TranSiam: Aggregating multi-modal visual features with locality for medical image segmentation

Xuejian Li [a,1], Shiqiang Ma [b,1], Junhai Xu [c], Jijun Tang [b], Shengfeng He [d], Fei Guo [a,*]

a School of Computer Science and Engineering, Central South University, China

b Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China

c College of Intelligence and Computing, Tianjin University, China

d School of Computing and Information Systems, Singapore Management University, Singapore

*Corresponding author 1 Contributed equally

Published in Expert Systems with Applications, 2024, 237, 121574. DOI: 0.1016/j.eswa.2023.121574

Abstract: Automatic segmentation of medical images plays an important role in the diagnosis of diseases. On single-modal data, convolutional neural networks have demonstrated satisfactory performance. However, multi-modal data encompasses a greater amount of information rather than single-modal data. Multi-modal data can be effectively used to improve the segmentation accuracy of regions of interest by analyzing both spatial and temporal information. In this study, we propose a dual-path segmentation model for multi-modal medical images, named TranSiam. Taking into account that there is a significant diversity between the different modalities, TranSiam employs two parallel CNNs to extract the features which are specific to each of the modalities. In our method, two parallel CNNs extract detailed and local information in the low-level layer, and the Transformer layer extracts global information in the high-level layer. Finally, we fuse the features of different modalities via a locality-aware aggregation block (LAA block) to establish the association between different modal features. The LAA block is used to locate the region of interest and suppress the influence of invalid regions on multi-modal feature fusion. TranSiam uses LAA blocks at each layer of the encoder in order to fully fuse multi-modal information at different scales. Extensive experiments on several multi-modal datasets have shown that TranSiam achieves satisfying results.

Keywords: Feature-level fusion, Local attention mechanism, Medical image segmentation, Multi-modal fusion

1. Introduction

Analysis of medical images plays an important role in the diagnosis of diseases. However, manual segmentation of medical images depends on the experience of the physician. Due to the advancement of computer technology, it is now possible for computers to automatically segment medical images. Medical images can be segmented accurately to detect diseased areas, human organs, and infected areas, improving the efficiency of diagnosis, such as brain tumor segmentation (Ma, Li et al., 2021, Yang et al., 2022), Vestibular Schwannoma segmentation, retinal vessel segmentation (Ding et al., 2021, Li et al., 2022), etc. As a result, medical image segmentation has a wide range of potential applications.

With the development of medical imaging technology, multi-modal medical images have become more prevalent in recent years. It is widely known that multi-modal data contains more information than single-modal data. The issue of effectively fusing information is a hot topic in data analysis. The fusion of multi-modal information is typically performed at three levels: input-level, feature-level, and decision-level. Fig. 1 illustrates the differences between these three fusion methods.

In feature-level fusion methods (Dolz, Desrosiers et al., 2018, Sun et al., 2021), modalities are usually concatenated or added element-by-element. Dolz, Gopinath et al. (2018) and Wang, Zhang et al. (2020) fused these features by dense Concatenation. Despite the simplicity and effectiveness of this fusion method, it does not fully explore the potential correlation between different modal features. As shown in Fig. 2, we decompose the fusion process into Concatenation and Element-wise Addition. An Element-wise Addition performs addition operations directly on the features of different modalities. Despite the semantic gap between the features of different modalities, this fusion method suffers from the same problem as input-level fusion. Unlike Element-wise Addition, concatenation involves mapping of features, since it uses convolution to further extract features after

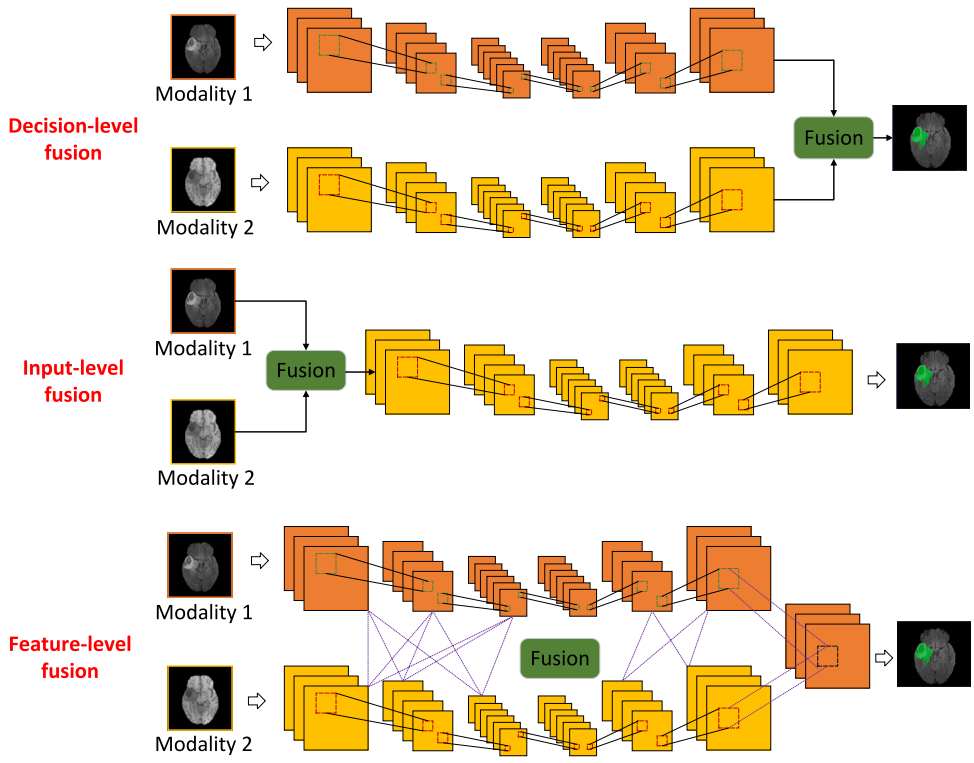

Fig. 1. Multi-modal data can be merged using three fusing methods. From top to bottom: decision-level fusion, input-level fusion, and feature-level fusion.

concatenating them. As shown in Fig. 2, convolution maps each independent channel feature to another feature space in order to further reduce the semantic gap, and then adds them element-by-element. Due to the independence of channel features, the concatenation fusion method allows for the fusion of multi-modal information as well by combining them element-by-element. Due to the lack of the mechanism to examine the correlation between different modal features before combining them, both fusion methods are unable to fully exploit the multi-modal information. To further explore the potential correlation between multi-modal information, we propose a LAA block based on self-attention (Vaswani et al., 2017). As Fig. 2, the LAA block has the capability of establishing not only local correlation between various modalities but also locating the region of interest and suppressing invalid regions from influencing the fusion process by incorporating the local attention mechanism before combining the different modalities.

In this study, the key contributions are as follows:

• We propose a local awareness fusion module, LAA block, which addresses the weaknesses of commonly used fusion techniques. By employing a local attention mechanism, the block aggregates local information across multi-modal data sets, exploits the association between them, and locates the region of interest.

• We present a segmentation model for multi-modal medical images that can effectively fuse multi-modal information, called TranSiam. It is capable of extracting multi-modal features and fusing them together with powerful feature extraction tools.

