DBTrans: A Dual-Branch Vision Transformer for Multi-Modal Brain Tumor Segmentation

Xinyi Zeng, Pinxian Zeng, Cheng Tang, Peng Wang, Binyu Yan, and Yan Wang(✉)

School of Computer Science, Sichuan University, Chengdu, China
wangyanscu@hotmail.com

Abstract. 3D Spatially Aligned Multi-modal MRI Brain Tumor Segmentation (SAMM-BTS) is a crucial task for clinical diagnosis. While Transformer-based models have shown outstanding success in this field due to their ability to model global features using the self-attention mechanism, they still face two challenges. First, due to the high computational complexity and deficiencies in modeling local features, the traditional self-attention mechanism is ill-suited for SAMM-BTS tasks that require modeling both global and local volumetric features within an acceptable computation overhead. Second, existing models only stack spatially aligned multi-modal data on the channel dimension, without any processing for such multi-channel data in the model's internal design. To address these challenges, we propose a Transformer-based model for the SAMM-BTS task, namely DBTrans, with dual-branch architectures for both the encoder and decoder. Specifically, the encoder implements two parallel feature extraction branches, including a local branch based on Shifted Window Self-attention and a global branch based on Shuffle Window Cross-attention to capture both local and global information with linear computational complexity. Besides, we add an extra global branch based on Shifted Window Cross-attention to the decoder, introducing the key and value matrices from the corresponding encoder block, allowing the segmented target to access a more complete context during up-sampling. Furthermore, the above dual-branch designs in the encoder and decoder are both integrated with improved channel attention mechanisms to fully explore the contribution of features at different channels. Experimental results demonstrate the superiority of our DBTrans model in both qualitative and quantitative measures. Codes will be released at https://github.com/Aru321/DBTrans.

Keywords: Spatially aligned multi-modal MRI (SAMM) · Brain tumor segmentation (BTS) · Transformer · Cross-Attention · Channel-Attention

1 Introduction

Glioma is one of the most common malignant brain tumors with varying degrees of invasiveness [1]. Brain Tumor Semantic segmentation of gliomas based on 3D spatially aligned Magnetic Resonance Imaging (SAMM-BTS) is crucial for accurate diagnosis

X. Zeng and P. Zeng—Contribute equally to this work.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 502–512, 2023.
https://doi.org/10.1007/978-3-031-43901-8_48

and treatment planning. Unfortunately, radiologists suffer from spending several hours manually performing the SAMM-BTS task in clinical practice, resulting in low diagnostic efficiency. In addition, manual delineation requires doctors to have high professionalism. Therefore, it is necessary to design an efficient and accurate glioma lesion segmentation algorithm to effectively alleviate this problem and relieve doctors' workload and improve radiotherapy quality.

With the rise of deep learning, researchers have begun to study deep learning-based image analysis methods [2, 37]. Specifically, many convolutional neural network-based (CNN-based) models have achieved promising results [3–8]. Compared with natural images, medical image segmentation often requires higher accuracy to make subsequent treatment plans for patients. U-Net reaches an outstanding performance on medical image segmentation by combining the features from shallow and deep layers using skip-connection [9–11]. Based on U-Net, Brugger. *et al.* [12] proposed a partially reversible U-Net to reduce memory consumption while maintaining acceptable segmentation results. Pei *et al.* [13] explored the efficiency of residual learning and designed a 3D ResUNet for multi-modal brain tumor segmentation. However, due to the lack of global understanding of images for convolution operation, CNN-based methods struggle to model the dependencies between distant features and make full use of the contextual information [14]. But for semantic segmentation tasks whose results need to be predicted at pixel-level or voxel-level, both local spatial details and global dependencies are extremely important.

