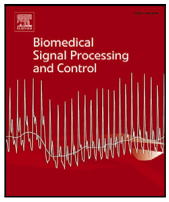
Biomedical Signal Processing and Control 95 (2024) 106383

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



MDANet: Multimodal difference aware network for brain stroke segmentation

Kezhi Zhang [a], Yu Zhu [a,*], Hangyu Li [a], Zeyan Zeng [b], Yatong Liu [a], Yuhao Zhang [b,c,d]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[b] Department of Neurology, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[c] National Clinical Research Center for Interventional Medicine, Shanghai 200032, China
[d] Shanghai Clinical Research Center for Interventional Medicine, Shanghai 200032, China

ARTICLE INFO

Keywords:
Computer-aided diagnosis
Brain stroke
Medical image segmentation
Deep learning

ABSTRACT

Stroke segmentation has great significance for clinical diagnosis and timely treatment. Medical images of strokes often come in the form of multiple modalities. But most existing methods simply stack these modalities as input, disregarding the connections and other clinical prior knowledge associated with each modality. In this paper, we present MDANet, a multimodal difference aware network for stroke segmentation based on multimodal input. The proposed network mainly consists of a difference aware module and a graph convolution fusion block. In the difference aware module, a parameter-shared encoder is adopted to extract features from different modality groups and generate difference feature maps by subtracting one group from another to enhance the perception of potential lesion areas. We further design a similarity loss to improve this ability. The graph convolution fusion block is developed to aggregate features from different modalities with a channel embedding strategy to model the features globally and a space embedding strategy for local modeling. The MDANet is trained and evaluated on the Ischemic Stroke Lesion Segmentation (ISLES) 2018 and 2022 datasets. Our approach achieves a dice score of 58.34 and 70.44, surpassing the performance of other advanced existing methods.

1. Introduction

Medical image segmentation is an essential topic for human society and health, which helps doctors make a diagnosis and carry out corresponding treatment more efficiently. Stroke segmentation is one of the most vital tasks in the field. Stroke has already been the second most common cause of death and the third most common cause of disability worldwide [1]. Computed Tomography Perfusion imaging (CTP) and Magnetic Resonance Imaging (MRI) are the usual medical imaging technique for the diagnosis of brain stroke in clinical practices, both of them can provided multiple imaging modalities. CTP maps include four modalities, namely cerebral blood flow (CBF), cerebral blood volume (CBV), mean transit time (MTT), Time To Peak of the Residue Function (TMax), as shown in Fig. 1. These perfusion maps evaluate brain health from blood volume and circulation efficiency. MRI evaluates the health status of the brain by establishing diffusion weighted map (DWI) and apparent diffusion coefficient (ADC) through the diffusion movement of water molecules in the brain [2], as shown in Fig. 2 These multimodal images can often provide reliable basis for clinical diagnosis.

Since the advent of deep learning [3], neural networks based on encoder–decoder architecture have achieved state-of-the-art performance in various medical image segmentation tasks [4–8]. However, most existing methods are designed only considering the scenario of single modality input. When facing multimodal inputs, most existing methods simply stack images of different modalities as input, or use multiple encoders to extract features from each modality and design another sub-network to fuse these features [9–11]. These methods rely on the network itself to model the relationship between different modalities, ignoring the characteristics of multimodal images themselves and their guiding role in clinical diagnosis, which probably lead to the model failing to make full use of the correlation between modalities or even introducing some unnecessary noise.