2. Related work

With the development of medical image segmentation techniques, many classical convolutional neural networks (CNNs) models had been proposed, including Unet (Ronneberger, Fischer, & Brox, 2015), CENet (Gu et al., 2019), Nested Unet (Zhou, Rahman Siddiquee, Tajbakhsh, & Liang, 2018), etc. Typically, CNN models employed a convolutional layer, a pooling layer, and a filtering layer to filter redundant information (Ma, Tang and Guo, 2021). It was important to note that convolution possesses the characteristics of local and shared weights, where the local-aware characteristic allowed these CNN models to extract detailed information efficiently at the low-level layer, however they could not establish long-distance pixel associations. The shared weight characteristic allowed the models to have less computation and parameters. Furthermore, CNN exhibited translational invariance and rotational equivariance. It was possible to summarize these characteristics as the CNN inductive bias. As a result of these inductive biases, CNNs were able to achieve excellent results on small datasets, including medical images.

Since Dosovitskiy et al. (2020) demonstrated the ability of Transformer to solve image classification tasks, Transformer has begun to be applied to image segmentation tasks. Using the self-attention mechanism, it establishes associations between global information and the MLP layer to extract features. Thus, Transformer is capable of extracting global information, which compensates for CNN's deficiencies. Transformer is not sensitive to detailed information and is computationally expensive when used at the low-level layer. Accordingly, Liu et al. (2021) proposed the Swin transformer as a method to calculate self-attention based on local windows. Swin transformer achieves state-of-the-art performance in a wide range of image tasks and reduces the computation requirements of Transformer significantly. Subsequently, Hassani, Walton, Li, Li, and Shi (2022) proposed a more flexible local window-based Transformer known as NAT transformer, which further reduced the computation time. However, they lack the inductive bias of CNNs and require a large amount of training data. Medical image data is difficult to obtain, and detailed information is crucial to the quality of medical care, so applying it to medical image segmentation is challenging. Therefore, we combine CNN and Transformer to extract detailed and local features in the low-level layer and global features in the high-level layer. In addition to reducing the computational cost, this method retains detailed information and can extract global information.

For multi-modal medical image fusion, previous works usually fused multi-modal data at input-level, decision-level and feature-level. Myronenko (2018) treated multi-modal images as different image channels.

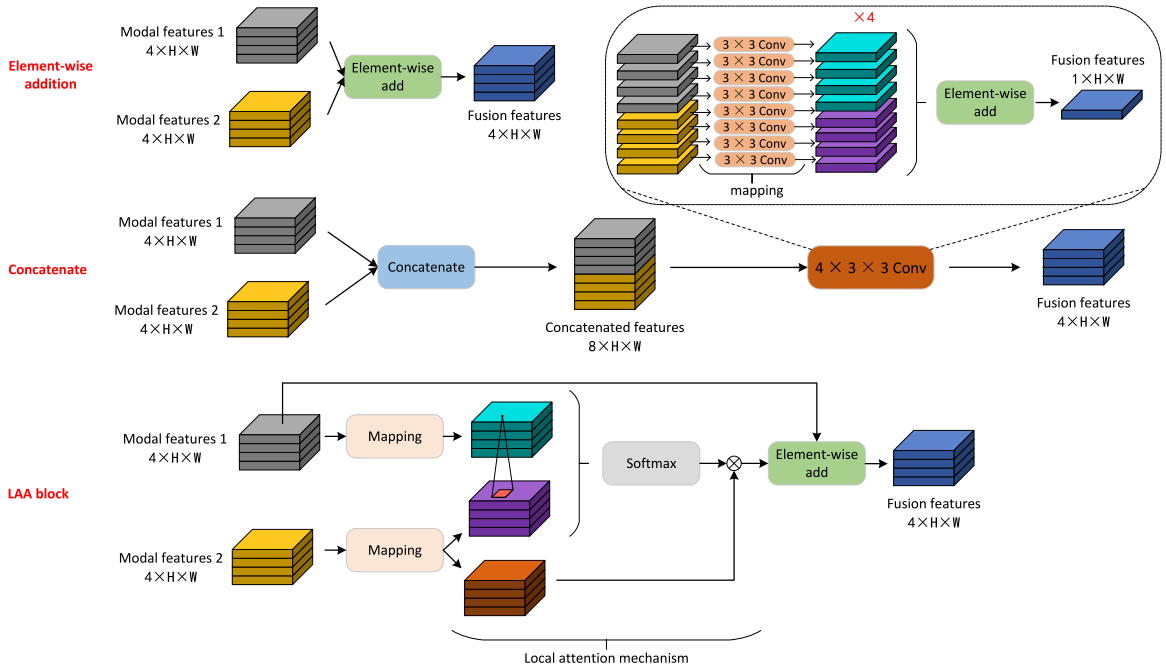

Fig. 2. The fusion process of Element-wise Addition, Concatenation, and LAA block. Here, $4 \times H \times W$ means the channel, height and width of the feature maps respectively; $4 \times 3 \times 3$ Conv represents the number and size of the convolution respectively. Here is an example of a LAA block fusion in which modal features 1 and 2 are fused.

Similarly, Jiang, Ding, Liu, and Tao (2019) also concatenated multi-modal images simply as different channels. Wang, Zhang et al. (2020) extracted features of different modalities by multiple models, and then densely concatenated the multi-modal features for fusing the multi-modal information. Similarly, there are also Dolz, Desrosiers et al. (2018) and Dolz, Gopinath et al. (2018). Based on the excellent performance of the transformer in the field of image processing and analysis, Hatamizadeh, Yang, Roth, and Xu (2021) and Zhang et al. (2022) combine CNN with the transformer to improve the segmentation results of multi-modal brain tumor. Different modal data usually contain complementary information, and effective fusion of these information can greatly improve the segmentation effect. However, previous work paid little attention to the fusion of multi-modal medical images. Therefore, we explore a novel fusion mechanism based on the commonly used feature-level fusion methods.

3. Methods

In this study, we propose TranSiam, an improved feature-level fusion model for multi-modal medical image fusion. TranSiam is a dual-path network that extracts features from various modalities independently. Additionally, we combine CNN and Transformer as a backbone in order to not only maintain detailed information, but also retrieve global information. We also design a LAA block as a means of efficiently fusing the features of different modalities. As part of the LAA block, we explore the possible correlation between multi-modal features, as well as locate the region of interest by using attention mechanisms.

3.1. TranSiam

TranSiam is composed of two sub-networks, sub-network 1 and sub-network 2, both of which are identical, as shown in Fig. 3. In the low-level layer of the sub-network, convolution is used to extract detailed information, and pooling is used to expand receptive fields and remove redundant information. It is important to note that excessive use of pooling layers can also result in the loss of detailed information. Therefore, we reduce the number of pooling layers. Nevertheless, this leads to a very large feature map resolution in the high-level layer, which means that convolution is not able to provide the global information that is required in the application of the feature map. As a result, in the high-level layer, we use Transformer to extract global information. In addition, skip connections are used to compensate for the information loss caused by the pooling layer. Feature extraction from two different modalities is accomplished by utilizing two sub-networks. In TranSiam, the multi-modal features of each layer are merged using two LAA blocks. Sub-network 1's features are mapped to another feature space as query vectors, while sub-network 2's features are mapped as key and value vectors. It is possible to mine the potential correlation between these vectors through the LAA block, and finally obtain the merged features based on the correlation between the vectors. In this way, the top LAA block is used to fuse the features of sub-network 2 to sub-network 1. Conversely, the bottom LAA block is used to fuse the features of sub-network 1 to sub-network 2. Using this bi-directional fusion method, the information between multi-modal data can be fully utilized and interference can be avoided between them during the fusion process. In addition, LAA blocks are used in each layer of TranSiam to achieve multi-scale fusion, which can greatly enhance the fusion effect. Finally, the predicted results may be obtained by combining the outputs of the two sub-networks.