In recent years, models based on the self-attention mechanism, such as Transformer, have received widespread attention due to their excellent performance in Natural Language Processing (NLP) [15]. Compared with convolution operation, the self-attention mechanism is not restricted by local receptive fields and can capture long-range dependencies. Many works [16–19] have applied Transformers to computer vision tasks and achieved favorable results. For classification tasks, Vision Transformer (ViT) [19] was a groundbreaking innovation that first introduced pure Transformer layers directly across domains. And for semantic segmentation tasks, many methods, such as SETR [20] and Segformer [21], use ViT as the direct backbone network and combine it with a task-specific segmentation head for prediction results, reaching excellent performance on some 2D natural image datasets. For 3D medical image segmentation, Vision Transformer has also been preferred by researchers. A lot of robust variants based on Transformer have been designed to endow U-Net with the ability to capture contextual information in long-distance dependencies, further improving the semantic segmentation results of medical images [22–27]. Wang *et al.* [25] proposed a novel framework named TransBTS that embeds the Transformer in the bottleneck part of a 3D U-Net structure. Peiris *et al.* [26] introduced a 3D Swin-Transformer [28] to segmentation tasks and first incorporated the attention mechanism into skip-connection.

While Transformer-based models have shown effectiveness in capturing long-range dependencies, designing a Transformer architecture that performs well on the SAMM-BTS task remains challenging. First, modeling relationships between 3D voxel sequences is much more difficult than 2D pixel sequences. When applying 2D models, 3D images need to be sliced along one dimension. However, the data in each slice is related to three views, discarding any of them may lead to the loss of local information, which may cause the degradation of performance [29]. Second, most existing MRI segmentation methods still have difficulty capturing global interaction information while effectively encoding local information. Moreover, current methods just stack modalities and pass them through a network, which treats each modality equally along the channel dimension and may ignore the contribution of different modalities. To address the above limitations, we propose a novel encoder-decoder model, namely DBTrans, for multi-modal medical image segmentation. In the encoder, two types of window-based attention mechanisms, i.e., Shifted Window-based Multi-head Self Attention (Shifted-W-MSA) and Shuffle Window-based Multi-head Cross Attention (Shuffle-W-MCA), are introduced and applied in parallel to dual-branch encoder layers, while in the decoder, in addition to Shifted-W-MSA mechanism, Shifted Window-based Multi-head Cross Attention (Shifted-W-MCA) is designed for the dual-branch decoder layers. These mechanisms in the dual-branch architecture greatly enhance the ability of both local and global feature extraction. Notably, DBTrans is designed for 3D medical images, avoiding the information loss caused by data slicing.

The contributions of our proposed method can be described as follows: 1) Based on Transformer, we construct dual-branch encoder and decoder layers that assemble two attention mechanisms, being able to model close-window and distant-window dependencies without any extra computational cost. 2) In addition to the traditional skip-connection structure, in the dual-branch decoder, we also establish an extra path to facilitate the decoding process. We design a Shifted-W-MCA-based global branch to build a bridge between the decoder and encoder, maintaining affluent information of the segmentation target during the decoding process. 3) For the multi-modal data adopt in the task of SAMM-BTS, we improve the channel attention mechanism in SE-Net by applying SE-weights to features from both branches in the encoder and decoder layers. By this means, we implicitly consider the importance of multiple MRI modalities and two window-based attention branches, thereby strengthening the fusion effect of the multi-modal information from a global perspective.

2 Methodology

Figure 1 shows the overall structure of the proposed DBTrans. It is an end-to-end framework that has a 3D patch embedding along with a U-shaped model containing an encoder and a decoder. The model takes MRI data of $D \times H \times W \times C$ with four modalities stacked along channel dimensions as the input. The 3D patch embedding converts the input data to feature embedding $e^1 \in R^{D_1 \times H_1 \times W_1 \times C_1}$ which will be further processed by encoder layers. At the tail of the decoder, the segmentation head takes the output of the last layer and generates the final segmentation result of $D \times H \times W \times K$.

