <u>69.99</u> | 82.94 | 74.82 | 86.19 | 78.07 | 80.70 | <u>72.46</u> | 82.04 | <u>72.95</u> |
| LightM-Unet(Axriv,2024) | <u>5.02</u> | 79.83 | 68.55 | 78.79 | 67.12 | <u>83.22</u> | <u>74.57</u> | 85.61 | 76.59 | 80.00 | 70.25 | 81.49 | 71.42 |
| Ours | 2.97 | 80.72 | 70.78 | 80.26 | 68.95 | 83.78 | 75.49 | 85.65 | <u>77.76</u> | <u>81.11</u> | 73.19 | 82.30 | 73.23 |

ence in the ED region and superior boundary performance, while avoiding redundant predictions for the NCR area.

We utilize $4 * 128^3$ voxel images as the benchmark for performance evaluation. The metrics include Training Memory (TM), Inference Memory (IM), Training Time (TT), Tflops for forward propagation, combined forward

and backward propagation Tflops, Memory Access Cost per second (MACs), and parameter count. As shown in Table.3, when applying to volumes of identical dimensions, SuperLightNet uses only 54.14% of the training memory required by Segmamba, with training time reduced to less than a quarter, while still achieving state-of-the-art seg-

Table 2. Quantitative evaluations on BraTS2021 validation set using Train-Valid-Eval (TVE) Mode. The best and second-best are marked with **bold** and underlined. Notice, The * denotes data from the training and validation processes, used to compare with evaluation data to assess overfitting and generalization performance.

| Iteration:40250 | BraTS2021 | | | | | |
|-----------------|-------------|------------------|--------|-------|--------------|--------------|
| Model | Params | *Train | *Valid | | Eval | |
| | | <i>soft-dice</i> | DSC | SDC | DSC | SDC |
| 3DUnet | 16.32 | 83.70 | 86.38 | 78.46 | 83.70 | 77.2 |
| TransBTS | 32.99 | 84.80 | 86.27 | 80.30 | 84.18 | 79.85 |
| nnUnet | 31.19 | 88.12 | 88.53 | 83.56 | 87.33 | 83.21 |
| SwinUNETR | 62.19 | 85.00 | 88.91 | 79.96 | <u>87.39</u> | 79.62 |
| TransUnet | 31.20 | 87.47 | 88.42 | 83.39 | 85.97 | 81.84 |
| MedNeXt_M | 17.55 | 89.09 | 87.09 | 82.54 | 85.63 | 69.37 |
| SegMamba | 67.42 | 87.84 | 88.55 | 82.84 | 86.36 | 81.31 |
| LightM-Unet | <u>5.02</u> | 85.87 | 86.51 | 77.19 | 84.59 | 75.88 |
| Ours | 2.97 | 86.87 | 88.85 | 84.05 | 87.45 | <u>82.99</u> |

mentation performance. Due to the lightweight design of our RMD encoder and LRS decoder, the forward propagation process requires only 0.094 Tflops, and the entire network operates at an ultra-lightweight level of 0.282 Tflops. This represents the 45.97% reduction in computational cost compared to LightM-Unet for the same task. Additionally, memory access cost is reduced by 45.34%, and the parameter is lowered by 40.83%, establishing SuperLightNet as a leading solution for lightweight medical brain tumor segmentation tasks.

4.5. Ablation Study

To achieve a balance between computational performance and efficiency, we systematically explore and optimize the multimodal brain tumor segmentation network structure to enhance segmentation accuracy while optimizing computational efficiency. The ablation experiments are divided into two stages: the first stage evaluates the performance of mainstream methods on the task in this study, and the second stage iteratively optimizes the encoder-decoder, with each stage progressively assessing the performance-effectiveness of the network structure.

In the first stage, we conduct experiments on mainstream encoder-decoder configuration combinations to establish a performance benchmark in Table 4. We integrate advanced models such as MedNeXt, KAN [18], and Mamba, and explore single- and multi-modal parallel structure variants of the KAN framework to assess its effectiveness in reducing parameter count. Results show that using the MedNeXt+KAN encoder achieves a high Dice score of 77.36%, though using both encoder and decoder consume a large amount of memory. Furthermore, Pe3 does not provide higher accuracy than Pe2, possibly due to the difficulty of learning in deeper networks. The Mamba+KAN combination scores 70.07%, indicating it could not effectively replace transformer and MLP structures and is highly dependent

on floating-point precision. ("Pe" refers to the repetition of the encoder block structure, and "Pd" refers to the repetition of the decoder structure.)