TranSiam is not limited to the fusion of two modal data. It can be extended to tasks with more modal data. Based on the segmentation task, we can classify the multi-modal data into two categories: task-dominant and task-relevant. A task-dominated modal data set is one that can dominate the task of segmentation, while a task-related data set is one that will affect the task of segmentation, and both types of data are used as inputs into two sub-networks. Using this approach, TranSiam can be applied to medical image segmentation tasks with a broader range of modalities. Take MRI data as an example, the imaging methods of different modalities are different. T2 signal and FLAIR signal are related to water content, and the intensity of T2 signal and FLAIR signal in edematous areas is stronger than that of surrounding normal tissues and is often highlighted. Unlike T2 and Flair sequences, T1 and T1ce sequences are not sensitive to edema, and it is difficult to distinguish edematous regions from normal tissue by T1 and T1ce sequences. Therefore, they can be used as task-relevant data to provide additional information for segmenting lesion areas containing edema, while flair and t2 are used as task-dominant data

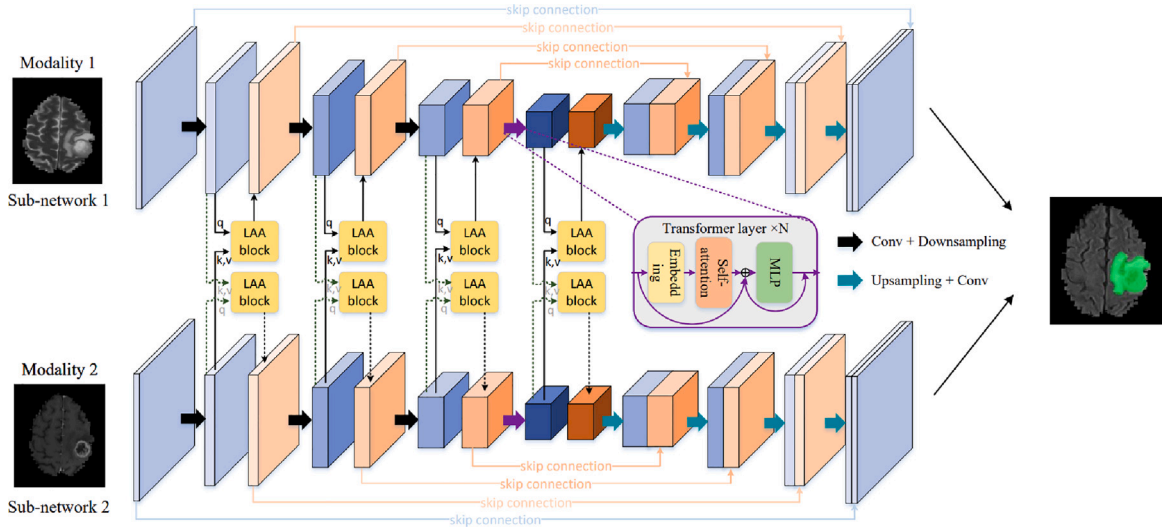

Fig. 3. An overview of TranSiam's structure. It consists of two sub-networks that extract different modal features separately. The LAA blocks are used in two layers to fuse multi-modal features bi-directionally.

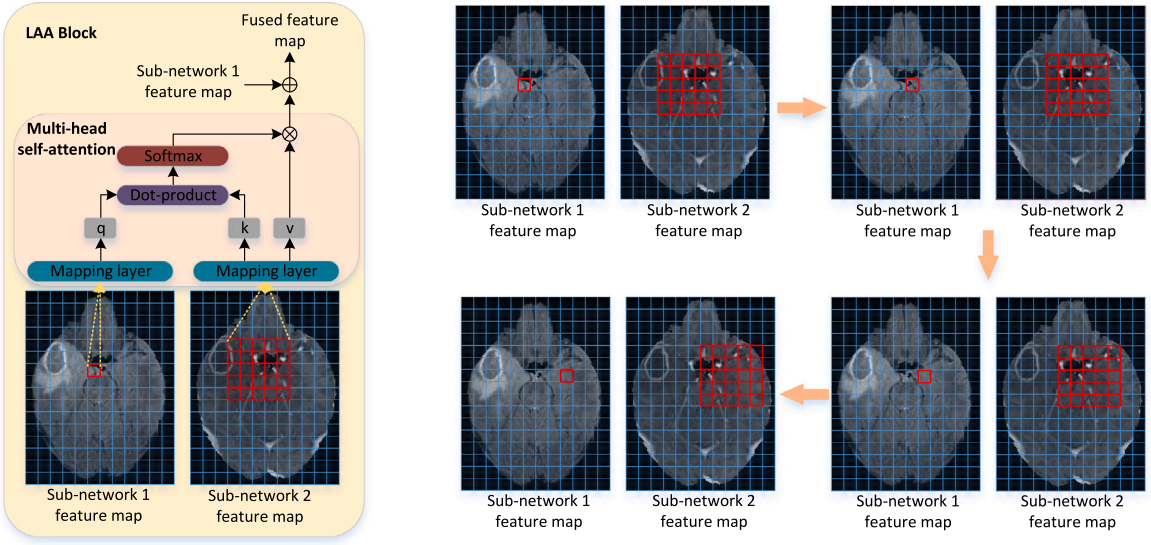

Fig. 4. An overview of the LAA block structure. The fusion window size is 5 × 5. From left to right: the fusion process of sub-network 2 features fused to sub-network 1 features, the process of sliding the fusion window.

3.2. Local attention fusion block

At the feature-level, existing multi-modal fusion methods generally lack the mechanism to establish the correlation between different modalities, so that these methods cannot fully fuse multi-modal information. Therefore, we propose the LAA block, which maps different modal features into multiple feature spaces and exploits the potential correlation between multi-modal features from different perspectives using the self-attention mechanism. Meanwhile, the LAA block emphasizes regions of interest and suppresses invalid regions from influencing fusion.

Fig. 4 shows the internal structure of the LAA block and its fusion window sliding process. Suppose that the feature map size of sub-network 1 and sub-network 2 is $B \times C \times H \times W$, where B represents the number of samples in a batch, C represents the number of channels, and H and W represent the height and width of the feature map, respectively. First, the LAA block uses a 3×3 convolution as the mapping layer to map the feature map of sub-network 1 to the new feature space as a query (q) vector. Similarly, the feature map of sub-network 2 is represented as a key (k) vector and a value (v) vector. In this case, the dimensions of the q, k, and v vectors are $B \times N \times HW \times C$, where N refers to the number of feature spaces. The LAA block maps the original features to multiple new feature spaces, and calculates the inner product of q vectors and k vectors by Eq. (1) in different feature spaces. In this way, the correlation between the feature map of sub-network 2 and sub-network 1 can be established in several different feature spaces, so as to obtain various association matrices. As a result of establishing an association between two modalities using multiple feature spaces, we are able to fully explore the potential correlation between them from different perspectives.

$$s = q * k^T \tag{1}$$

where q and k represent query vectors and key vectors, respectively. Unlike the standard mechanism of self-attention, this q vector only queries the features of the k vector in the local window centered on it.