2.1 Dual-Branch in Encoder

As shown in Fig. 1(b), the encoder consists of four dual-branch encoder layers (one bottleneck included). Each encoder layer contains two consecutive encoder blocks and a 3D patch merging to down-sample the feature embedding. Note that there is only one encoder block in the bottleneck. The encoder block includes a dual-branch architecture with the local feature extraction branch, the global feature extraction branch, and

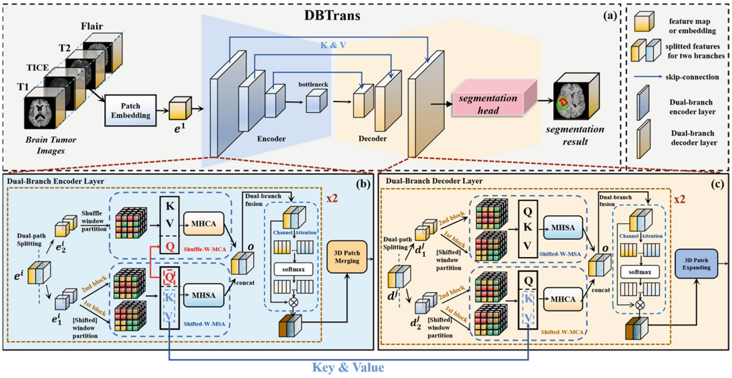

Fig. 1. The overall framework of the proposed DBTrans. The encoder contains dual-branch encoder layers (including a bottleneck), and the decoder contains dual-branch encoder layers. Skip connections based on cross attention are built between encoder and decoder.

the channel-attention-based dual-branch fusion module. After acquiring the embedding $e^i \in R^{D_i \times H_i \times W_i \times C_i}$ in the i-th encoder layer ($i \in [1, 4]$), we split it along the channel dimension, obtaining $e^i_1, e^i_2 \in R^{D_i \times H_i \times W_i \times [C_i/2]}$ which are then separately fed into two branches.

Shifted-W-MSA-Based Local Branch. The image embedding $e^i_1 \in R^{D_i \times H_i \times W_i \times [C_i/2]}$ is fed into the local branch in the encoder block. In the process of window partition (denoted as WP), e^i_1 is split into non-overlapping windows after a layer normalization (LN) operation to obtain the window matrix m^i_1. Since we set M as 2, the input of $4 \times 4 \times 4$ size is uniformly divided into 8 windows of $2 \times 2 \times 2$. Following 3D Swin-Transformer [27], we introduce MSA based on the shifted-window partition to the second block in an encoder layer. During window partition of Shifted-W-MSA, the whole feature map is shifted by half of the window size, i.e., $\left(\frac{\mathrm{M}}{2}, \frac{\mathrm{M}}{2}, \frac{\mathrm{M}}{2}\right)$. After the window partition, W-MSA is applied to calculate multi-head self-attention within each window. Specifically, for window matrix m^i_1, we first apply projection matrices W^i_Q,W^i_K, and W^i_V to obtain Q^i_1,K^i_1, and V^i_1 matrices (the process is denoted as $Proj$). After projection, multi-head self-attention calculation is performed on Q^i_1,K^i_1, and V^i_1 to get the attention score of every window. Finally, we rebuild the feature map from the windows, which serves as the inverse process of the window partition. After calculating the attention score, other basic components in Transformer are also employed, that is, layer normalization (LN), as well as a multi-layer perceptron (MLP) with two fully connected layers and a Gaussian Error Linear Unit (GELU). Residual connection is applied after each module. The whole process can be expressed as follows:

$$\begin{gathered}
m^i_1 = \left[\boldsymbol{Shifted}\text{-}\right] WP\left(LN\left(e^i_1\right)\right), \\
Q^i_1, K^i_1, V^i_1 \in R^{\frac{D_i H_i W_i}{M^3} \times M^3 \times [C_i/2]} = Proj^i\left(m^i_1\right) = W^i_Q \cdot m^i_1, W^i_K \cdot m^i_1, W^i_V \cdot m^i_1, \\
\hat{z}^i_1 = W\text{-}MSA\left(Q^i_1, K^i_1, V^i_1\right), \\
o^i_1 = MLP^i\left(LN\left(\left(\hat{z}^i_1 + e^i_1\right)\right)\right) + \left(\hat{z}^i_1 + e^i_1\right),
\end{gathered} \tag{1}$$

where $\hat{z}_1^i$ represents the attention score after $W\text{-}MSA$ calculation, "$\left[\boldsymbol{Shifted}\text{-}\right]$" represents that we use the shifted-window partition and restoration in the second block of every encoder layer. o_1^i is the final output of the local branch.