We use KAN to explore a similar 4-Parallel structure, which significantly reduces computational resources by splitting channels, thus further releasing memory. Using KAN network structure still achieves a DSC of 75.64%. The data also shows that the MedNeXt structure is not suitable for modal channel splitting after employing depthwise convolution. Although the parallel structure demonstrates excellent memory efficiency, its low performance led us to abandon this approach.

Additionally, we test the performance of the Tri-view Hadamard Product Attention (THPA) encoder and the UNETR decoder on both Parallel and U structures. It is observed that in the U-Pe3-Pd3 THPA+UNETR experiment, a performance of 78.84% surpasses the second-best U-Pe2-Pd2 MedNeXt+KAN MedNeXt structure in Table 4, demonstrating that the multi-view Hadamard product is highly effective in feature extraction for voxel data.

In Table 5, we present the network encoder-decoder improvement plans based on the results from Table 4. We use the U structure and the widely-used UNETR decoder for such tasks. The encoder uses the best-performing THPA from Table 4, and a random path-guided strategy is employed to reduce two-thirds of the intra-block computation. Subsequently, inspired by the bidirectional scanning to THPA, increasing the random routing from the original 3 paths to 6 paths, improving it to THPAFR without increasing the parameter count, though the performance slightly declined. To address this, we compress the routing back to 3 paths and assigned corresponding reverse scanning for each path, resulting in THPAFR2, achieving a DSC of 79.53% and an SDC of 69.32%, with a notable improvement in SDC.

The aim of this study is lightweight optimization, and the reverse scanning significantly increases the computational cost. In response, We maintain the original network structure of THPAFR2 while examining the performance saturation achieved with a 384-dimensional hidden layer compared to the 768-dimensional layer. The results show that within the given epochs, DSC improved by 0.23%, while SDC decreased by 0.09%. This modification reduces the parameter count by 46%, while retaining the network's performance. Next, thanks to the very effective residual structure, we design THPAFR3 based on it, using a drop residual (dropRes) mode residual structure to retain half of the channel data and apply feature extraction to the remaining portion, thereby extending the continuity of contextual information. This change increases the DSC to 79.99%.

At this point, the network's parameters are mainly concentrated in the decoder, and UNETR is mostly used for transformer-based decoders. It is effective but parameter-

Table 3. Quantitative evaluations with $4 * 128^3$ In Res, include training memory(TM), inference memory(IM), training time(TT), Tflops for forward propagation, combined forward and backward propagation Tflops, memory access cost per second(MACs), and parameter count.

| Model | Tflops | | MACs(G) | Params(M) | In Res | TM(M) | IM(M) | TT(ms) | DSC |
|----------------------------|--------------|--------------|-------------|-------------|-------------|--------------|-------------|------------|--------------|
| | fwd | fwd+bwd | | | | | | | |
| 3DUnet(MICCAI,2016) | 3.80 | 11.40 | 1899.4 | 16.32 | $4 * 128^3$ | 28172 | 7013 | 448 | 72.77 |
| TransBTS(MICCAI,2021) | 0.51 | 1.54 | 256.7 | 32.99 | $4 * 128^3$ | 22216 | 2799 | 386 | 80.76 |
| nnUnet(Nature Method,2021) | 0.97 | 2.91 | 482.9 | 31.19 | $4 * 128^3$ | <u>13973</u> | <u>3227</u> | <u>232</u> | 81.41 |
| SwinUNETR(MICCAI,2021) | 1.48 | 4.45 | 743.0 | 62.19 | $4 * 128^3$ | 35482 | 6621 | 667 | 81.59 |
| MedNeXt_M(MICCAI,2023) | 0.50 | 1.49 | 239.7 | 17.55 | $4 * 128^3$ | 28037 | 5433 | 887 | 81.74 |
| TransUnet(MIA,2024) | 0.97 | 2.90 | 482.9 | 31.20 | $4 * 128^3$ | 12944 | 3229 | 248 | 78.86 |
| SegMamba(MICCAI,2024) | 2.92 | 8.76 | 1458.0 | 67.42 | $4 * 128^3$ | 37140 | 5575 | 900 | <u>82.04</u> |
| LightM-Unet(Axriv,2024) | <u>0.174</u> | <u>0.522</u> | <u>83.8</u> | <u>5.02</u> | $4 * 128^3$ | 38871 | 6159 | 1350 | 81.49 |
| Ours | 0.094 | 0.282 | 45.8 | 2.97 | $4 * 128^3$ | 20108 | 3485 | 219 | 82.3 |