In this case, the association matrix (s) has dimensions of $B \times N \times HW \times HW$. Using the Softmax layer, the association matrix is transformed into an attention matrix, and the element-multiplication operation is performed on the v vector to obtain the weighted features. By implementing this attention mechanism, the sub-network 2 feature map can be highlighted for high correlations with the sub-network 1 feature map, thus suppressing the influence of invalid regions on the fusion process. Finally, the LAA block performs Element-wise Addition operations with the features of sub-network 1 to obtain the final fusion feature. Eq. (2) illustrates the fusion process in detail.

$$f = \frac{softmax(s)}{\sqrt{d_k}} v + x \tag{2}$$

where d_k represents the dimension of the k vector and x represents the feature map of sub-network 1. By including this local window-based self-attention mechanism in the fusion module, we can establish associations between multi-modal features while eliminating the interference of invalid information. In addition, the LAA block is flexible. By resizing the local window, it can be flexibly applied to different datasets. It is possible to adopt large windows for large targets, and small windows for small targets, which has a good application potential.

3.3. Discussion

To explore the differences between the LAA block and common fusion methods in feature-layer, such as Concatenation and Element-wise Addition. We split the fusion process of Concatenation and Element-wise Addition, as shown in Fig. 2. Suppose the number of channels of the original feature map is 4, and the size of the feature map is $H \times W$. By the Element-wise Addition fusion method, the fusion features with size $4 \times H \times W$ can be obtained. Compared with the Element-wise Addition, the Concatenation method contains a mapping operation. It maps the features of each channel to an additional feature space, and then combines these channel features by Element-wise Addition to generate a fusion feature map with size $1 \times H \times W$. Repeating this process four times can obtain fusion features with size $4 \times H \times W$. Essentially, the Concatenation method maps the features of different modalities into four different feature spaces, and then fuses these modal features in each feature space by Element-wise Addition. Therefore, the Concatenation method is more comprehensive in fusing features than the Element-wise Addition method. Despite the fact that these two methods can obtain fusion features, the features of different modalities must be independent. They fail to explore the potential association between multi-modal features. Therefore, we propose the LAA block. As shown in Fig. 2, the LAA block establishes the association between multi-modal features through a local attention mechanism before Element-wise Addition, and enhances the information exchange between modalities. The detailed structure of the LAA block is shown in Fig. 4. It can be seen that the LAA block uses multi-head local self-attention, which means that the LAA block also maps the modal features to multiple feature spaces and then explores the local correlation between the multi-modal features in each feature space. Therefore, the LAA block also has the advantages of the Concatenation method. In addition, the LAA block can locate regions of interest and suppress the influence of invalid regions.

Table 1
Summary of the dataset including BraTS 2018, BraTS 2019, BraTS 2020 and Vestibular Schwannoma.

Dataset	Training set	Validation set	Image size
BraTS 2018	285	66	240 × 240 × 155
BraTS 2019	335	125	240 × 240 × 155
BraTS 2020	369	125	240 × 240 × 155
Vestibular Schwannoma	242	48	512 × 512 × 120

3.4. Loss function

We employ a joint loss to enhance our network's training efficiency and segmentation accuracy. The joint loss consists of cross-entropy loss and Dice loss (Milletari, Navab, & Ahmadi, 2016). The Dice loss focuses on the prediction results of the foreground area and reduces the effect of class imbalance. However, Dice loss is difficult to converge when the foreground area is small. Therefore, we combine the cross-entropy loss L_{CE} with the Dice loss L_{dice} to alleviate the difficulty. The joint loss can be formulated as follows:

$$L_{joint} = \alpha_1 L_{CE} + \alpha_2 L_{dice} \tag{3}$$

$$L_{dice} = \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{4}$$

$$L_{CE} = -\sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} ln \hat{y}_{nk} + (1 - y_{nk}) ln(1 - \hat{y}_{nk})) \tag{5}$$

where the target matrix is an $N \times K$ matrix. y is the ground truth. $\hat{y}$ refers to the prediction. $alpha_1$ and $alpha_2$ are the loss weights of the multi-class cross-entropy loss and Dice loss. $alpha_1$ and $alpha_2$ are set to 0.3 and 0.7, respectively.

4. Experiments and results

4.1. Dataset

We evaluate the performance of TranSiam on the BraTS 2018, BraTS 2019 dataset, BraTS 2020 dataset and Vestibular Schwannoma dataset (see Table 1). Using the BraTS 2020 dataset as an example, the BraTS 2020 dataset is divided into two parts: training set and validation

Table 2
Ablation experiment of LAA block on BraTS 2020 dataset, w/o means without, w/ means with.

Method	Dice score (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	Hausdorff Dist. (mm) ↓
TranSiam w/o LAA	85.43	81.05	**99.93**	5.37
TranSiam w/ LAA	**89.46**	**88.38**	99.91	**5.02**

set. The training set contains 369 3D MRI images and the validation set contains 125 3D MRI images. The validation set does not provide labels, and we need to upload the results of the inference to the online evaluation system for obtaining evaluation results. The size of all MRI images is 240 × 240 × 155. The training and validation sets both include glioblastoma (GBM/HGG) and lower grade glioma (LGG). In this paper, we segment the whole tumor area. In addition, the BraTS dataset contains data from four modalities: T1ce, T1, Flair, and T2. We divided them into two categories: task-dominant and task-relevant. Task-dominant contains Flair and T2, because they are more sensitive to the whole tumor area. T1ce and T1 are seen as task-relevant. Finally, these two types of data are used as the input of two sub-networks respectively.

Vestibular Schwannoma dataset contains a labeled dataset of MRI images collected on 242 consecutive patients with Vestibular Schwannoma (VS) undergoing Gamma Knife stereotactic radiosurgery (GK SRS). The structural images included contrast-enhanced T1-weighted (ceT1) images and high-resolution T2-weighted (hrT2) images. The size of all MRI images is $512 \times 512 \times 120$. The ceT1 and hrT2 are unaligned, but the dataset provides the transformation matrix. They are used as inputs to our model after alignment. The training and test set contain 242 samples and 48 samples, respectively.

4.2. Evaluation metrics

We employ four conventional metrics to quantitatively evaluate the segmentation performance of TranSiam. They are Dice similarity coefficient (%), Specificity (%), Sensitivity (%) and Hausdorff distance (HD95) (mm) respectively. Dice similarity coefficient calculates the volume overlap between the prediction mask and the ground truth. Sensitivity and specificity are statistical measures of the performance of binary classification tests. Hausdorff distance calculates the distance between the prediction mask and the ground truth in metric space.

4.3. Experiment details

Our experiments are mainly trained on an NVIDIA V100. All models are trained for 100 epochs. The SGD is used as our optimizer. The initial learning rate is set to 0.03, the weight decay is 0.0001 and the momentum is 0.9. We use conventional strategies for image augmentation including random flip, random rotate, random scale and random crop. The probability of each strategy being used is 0.5. In the test phase, we use the weights of the last 5 epochs to make predictions on the test set separately. We take the average of 5 results as the final result to enhance the generalization ability of our model.

4.4. Ablation study

In this section, we design ablation experiments to evaluate the performance of each component of TranSiam. As shown in Table 2, we remove or add the LAA block to TranSiam. The Dice of TranSiam without LAA block reaches 85.43%, which is much lower than that of TranSiam with LAA block at 89.46%. In addition, the Sensitivity of TranSiam with LAA block reaches 88.38%, which is 7.33% higher than that of TranSiam without LAA block. This means that the LAA block is used to fuse multi-modal features is important and can significantly improve the segmentation accuracy of the model.