Shuffle-W-MCA-Based Global Branch. Through the local branch, the network still cannot model the long-distance dependencies between non-adjacent windows in the same layer. Thus, Shuffle-W-MCA is designed to complete the complementary task. After the window-partition process that converts the embedding $e_2^i \in R^{D_i \times H_i \times W_i \times [C_i/2]}$ to $m_2^i \in R^{\frac{D_i H_i W_i}{M^3} \times M^3 \times [C_i/2]}$, inspired by ShuffleNet [35], instead of moving the channels, we propose to conduct shuffle operations on the patches in different windows. The patches at the same relative position in different windows are rearranged together in a window, and their position is decided by the position of the window they originally belong to. Then, this branch takes the query from the local branch, while generating keys and values from m_2^i to compute cross-attention scores. Such a design aims to enable the network to model the relationship between a window and other distant windows. Note that we adopt the projection function $Proj$ from the local branch, indicating that the weights of the two branches are shared. Through the shuffle operation, the network can generate both local and global feature representations without setting additional weight parameters. The whole process can be formulated as follows:

$$\begin{aligned} Q_2^i, K_2^i, V_2^i \in R^{\frac{D_i H_i W_i}{M^3} \times M^3 \times [C_i/2]} &= Proj^i\left(Shuffle(m_2^i)\right), \\ \hat{z}_2^i &= W\text{-}MCA\left(Q = Q_1^i, K = K_2^i, V = V_2^i\right), \\ o_2^i &= MLP^i\left(LN\left(\hat{z}_2^i + e_2^i\right)\right) + \left(\hat{z}_1^i + e_1^i\right). \end{aligned} \tag{2}$$

In the second block, we get the final output o_2^i of the layer through the same process.

2.2 Dual-Branch in Decoder

As shown in Fig. 1(c), the decoder takes the output of the encoder as the input and generates the final prediction through three dual-branch decoder layers as well as a segmentation head. Each decoder layer consists of two consecutive decoder blocks and a 3D patch expanding layer. As for the j-th decoder layer ($j \in [1, 3]$), the embedding $d^j \in R^{D_j \times H_j \times W_j \times C_j}$ is also divided into the feature maps $d_1^j, d_2^j \in R^{D_j \times H_j \times W_j \times [C_j/2]}$. We further process d_1^j, d_2^j using a dual-branch architecture similar to that of the encoder, but with an additional global branch based on Shifted-W-MCA mechanism. Finally, the segmentation head generates the final segmentation result of $D \times H \times W \times K$, where K represents the number of classes. The local branch based on Shifted-W-MSA is the same as that in the encoder and will not be introduced in this section.

Shifted-W-MCA-Based Global Branch. Apart from employing Shifted-W-MSA to form the local branch of the decoder layer, we design a novel Shifted-W-MCA mechanism for the global branch to ease the information loss during the decoding process and take full advantage of the features from the encoder layers. The global branch receives the query matrix from the split feature map $d_2^j \in R^{D_j \times H_j \times W_j \times [C_j/2]}$, while receiving key and value matrices from the encoder block in the corresponding stage, denoted as

Q_2^j, K_{e_i}, and V_{e_i}. The process of Shifted-W-MCA can be formulated as follows:

$$\begin{gathered} Q_2^j, K_2^j, V_2^j = Proj^j\left(\left[\boldsymbol{Shifted}\text{-}\right]WP\left(LN\left(d_2^i\right)\right)\right), \\ \hat{z}_2^j = W\text{-}MCA\left(Q = Q_2^j, K = K_{e_1^{4-j}}, V = V_{e_1^{4-j}}\right), \\ o_2^j = MLP^i\left(LN\left(\hat{z}_2^j + d_2^j\right)\right) + \left(\hat{z}_1^j + d_1^j\right), \end{gathered} \tag{3}$$

where $\hat{z}_2^j$ denotes the attention score after MCA calculation, o_2^j denotes the final output of the global branch.