Table 4. Analysis of networks, showing various structural performance results for this task. “OOM!” indicates out of memory, “*” denotes validation at 1000 epochs, “4I” means parallel structure

| Structure | Encoder | Decoder | DSC | SDC |
|------------|------------------|------------------|---------------|---------------|
| U-Pe2-Pd2 | MedNeXt+KAN | MedNeXt | 77.09 | 64.12 |
| U-Pe2-Pd2 | MedNeXt+KAN | MedNeXt | <u>77.36*</u> | <u>64.78*</u> |
| U-Pe3-Pd2 | MedNeXt+KAN | MedNeXt | 76.41 | 62.63 |
| U-Pe3-Pd2 | MedNeXt+KAN | MedNeXt+KAN | OOM! | OOM! |
| U-Pe2-Pd2 | Mamba+KAN | Mamba+KAN | 70.07 | 53.22 |
| U-Pe2-Pd2 | KAN | KAN | 77.30 | 63.78 |
| U-Pe2-Pd2 | KAN+SGD | KAN+SGD | 73.93 | 57.53 |
| 4I-Pe4-Pd4 | KAN Parallel | KAN Parallel | 75.64 | 62.48 |
| 4I-Pe4-Pd4 | no KAN Parallel | no KAN Parallel | 74.89 | 60.19 |
| 4I-Pe4-Pd4 | MedNeXt Parallel | MedNeXt Parallel | 69.95 | 48.77 |
| 4I-Pe3-Pd3 | THPA | UNETR | 76.53 | 64.23 |
| U-Pe3-Pd3 | THPA | UNETR | 78.84 | 68.48 |

Table 5. Ablation study of network configurations. **Bold** indicates optimal data, underline denotes the second-best data.

| Encoder | Decoder | Params | DSC | SDC | Droprr | Dimension |
|---------|---------|--------------|--------------|--------------|--------|-----------|
| THPA | UNETR | 15.8M | 78.84 | 68.48 | 0 | 768 |
| THPAFR | UNETR | 15.8M | 77.58 | 66.43 | 0 | 768 |
| THPAFR2 | UNETR | 15.8M | 76.83 | 65.75 | 0 | 768 |
| THPAFR2 | UNETR | <u>8.6M</u> | <u>78.63</u> | 67.75 | 0 | 384 |
| THPAFR3 | UNETR | 7.56M | 79.02 | <u>68.15</u> | 1 | 384 |
| THPAFR3 | LRSUp | 2.86M | 78.39 | 65.95 | 1 | 384 |
| THPAFR3 | LRSUp2 | 4.19M | 78.68 | 66.12 | 1 | 384 |
| THPAFR3 | LRSUp3 | 3.96M | <u>78.74</u> | <u>67.87</u> | 1 | 384 |
| THPAFR3 | LRSUp4 | <u>2.97M</u> | 79.05 | 68.20 | 1 | 384 |

heavy. To address this, we redesign the LRSUp upsampling structure. First, we use linear upsampling and grouped MLP to learn features while employing weighted skip and weighted residual structures to retain contextual information. We achieve a network with 2.86M parameters while maintaining segmentation performance. LRSUp2 uses a deconvolution mode to replace linear upsampling, and LRSUp3 uses the THPAFR3 encoding blocks as decoding

blocks. LRSUp4 replaces grouped weighting and fully weighted residuals and skips, and also replaces the channel-matching grouped MLP in LRSUp with a grouped MLP of channel count $C/12$, aiming to retain more inter-slice domain relations while keeping it lightweight.