Table 3
Ablation experiment of LAA block number on BraTS 2020 dataset.

Pair number	Dice score (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	Hausdorff Dist. (mm) ↓
1	88.68	88.17	99.91	6.02
2	89.10	88.00	99.91	5.94
3	89.32	88.20	**99.92**	5.30
4	**89.46**	**88.38**	99.91	**5.02**

Table 4
Comparison of different sizes of fusion window size on BraTS 2020 dataset.

Window size	Dice score (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	Hausdorff Dist. (mm) ↓
3	89.25	88.24	**99.92**	5.11
5	89.38	88.35	99.91	5.12
7	89.34	88.43	99.90	5.00
9	89.27	88.21	99.90	4.87
11	**89.46**	88.38	99.91	5.02
13	89.39	**88.62**	99.91	5.86
15	89.26	88.13	**99.92**	5.47
17	89.42	88.18	**99.92**	**4.73**

In addition, we design ablation experiments for the number of LAA block, which is shown in Table 3. Since our LAA block is based on the attention mechanism of local window, the LAA block builds only the local correlation between multi-modal features at the low-level layer. As the network deepens, the window size of the LAA block remains constant, so using the LAA block at each layer of the network enables the fusion of multi-scale features. Since LAA blocks are always used in pairs, which are used to fuse different modal features separately. So, we gradually add LAA block in pairs from the high-level layer to the low-level layer of the model to verify the effectiveness of multi-scale fusion. As shown in Table 3, with the increase of the number of LAA block, the Dice of TranSiam gradually is increased from 88.68% to 89.46%, and the HD95 gradually is decreased from 6.02 mm to 5.02 mm. Besides, the Sensitivity of TranSiam reaches the highest 88.38% when the number of LAA block is 4. The results indicate that multi-scale fusion is meaningful.

4.5. Evaluation of fusion window size

In the LAA block, the different sizes of fusion windows have different effects on the fusion efficiency. In this section, we try different sizes of fusion windows to fuse multi-modal features. From Table 4, we can see that a model with the fusion window of size 3 achieves a segmentation Dice of 89.25%. This suggests that the fusion between multi-modal features does not require an excessively large receptive field for multi-modal fusion tasks, and the correlation of local information between them is stronger. As the size of the fusion window increases, Dice reaches its highest at the window size of 11. In TranSiam, the default size of fusion window is 11.

4.6. Comparison with other multi-modal methods

Recent works usually use Concatenation or Element-wise Addition to fuse multi-modal features on the feature-level. However, they both lack the mechanism to establish the association between multi-modal features before fusing them. Therefore, we propose the LAA block, which uses local self-attention to explore the local correlation between multi-modal features before fusing them. In this section, we design comparison experiments to verify the effectiveness of the improvement. LAA block is a feature level fusion method, and considering that the backbone of the model may have an impact on the performance of fusion, we implement other feature-level fusion methods on TranSiam. We reproduce classical medical image segmentation algorithms and

Table 5
Comparison with other multi-modal methods on BraTS 2020 validation dataset.

Types	Methods	Dice score (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	Hausdorff Dist. (mm) ↓
Input-level fusion	Unet (Ronneberger et al., 2015)	88.63	86.91	**99.92**	5.49
	CENet (Gu et al., 2019)	88.98	87.23	**99.92**	5.14
	Probabilistic UNet (Kohl et al., 2018)	81.9	84.61	99.81	41.52
	TransBTS w/o TTA (Wang, Chen et al., 2021)	89.00	–	–	6.47
	TransUnet (Chen et al., 2021)	88.97	87.62	**99.92**	5.44
	SETR (Zheng et al., 2021)	85.85	84.98	99.89	14.37
	Swin Transformer (Liu et al., 2021)	86.10	85.62	99.87	8.54
	NAT (Hassani et al., 2022)	87.59	86.97	99.89	7.39
	Swin Unet (Cao et al., 2023)	86.39	86.07	99.88	9.86
	UNetFormer (Wang, Li et al., 2022)	87.83	85.87	**99.92**	6.15
	Uctransnet (Wang, Cao, Wang and Zaiane, 2022)	88.85	87.71	99.91	5.38
Feature-level fusion	Concatenation	88.97	87.62	**99.92**	5.44
	Element-wise Addition	88.44	87.81	99.90	6.81
	MAML (Zhang et al., 2021)	88.99	87.77	99.91	6.27
	Modality Pairing (Wang, Zhang et al., 2020)	88.90	87.68	99.90	6.81
	TranSiam (Ours)	**89.46**	**88.38**	99.91	**5.02**

Note: "–" denotes not mentioned in the original paper.

Table 6
Comparison with other multi-modal methods on BraTS 2018 validation dataset.

Types	Methods	Dice score (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	Hausdorff Dist. (mm) ↓
Input-level fusion	Bit-plane UNet (Tuan, Tuan, & Bao, 2019)	81.87	77.34	**99.53**	9.42
	voxel-Gan (Rezaei, Yang, & Meinel, 2019)	84	86	99	6.41
	CGan (Rezaei et al., 2018)	81	75	99	7.30
	Deep Hourglass (Benson, Pound, French, Jackson, & Pridmore, 2019)	82	–	–	26.41
	AGResU-Net (Zhang, Jiang, Dong, Hou, & Liu, 2020)	87.2	–	–	5.62
	AFPNet (Zhou, He and Jia, 2020)	84.94	–	–	–
	DenseAFPNet (Zhou, He, Shi, Du and Chen, 2020)	86.42	–	–	–
	3D Unet (Wang, Jiang, Zheng, Meng and Biswal, 2020)	86.38	–	–	12
	Weninger, Rippel, Koppers, and Merhof (2019)	88.9	88.7	99.5	6.97
	Ma and Yang (2019) (Ordinary fusion)	85.1	–	–	8.64
Feature-level fusion	Ma and Yang (2019) (Complementary fusion)	87.2	–	–	6.12
	Marcinkiewicz, Nalepa, Lorenzo, Dudzik, and Mrukwa (2019)	89.8	**90.96**	99.35	–
	Concatenation	89.46	88.95	99.23	5.21
	Element-wise Addition	89.51	88.84	99.49	4.6
	MAML (Zhang et al., 2021)	89.04	88.45	99.44	4.97
	Modality Pairing (Wang, Zhang et al., 2020)	89.57	88.66	99.5	5.08
	TranSiam (Ours)	**90.11**	90.0	99.46	**4.55**

recent Transformer-based segmentation algorithms (Input-level), including Unet, CENet, SETR, and Swin Transformer, NAT, Swin Unet, UNetFormer, Uctransnet, etc, under the same experimental settings and model parameters. We compare TranSiam with classical segmentation methods on the BraTS 2019 dataset, BraTS 2020 dataset and the Vestibular Schwannoma dataset. Tables 6, 7, 5 and 8 show the segmentation results of TranSiam and other classical segmentation methods on the BraTS 2018, BraTS 2019 dataset, BraTS 2020 dataset and Vestibular Schwannoma dataset, respectively.