2.3 Channel-Attention-Based Dual-Branch Fusion

As shown in Fig. 1(b) and Fig. 1(c), the dual-branch fusion module is based on the channel attention. For the block of the m -th ($m \in [1, 3]$) encoder or decoder layer, the dual-branch fusion module combines the features $o_1^m, o_2^m \in R^{D_m \times H_m \times W_m \times [C_m/2]}$ from the two extraction branches, obtaining a feature map filled with abundant multi-scale information among different modalities. Subsequently, the dependencies between the feature channels within the individual branches are implicitly modeled with the SE-Weight assignment first proposed in SE-Net [30]. Different from SE-Net, we dynamically assign weights for both dual-branch fusion and multi-modal fusion. The process of obtaining the attention weights can be represented as the formula (5) below:

$$Z_p = SE\_Weight\left(o_p^m\right), p = 1, 2, \tag{4}$$

where $Z_p \in R^{[C/2] \times 1 \times 1 \times 1}$ is the attention weight of a single branch. Then, the weight vectors of the two branches are re-calibrated using a Softmax function. Finally, the weighted channel attention is multiplied with the corresponding scale feature map to obtain the refined output feature map with richer multi-scale feature information:

$$\begin{gathered} attn_p = Softmax\left(Z_p\right) = \frac{\exp\left(Z_p\right)}{\sum_{p=1}^{2} \exp\left(Z_p\right)}, \\ Y_p = o_p^m \odot attn_p, p = 1, 2, \\ O = Cat([Y_1, Y_2]), \end{gathered} \tag{5}$$

where "$\odot$" represents the operation of element-wise multiplication and "*Cat*" represents the concatenation. The concatenated output O, serving as the dual-branch output of a block in the encoder or decoder, implicitly integrates attention interaction information within individual branches across different channels/modalities, as well as across different branches.

2.4 Training Details

For the loss function, the widely used cross entropy (CE) loss and Dice loss [32] are introduced to train our DBTrans. Here we use parameter γ to balance the two loss parts. Our network is implemented based on PyTorch, and trained for 300 epochs using a single

RTX 3090 with 24G memory. The weights of the network were updated using the Adam optimizer, the batch size was set to 1, and the initial learning rate was set to 1×10^{-4}. A cosine decay scheduler was used as the adjustment strategy for the learning rate during training. We set the embedding dimensions C_0 as 144. Following previous segmentation methods, the parameter γ is set to 0.5.

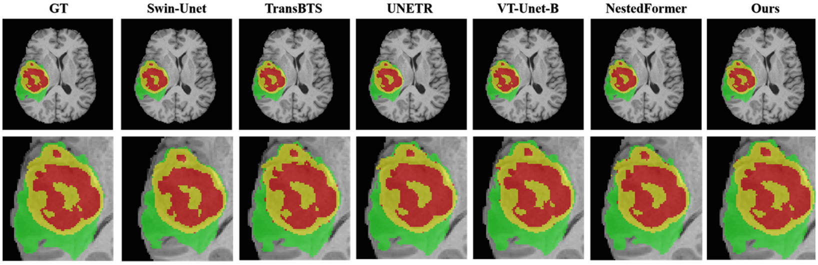

Fig. 2. Qualitative segmentation results on the test samples. The bottom row zooms-in the segmentation regions. Green, yellow, and red represent the peritumoral edema (ED), enhancing tumor (ET) and non-enhancing tumor/necrotic tumor (NET/NCR).