LRSUp4 demonstrates clear advantages over LRSUp in rapid learning, achieving a 0.66% improvement in DSC and a 2.25% improvement in SDC at epoch 100. The significant increase in SDC is highly correlated with the $C/12$ grouped MLP improvement. Compared with THPAFR3+UNETR, LRSUp4 achieves similar performance with a 60.71% reduction in parameter count. RMD encoder refers to the THPAFR3 encoder, and LRS decoder refers to the LRSUp4 decoder.

5. Conclusion

This paper proposes a lightweight parameter aggregation network for multimodal brain tumor segmentation, which designs a random multiview drop encoder and a learnable residual skip decoder. The random multiview drop encoder is designed to learn the spatial structure of multimodal images through a random multi-view approach. The learnable residual skip decoder is designed to incorporate learnable residual and group skip weights. Experimental results demonstrate that the proposed method achieves a leading reduction in parameter count by 95.59%, a 96.78% improvement in computational efficiency, a 96.86% enhancement in memory access performance, and an average performance gain of 0.21% on the BraTS2019 and BraTS2021 datasets in comparison with the state-of-the-art methods. The proposed lightweight segmentation method with low computation can assist doctors in better and efficient planning surgeries and calculating radiation therapy doses. In the future work, the lightweight network with more accuracy should be further studied.

Acknowledgment

This work was supported by national natural science foundation of China (No.62202346), China scholarship council (No.202208420109), Wuhan applied basic frontier research project (No.2022013988065212), 2024 Wuhan textile university special fund project (No.2024442), and central government-guided local science and technology development special fund of Hubei province (No.2024EIA003).

References

- [1] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13, 2017. 5
- [2] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, RT Shinohara, Christoph Berger, SM Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation. *progression assessment, and overall survival prediction in the BRATS challenge*, 10, 2018. 5
- [3] MK Balwant. A review on convolutional neural networks for brain tumor segmentation: methods, datasets, libraries, and future directions. *Irbm*, 43(6):521–537, 2022. 5
- [4] 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 *European conference on computer vision*, pages 205–218. Springer, 2022. 2
- [5] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170: 446–455, 2018. 3
- [6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2
- [7] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2
- [8] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021. 2
- [9] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 2, 4
- [10] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–426. Springer, 2023. 3
- [11] Zhenqi He, Mathias Unberath, Jing Ke, and Yiqing Shen. Transnuseg: A lightweight multi-task transformer for nuclei segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 206–215. Springer, 2023. 3
- [12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3
- [13] Nabil Ibtehaz and Daisuke Kihara. Acc-unet: A completely convolutional unet model for the 2020s. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 692–702. Springer, 2023. 3
- [14] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 1, 2
- [15] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2*, pages 138–149. Springer, 2016. 1
- [16] Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M Jorge Cardoso, and Tom Vercauteren. On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25–30, 2017, Proceedings 25*, pages 348–360. Springer, 2017. 1
- [17] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Cehngwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*, 2024. 3
- [18] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks, 2024. 7
- [19] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2
- [20] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 5
- [21] 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)*, pages 565–571. Ieee, 2016.

[1](#)

- [22] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016. [2](#)
- [23] 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. [1](#)
- [24] Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024. [3](#)
- [25] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 481–490. Springer, 2023. [2](#), [3](#)
- [26] Na Ta, Haipeng Chen, Xianzhu Liu, and Nuo Jin. Let-net: locally enhanced transformer network for medical image segmentation. *Multimedia Systems*, 29(6):3847–3861, 2023. [3](#)
- [27] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. [1](#)
- [28] Yiqing Wang, Zihan Li, Jieru Mei, Zihao Wei, Li Liu, Chen Wang, Shengtian Sang, Alan L Yuille, Cihang Xie, and Yuyin Zhou. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 486–496. Springer, 2023. [2](#)
- [29] Wang Wenxuan, Chen Chen, Ding Meng, Yu Hong, Zha Sen, and Li Jianguyun. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pages 109–119, 2021. [2](#)
- [30] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6030–6038, 2024. [5](#)
- [31] xiaoju ye. calflops: a flops and params calculate tool for neural networks in pytorch framework, 2023. [5](#)
- [32] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. [3](#)
- [33] 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. [1](#)