Table 5 shows the performance of common methods on the BraTS 2020 dataset. In particular, Average (Decision-level) represents the fusion method on the decision-level, where we calculate the average of multiple outputs as the final result. Besides, Concatenation (Input-level) represents the fusion method on the input-level, where we concatenate multi-modal images as different image channels. It achieves Dice of 88.97%, which is significantly better than the Decision-level fusion method. Modality Pairing (Feature-level) indicates densely concatenate multi-modal features on the feature-level. It achieves Dice of 88.90%, which is better than Element-wise Addition (Feature-level). This means that the feature mapping process is important. TranSiam (Feature-level) achieves the best Dice, Sensitivity and HD95. The Dice of TranSiam can reach 89.46% on the BraTS 2020 dataset, which is higher than the CNN-based and Transformer-based classical segmentation methods. In addition, the Sensitivity and HD95 of TranSiam reach 88.38% and 5.02 mm respectively, which are better than other methods. This means that the LAA block establish the association between multi-modal features is meaningful.

To evaluate the performance and generalization of the TranSiam, experiments are also conducted on the BraTS 2018 validation dataset and the BraTS 2019 validation dataset. The performance is shown in Tables 6 and 7. It can be seen that our TranSiam achieves satisfying performance in several metrics compared to classical segmentation methods. TranSiam achieves Dice score of 89.44%, sensitivity of 88.5%, specificity of 99.42%, Hausdorff distance of 5.09 mm on the BraTS 2019 validation dataset and achieves Dice score of 90.11%, sensitivity of 90.0%, specificity of 99.46%, Hausdorff distance of 4.55 mm on the BraTS 2018 validation dataset. Compared to Modality Pairing, Element-wise Addition, Concatenation and MAML, TranSiam shows satisfying performance. Furthermore, compared to classical medical image segmentation models such as 3D Unet, V-Net, KiUnet and Attention Unet, our method is also competitive.

On the Vestibular Schwannoma dataset, we also design comparison experiments to verify the generalization of TranSiam, as shown in Table 8. Since Vestibular Schwannoma is usually small in size, the fusion window size of the LAA block is set to 5. For Vestibular Schwannoma segmentation, the Dice 91.67% of Modality Pairing (Feature-level) is also higher than that of Element-wise Addition (Feature-level). TranSiam achieves 92.62%, 77.17%, and 96.88% for Dice, Sensitivity, and Precision, respectively, which are higher than other segmentation methods. In summary, TranSiam can greatly improve the segmentation effect by using LAA block to fuse multi-modal features. The results show that it is important to utilize multi-modal features efficiently.

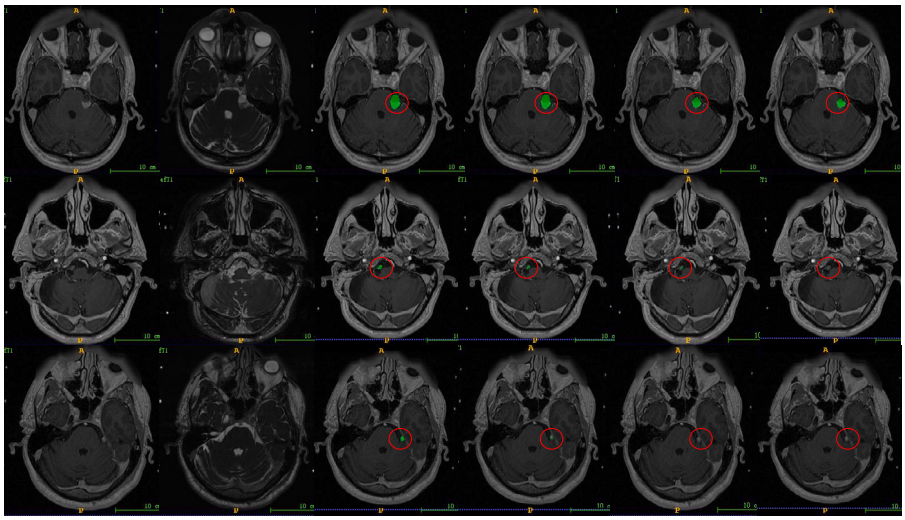

Fig. 5. Visual analysis of fusion methods. From left to right: cet1, hrT2, label, LAA block, Element-wise Addition, and concatenation. Tumors are marked in green.

Table 7
Comparison with other multi-modal methods on BraTS 2019 validation dataset.

Types	Methods	Dice score (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	Hausdorff Dist. (mm) ↓
Input-level fusion	MD-Unet (Ge et al., 2021)	87.2	87.2	99.3	5.41
	3D Unet (Ronneberger et al., 2015)	87.38	–	–	9.43
	V-Net (Milletari et al., 2016)	88.73	–	–	6.25
	2D KiUnet (Valanarasu, Sindagi, Hacihaliloglu, & Patel, 2022)	86.12	–	–	12.79
	3D KiUnet (Valanarasu et al., 2022)	87.6	–	–	8.94
	TransBTS w/o TTA (Wang, Chen et al., 2021)	88.89	–	–	7.59
	Swin Unet (Cao et al., 2023)	89.38	–	–	7.5
	Attention Unet (Oktay et al., 2018)	88.81	–	–	7.75
	Wang, Jiang et al. (2020)	89.40	–	–	5.67
	Li, Luo, and Wang (2020)	88.60	–	–	6.23
	Myronenko (2018)	89.40	–	–	5.89
Feature-level fusion	Concatenation	89.20	87.82	**99.51**	**4.77**
	Element-wise Addition	88.98	87.74	99.44	5.1
	MAML (Zhang et al., 2021)	88.86	87.48	99.46	6.5
	Modality Pairing (Wang, Zhang et al., 2020)	88.75	87.27	99.33	5.86
	TranSiam (Ours)	**89.44**	**88.50**	99.42	5.09

Table 8
Comparison with other multi-modal methods on Vestibular Schwannoma dataset.

Types	Methods	Dice score (%) ↑	Sensitivity (%) ↑	Precision (%) ↑
Input-level fusion	Unet (Ronneberger et al., 2015)	91.35	77.02	94.96
	CENet (Gu et al., 2019)	90.44	75.97	94.52
	TransUnet (Chen et al., 2021)	91.00	77.15	93.88
	SETR (Zheng et al., 2021)	85.51	73.99	90.43
	Swin Transformer (Liu et al., 2021)	81.50	71.86	86.43
	NAT (Hassani et al., 2022)	83.10	71.27	89.87
	Swin Unet (Cao et al., 2023)	77.25	76.15	81.06
	UNetFormer (Wang, Li et al., 2022)	89.86	88.39	92.49
	Uctransnet (Wang, Cao et al., 2022)	87.93	**88.51**	88.33
Feature-level fusion	Concatenation	90.29	72.49	**98.55**
	Element-wise Addition	91.21	76.74	94.80
	MAML (Zhang et al., 2021)	90.67	**89.07**	93.63
	Modality Pairing (Wang, Zhang et al., 2020)	91.67	77.25	95.37
	TranSiam (Ours)	**92.62**	77.17	96.88

4.7. Visual analysis

To readily compare the advantages of the LAA block compared to other fusion methods, we visualize the prediction results on the Vestibular Schwannoma dataset. Based on the TranSiam, we only replace the fusion method under the same experimental settings. We compare the LAA block with commonly used fusion methods on the feature-level, including Element-wise Addition and Concatenation. As shown as the first row of Fig. 5, complete segmentation of the tumor region is required to effectively combine ceT1 and hrT2, because these two modalities reflect complementary tumor information. Compared with Element-wise Addition and Concatenation, the LAA block can utilize the multi-modal data more effectively. In addition, since the LAA block is a local window-based fusion mechanism, it does not lose small target information excessively. As shown in the second and third rows of Fig. 5, the LAA block is sensitive to segment the small tumor region.