3 Experiments and Results

Datasets. We use the Multimodal Brain Tumor Segmentation Challenge (BraTS 2021 [33, 34, 38]) as the benchmark training set, validation set, and testing set. We divide the 1251 scans provided into 834, 208, and 209 (in a ratio of 2:1:1), respectively for training and testing. The ground truth labels of GBM segmentation necrotic/active tumor and edema are used to train the model. The BraTS 2021 dataset reflects real clinical diagnostic species and has four spatially aligned MRI modality data, namely T1, T1CE, T2, and Flair, which are obtained from different devices or according to different imaging protocols. The dataset contains three distinct sub-regions of brain tumors, namely peritumoral edema, enhancing tumor, and tumor core. The data augmentation includes random flipping, intensity scaling and intensity shifting on each axis with probabilities set to 0.5, 0.1 and 0.1, respectively.

Comparative Experiments. To evaluate the effectiveness of the proposed DBTrans, we compare it with the state-of-the-art brain tumor segmentation methods including six Transformer-based networks Swin-Unet [27], TransBTS [25], UNETR [22], nnFormer [23], VT-Unet-B [26], NestedFormer [36] as well as the most basic CNN network 3D U-Net [31] as the baseline. During inference, for any size of 3D images, we utilize the overlapping sliding windows technique to generate multi-class prediction results and take average values for the voxels in the overlapping region. The evaluation strategy adopted in this work is consistent with that of VT-Unet [26]. For other methods, we used the corresponding hyperparameter configuration mentioned in the original papers and reported the average metrics over 3 runs. Table 1 presents the Dice scores and 95% HDs of different methods for segmentation results on three different tumor regions (i.e.,

Table 1. Quantitative comparison with other state-of-the-arts methods in terms of dice score and 95% Hausdorff distance. Best results are bold, and second best are underlined.

Method	#param	FLOPS	Dice Score				95% Hausdorff Distance			
			ET	TC	WT	AVG	ET	TC	WT	AVG
3D U-Net[31]	11.9M	557.9G	83.39	86.28	89.59	86.42	6.15	**6.18**	11.49	7.94
Swin-Unet[27]	52.5M	93.17G	83.34	87.62	89.81	89.61	6.19	6.35	11.53	8.03
TransBTS[25]	33M	333G	80.35	85.35	89.25	84.99	7.83	8.21	15.12	10.41
UNETR[22]	102.5M	193.5G	79.78	83.66	90.10	84.51	9.72	10.01	15.99	11.90
nnFormer[23]	39.7M	110.7G	82.83	86.48	90.37	86.56	8.00	7.89	11.66	9.18
VT-Unet-B[26]	20.8M	165.0G	85.59	87.41	91.02	88.07	6.23	6.29	10.03	7.52
NestedFormer[36]	**10.48M**	**71.77G**	85.62	88.18	90.12	87.88	**6.08**	6.43	10.23	7.63
DBTrans(Ours)	24.6M	146.2G	**86.70**	**90.26**	**92.41**	**89.69**	6.13	6.24	**9.84**	**7.38**

ET, TC and WT) respectively, where a higher Dice score indicates better results and a lower 95% HD indicates better performance. By observation, our approach achieved the best performance on average among these methods in all three tumor regions. Compared with the state-of-the-art method VT-Unet, our method increased the average Dice score by 1.62 percentage points and achieved the lowest average 95% HD. Moreover, to verify the significance of our improvements, we calculate the variances of all results and conduct statistical tests (i.e., paired t-test). The results show that p-values on Dice and 95%HD are less than 0.05 in most comparison cases, indicating that the improvements are statistically significant. Compared with other methods, our DBTrans can not only capture more precise long-term dependencies between non-adjacent slices through Shuffle-W-MCA, but also accomplish dual-branch fusion and multi-modal fusion through dynamic SE-Weight assignment, obtaining better feature representations for segmentation.

Table 2. Quantitative comparison of ablation models in terms of average dice score.

Name	Index	DB-E	DB-D	Dual-branch fusion	Avg-Dice	Param
SwinUnet-1	(1)	×	×	×	86.73	52.5M
SwinUnet-2	(2)	√	×	×	87.52	43.2M
SwinUnet-3	(3)	×	√	×	88.26	46.1M
SwinUnet-4	(4)	√	√	×	88.86	27.7M
proposed	**(5)**	√	√	√	**89.69**	**24.6M**

Figure 2 also shows the qualitative segmentation results on the test samples of test set patients, which further proves the feasibility and superiority of our DBTrans model. From the zoom-in area, we can observe that our model can more accurately segment tumor structures and delineate tumor boundaries compared with other methods.

Ablation Study. To further verify the contribution of each module, we establish the ablation models based on the modules introduced above. Note that, DP-E represents the dual-branch encoder layer, while DB-D represents the dual-branch decoder layer. When the dual-branch fusion is not included, we do not split the input, and simply fuse the features from the two branches using a convolution layer. In all, there are 5 models included in this ablation study: (1) SwinUnet-1 (baseline): We use Swin-Transformer layers without any dual-branch module. (2) SwinUnet-2: Based on (1), we add dual-branch encoder layers to the model. (3) SwinUnet-3: Based on (1), we add dual-branch decoder layers to the model. (4) SwinUnet-4: Based on (1), add both the encoder and decoder without the dual-branch fusion module. (5) Our proposed DBTrans model. As shown in Table 2, After applying our proposed dual-branch encoder and decoder layers to the baseline model, the average Dice score notably increased by 2.13. Subsequently, applying the dual-branch fusion module also prominently contributes to the performance of the model by an improvement of 0.83 on the Dice score. Notably, our dual-branch designs achieve higher performance while also reducing the number of parameters required. This is because we split the original feature embedding into two parts, thus the channel dimensions of features in two branches are halved.

4 Conclusion

In this paper, we innovatively proposed an end-to-end model named DBTrans for multi-modal medical image segmentation. In DBTrans, first, we well designed the dual-branch structures in encoder and decoder layers with Shifted-W-MSA, Shuffle-W-MCA, and Shifted-W-MCA mechanisms, facilitating feature extraction from both local and global views. Moreover, in the decoder, we establish a bridge between the query of the decoder and the key/value of the encoder to maintain the global context during the decoding process for the segmentation target. Finally, for the multi-modal superimposed data, we modify the channel attention mechanism in SE-Net, focusing on exploring the contribution of different modalities and branches to the effective information of feature maps. Experimental results demonstrate the superiority of DBTrans compared with the state-of-the-art medical image segmentation methods.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (NSFC 62371325, 62071314), Sichuan Science and Technology Program 2023YFG0263, 2023YFG0025, 2023NSFSC0497, and Opening Foundation of Agile and Intelligent Computing Key Laboratory of Sichuan Province.

References

1. Gordillo, N., Montseny, E., et al.: State of the art survey on MRI brain tumor segmentation. Magn. Reson. Imaging **31**(8), 1426–1438 (2013)
2. Luo, Y., Zhou, L., Zhan, B., et al.: Adaptive rectification based adversarial network with spectrum constraint for high-quality PET image synthesis. Med. Image Anal. **77**, 102335 (2022)