We also compare TranSiam with other classical segmentation models on the Vestibular Schwannoma dataset. Fig. 6 shows the visual

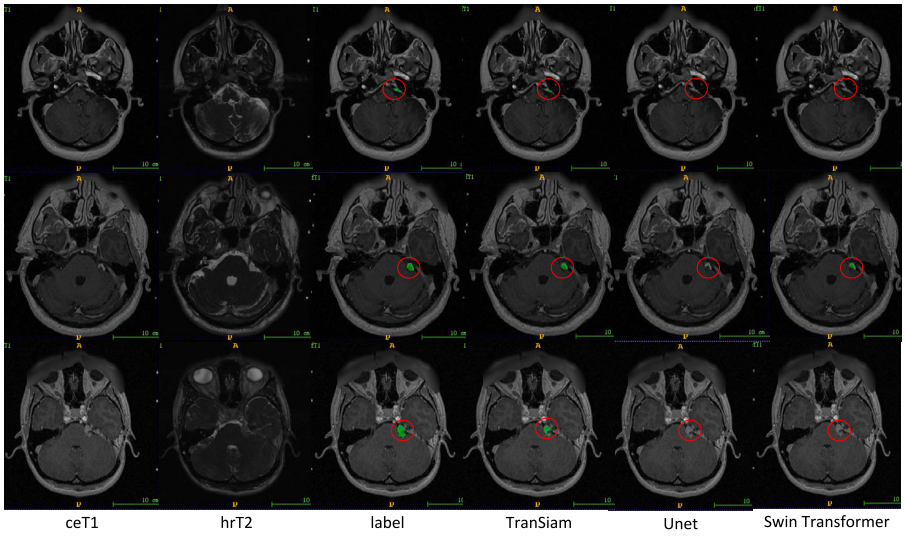

Fig. 6. Visual analysis of segmentation methods. From left to right: cet1, hrT2, label, TranSiam, Unet, and Swin Transformer. Tumors are marked in green.

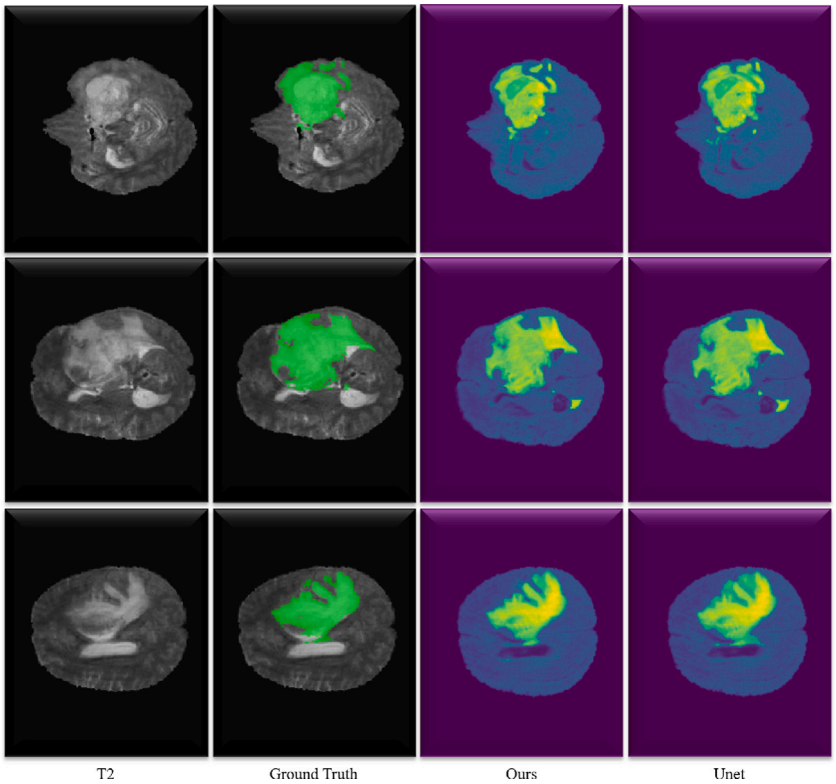

Fig. 7. An illustration of the activation map of the last convolutional layer. From left to right: T2, ground truth, our approach and Unet. The bright yellow area represents areas with higher activation values (attention). Our method uses LAA blocks to learn complementary information between different modalities, providing richer visual representations for segmentation tasks. The comparison between our method and Unet's activation map shows that our method can make the model more focused on lesion region.

results of Unet, Swin Transformer and TranSiam. It can be seen that the TranSiam can segment more tumor regions. The visual comparison of MRI brain tumour segmentation results shows that it is useful to fully fuse multi-modal data for medical image segmentation.

To further investigate the multi-modal correlation feature representation obtained by the LAA block, we visualized the feature maps in Fig. 7 and compared them qualitatively with the classical Unet model. It can be seen as a heat map where highlighted areas receive more attention and other areas have smaller pixel values. Compared to Unet, the feature maps produced by TranSiam have clearer contours in the lesion area, which helps the model to achieve more accurate segmentation results. The quantitative visual results show that the proposed LAA block can be used to perform a joint analysis on the local information of different modes, which can help the model to obtain a richer visual feature representation.

5. Conclusion

In this paper, we propose a feature-level fusion model for multi-modal medical image segmentation, called TranSiam. It has two paths, which are used to extract features from different modalities and avoid the interference of differences between them. For each path, it combines Transformer and CNN as feature extractors to extract both detailed and global information. At the feature-level, we design the LAA block to explore the potential association between different modal features. The LAA block compensates for the shortcomings of commonly used multi-modal fusion methods at the feature-level. However, in order to save computational resources, our approach is a 2D segmentation method that does not explore the association of spatial contexts in 3D medical images. In addition, we do not use any post-processing, resulting in the Specificity of our model is poor.

In the future, we will provide additional technological support to enhance the security and service availability of the proposed work in clinical environments. Inspired by the work of Conti, Militello, Rundo, and Vitabile (2021), we will design self-organizing systems with a nonfixed structure to improve service continuity, allowing them to adaptively change their structure and organization in response to internal and external environmental changes. In service-oriented networks (SONs), communication among nodes through stimulation or suppression chains will give rise to emerging behaviors that defend against external intrusions, attacks, and failures, further enhancing the robustness of our approach in clinical operational scenarios. Moreover, we are exploring how to reduce computational resources while fusing spatial contextual information of images and designing specialized post-processing to improve the effectiveness of our model segmentation.

CRediT authorship contribution statement

Xuejian Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Project administration. **Shiqiang Ma:** Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Project administration. **Junhai Xu:** Supervision, Writing – review & editing. **Jijun Tang:** Supervision, Writing – review & editing. **Shengfeng He:** Supervision, Writing – review & editing. **Fei Guo:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is supported by a grant from the National Natural Science Foundation of China (NSFC 623222215, 62172296), Excellent Young Scientists Fund in Hunan Province (2022JJ20077), Scientific Research Fund of Hunan Provincial Education Department (22A0007), Sponsored by CCF-Tencent Open Fund (NO. IAGR20220109). This work was supported in part by the High Performance Computing Center of Central South University.