3. Wang, K., Zhan, B., Zu, C., Wu, X., et al.: Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Med. Image Anal. **79**, 102447 (2022)
4. Ma, Q., Zu, C., Wu, X., Zhou, J., Wang, Y.: Coarse-To-fine segmentation of organs at risk in nasopharyngeal carcinoma radiotherapy. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 358–368. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_34
5. Tang, P., Yang, P., Nie, D., et al.: Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowl.-Based Syst. **241**, 108215 (2022)
6. Zhang, J., Zhang, Z., Wang, L., et al.: Kernel-based feature aggregation framework in point cloud networks. Pattern Recogn. **1**(1), 1–15 (2023)
7. Shi, Y., Zu, C., Yang, P., et al.: Uncertainty-weighted and relation-driven consistency training for semi-supervised head-and-neck tumor segmentation. Knowl.-Based Syst. **272**, 110598 (2023)
8. Wang, K., Wang, Y., Zhan, B., et al.: An efficient semi-supervised framework with multi-task and curriculum learning for medical image. Int. J. Neural Syst. **32**(09), 2250043 (2022)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Zhou, T., Zhou, Y., He, K., et al.: Cross-level feature aggregation network for polyp segmentation. Pattern Recogn. **140**, 109555 (2023)
11. Du, G., Cao, X., Liang, J., et al.: Medical image segmentation based on u-net: a review. J. Imaging Sci. Technol. **64**(2), 020508-1–020508-12 (2020)
12. Brügger, R., Baumgartner, C.F., Konukoglu, E.: A partially reversible U-Net for memory-efficient volumetric image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 429–437. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_48
13. Pei, L., Liu, Y.: Multimodal brain tumor segmentation using a 3D ResUNet in BraTS 2021. In: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, Revised Selected Papers, Part I, pp. 315–323. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_26
14. Zeng, P., Wang, Y., et al.: 3D CVT-GAN: a 3D convolutional vision transformer-GAN for PET reconstruction.In: Wang, L., et al. (eds.) MICCAI 2022, Proceedings, pp. 516-526. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_49
15. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
16. Parmar, N., Vaswani, A,, Uszkoreit, J., et al.: Image transformer. In: International Conference on Machine Learning, pp. 4055–4064. PMLR (2018)
17. Luo, Y., et al.: 3D Transformer-GAN for high-quality PET reconstruction. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 276–285. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_27
18. Wen, L., Xiao, J., Tan, S., et al.: A transformer-embedded multi-task model for dose distribution prediction. Int. J. Neural Syst. **33**, 2350043–2350043 (2023)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations (2021)
20. Zheng, S., Lu, J., Zhao, H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)

21. Xie, E., Wang, W., Yu, Z., et al.: SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural. Inf. Process. Syst. **34**, 12077–12090 (2021)
22. Hatamizadeh, A., Tang, Y., Nath, V., et al.: Unetr: transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
23. Zhou, H.Y., Guo, J., Zhang, Y., et al.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
24. Lin, A., Chen, B., Xu, J., et al.: Ds-transunet: dual swin transformer u-net for medical image segmentation. IEEE Trans. Instrum. Meas. **71**, 1–15 (2022)
25. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
26. Peiris, H., Hayat, M., Chen, Z., et al.: A robust volumetric transformer for accurate 3d tumor segmentation. In: Wang, L., et al. (eds.) MICCAI 2022, Proceedings, Part V, pp. 162–172. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_16
27. Cao, H., Wang, Y., Chen, J., et al.: Swin-unet: unet-like pure transformer for medical image segmentation. In: Computer Vision–ECCV 2022 Workshops. Proceedings, Part III, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
28. Liu, Z., Ning, J., Cao, Y., et al.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202–3211 (2022)
29. Gholami, A., et al.: A novel domain adaptation framework for medical image segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 289–298. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_26
30. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
31. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
32. Jadon S.: A survey of loss functions for semantic segmentation. In: Proceedings of IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–7. IEEE (2020)
33. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021)
34. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Nat. Sci. Data **4**, 170117 (2017)
35. Zhang, X., Zhou, X., Lin, M, et al.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
36. Xing, Z., Yu, L., Wan, L., et al.: NestedFormer: nested modality-aware transformer for brain tumor segmentation. In: Wang, L., et al. (eds.) MICCAI 2022, Proceedings, Part V, pp. 140–150. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_14
37. Wang, Y., Zhou, L., Yu, B., Wang, L., et al.: 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. IEEE Trans. Med. Imaging **38**(6), 1328–1339 (2019)
38. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)