References

Benson, E., Pound, M. P., French, A. P., Jackson, A. S., & Pridmore, T. P. (2019). Deep hourglass for brain tumor segmentation. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, & T. van Walsum (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 419–428). Cham: Springer International Publishing.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2023). Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer vision–ECCV 2022 workshops: Tel Aviv, Israel, October 23–27, 2022, proceedings, Part III* (pp. 205–218). Springer.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Conti, V., Militello, C., Rundo, L., & Vitabile, S. (2021). A novel bio-inspired approach for high-performance management in service-oriented networks. *IEEE Transactions on Emerging Topics in Computing*, *9*(4), 1709–1722. http://dx.doi.org/10.1109/TETC.2020.3018312.

Ding, J., Zhang, Z., Tang, J., & Guo, F. (2021). A multichannel deep neural network for retina vessel segmentation via a fusion mechanism. *Frontiers in Bioengineering and Biotechnology*, *9*, Article 697915.

Dolz, J., Desrosiers, C., & Ben Ayed, I. (2018). IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet. In *International workshop and challenge on computational methods and clinical applications for spine imaging* (pp. 130–143). Springer.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., & Ayed, I. B. (2018). HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, *38*(5), 1116–1126.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Ge, R., Cai, H., Yuan, X., Qin, F., Huang, Y., Wang, P., et al. (2021). MD-UNET: Multi-input dilated U-shape neural network for segmentation of bladder cancer. *Computational Biology and Chemistry*, *93*, Article 107510. http://dx.doi.org/10.1016/j.compbiolchem.2021.107510, URL https://www.sciencedirect.com/science/article/pii/S1476927121000773.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., et al. (2019). Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, *38*(10), 2281–2292.

Hassani, A., Walton, S., Li, J., Li, S., & Shi, H. (2022). Neighborhood attention transformer. arXiv preprint arXiv:2204.07143.

Hatamizadeh, A., Yang, D., Roth, H., & Xu, D. (2021). UNETR: Transformers for 3D medical image segmentation.

Jiang, Z., Ding, C., Liu, M., & Tao, D. (2019). Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *International MICCAI brainlesion workshop* (pp. 231–241). Springer.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., et al. (2018). A probabilistic U-net for segmentation of ambiguous images. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in neural information processing systems, Vol. 31*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2018/file/473447ac58e1cd7e96172575f48dca3b-Paper.pdf.

Li, X., Ding, J., Tang, J., & Guo, F. (2022). Res2Unet: A multi-scale channel attention network for retinal vessel segmentation. *Neural Computing and Applications*, *34*(14), 12001–12015.

Li, X., Luo, G., & Wang, K. (2020). Multi-step cascaded networks for brain tumor segmentation. In A. Crimi, & S. Bakas (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 163–173). Cham: Springer International Publishing.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Ma, S., Li, X., Tang, J., & Guo, F. (2021). A zero-shot method for 3d medical image segmentation. In *2021 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

Ma, S., Tang, J., & Guo, F. (2021). Multi-task deep supervision on attention R2U-net for brain tumor segmentation. *Frontiers in Oncology*, *11*, Article 704850.

Ma, J., & Yang, X. (2019). Automatic brain tumor segmentation by exploring the multi-modality complementary information and cascaded 3D lightweight CNNs. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, & T. van Walsum (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 25–36). Cham: Springer International Publishing.

Marcinkiewicz, M., Nalepa, J., Lorenzo, P. R., Dudzik, W., & Mrukwa, G. (2019). Segmenting brain tumors from MRI using cascaded multi-modal U-nets. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, & T. van Walsum (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 13–24). Cham: Springer International Publishing.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3d vision (3DV)* (pp. 565–571). http://dx.doi.org/10.1109/3DV.2016.79.

Myronenko, A. (2018). 3D MRI brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop* (pp. 311–320). Springer.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Rezaei, M., Harmuth, K., Gierke, W., Kellermeier, T., Fischer, M., Yang, H., et al. (2018). A conditional adversarial network for semantic segmentation of brain tumor. In A. Crimi, S. Bakas, H. Kuijf, B. Menze, M. Reyes (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 241–252). Cham: Springer International Publishing.

Rezaei, M., Yang, H., & Meinel, C. (2019). Voxel-GAN: Adversarial framework for learning imbalanced brain tumor segmentation. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, & T. van Walsum (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 321–333). Cham: Springer International Publishing.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Sun, Q., Fang, N., Liu, Z., Zhao, L., Wen, Y., & Lin, H. (2021). Hybridctrm: Bridging cnn and transformer for multimodal brain image segmentation. *Journal of Healthcare Engineering*, *2021*.

Tuan, T. A., Tuan, T. A., & Bao, P. T. (2019). Brain tumor segmentation using bit-plane and UNET. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, & T. van Walsum (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 466–475). Cham: Springer International Publishing.

Valanarasu, J. M. J., Sindagi, V. A., Hacihaliloglu, I., & Patel, V. M. (2022). KiU-Net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Transactions on Medical Imaging*, *41*(4), 965–976. http://dx.doi.org/10.1109/TMI.2021.3130469.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, H., Cao, P., Wang, J., & Zaiane, O. R. (2022). Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. *36*, In *Proceedings of the AAAI conference on artificial intelligence* (3), (pp. 2441–2449).

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021). Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention* (pp. 109–119). Springer.

Wang, F., Jiang, R., Zheng, L., Meng, C., & Biswal, B. (2020). 3D U-Net based brain tumor segmentation and survival days prediction. In A. Crimi, & S. Bakas (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 131–141). Cham: Springer International Publishing.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., et al. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 196–214.

Wang, Y., Zhang, Y., Hou, F., Liu, Y., Tian, J., Zhong, C., et al. (2020). Modality-pairing learning for brain tumor segmentation. In *International MICCAI brainlesion workshop* (pp. 230–240). Springer.

Weninger, L., Rippel, O., Koppers, S., & Merhof, D. (2019). Segmentation of brain tumors and patient survival prediction: Methods for the BraTS 2018 challenge. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, & T. van Walsum (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 3–12). Cham: Springer International Publishing.

Yang, Q., Guo, X., Chen, Z., Woo, P. Y. M., & Yuan, Y. (2022). D2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, *41*(10), 2953–2964. http://dx.doi.org/10.1109/TMI.2022.3175478.

Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., et al. (2022). *mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation*. Cham: Springer.

Zhang, J., Jiang, Z., Dong, J., Hou, Y., & Liu, B. (2020). Attention gate ResU-Net for automatic MRI brain tumor segmentation. *IEEE Access*, *8*, 58533–58545. http://dx.doi.org/10.1109/ACCESS.2020.2983075.

Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., et al. (2021). Modality-aware mutual learning for multi-modal medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24* (pp. 589–599). Springer.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). 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).

Zhou, Z., He, Z., & Jia, Y. (2020). AFPNet: A 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images. *Neurocomputing*, *402*, 235–244. http://dx.doi.org/10.1016/j.neucom.2020.03.097, URL https://www.sciencedirect.com/science/article/pii/S0925231220304847.

Zhou, Z., He, Z., Shi, M., Du, J., & Chen, D. (2020). 3D dense connectivity network with atrous convolutional feature pyramid for brain tumor segmentation in magnetic resonance imaging of human heads. *Computers in Biology and Medicine*, *121*, Article 103766. http://dx.doi.org/10.1016/j.compbiomed.2020.103766, URL https://www.sciencedirect.com/science/article/pii/S0010482520301396.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Springer.