Through the observation of CT perfusion images Fig. 1, the pixel intensity of the lesion area is lower than the surrounding normal tissue in CBF and CBV. But higher in MTT and TMax. The pixel intensity mismatch between modalities can also be observed in multimodal MRI images, as shown in Fig. 2. This observation provides a valuable reference for the location of the lesion, by exploring the regions with

* Corresponding author.
E-mail addresses: zhuyu@ecust.edu.cn (Y. Zhu), zhang.yuhao@zs-hospital.sh.cn (Y. Zhang).

https://doi.org/10.1016/j.bspc.2024.106383
Received 4 December 2023; Received in revised form 15 February 2024; Accepted 23 April 2024
Available online 7 May 2024

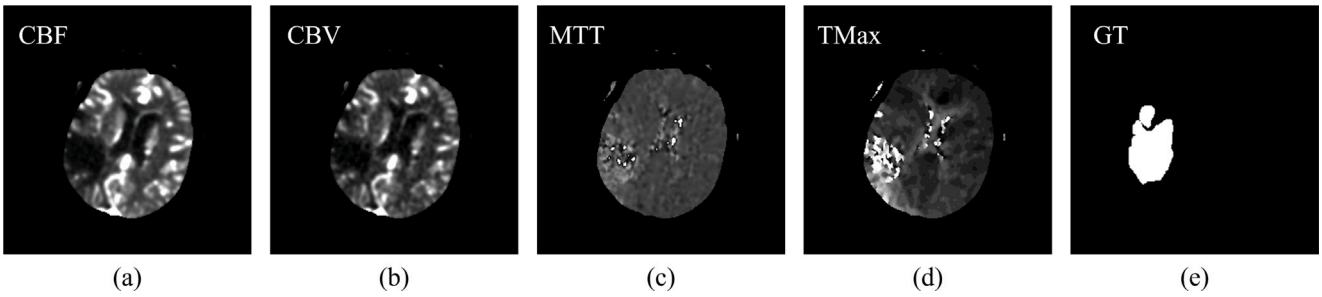


Fig. 1. An example of multi-modal perfusion maps of brain stroke segmentation. (a) The CBF modality (b) The CBV modality (c) The MTT modality (b) The TMax modality (b) The Ground Truth.

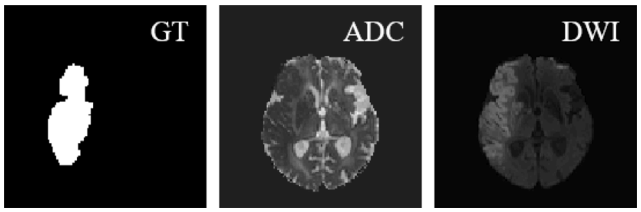


Fig. 2. An example of multi-modal MRI brain slice.

mismatched features between modalities, regions with possible lesions can be better located and segmented. Based on the above analysis, we propose MDANet, a multimodal difference aware segmentation network for brain stroke segmentation. The main contributions of this study are as follows:

(1) We propose a difference aware module. Our method adopts a parameter-sharing encoder based on multi-scale convolution to group and encode features from multimodal inputs. We subtract features of one modality group from another to extract difference feature maps and design a difference aware skip connection (DASC) to enhance sensitivity toward lesions by fusing the difference feature map in different scales.

(2) A graph convolution fusion block (GCFB) is designed to fuse the features from various modalities and improve the overall representation of the network. Specifically, we develop two distinct graph embedding strategies. GCFB(channel) builds the graph along the channel to infer the global feature and GCFB(space) generates the graph by partitioning the feature map to aggregate local features.

(3) We design a mask similarity loss function to align features in non-lesion areas of different modalities, which mitigates the interference generated by feature map subtraction. We evaluate the performance of the proposed method on the ISLES2018 and ISLES2022 datasets. The results demonstrate its promising capabilities and outperform other existing advanced methods.

2. Related works

2.1. Medical image segmentation

Image segmentation is one of the main tasks in the computer vision field. In the past few years, deep learning based on convolutional neural networks (CNN) has rapidly become the mainstream of related research. Fully convolutional network (FCN) [12] was the first end-to-end pixel-level segmentation network based on deep learning. Inspired by its encoder–decoder structure, U-Net [13] proposed a skip connection mechanism to combine the features of encoder and decoder and firstly introduced deep learning into medical image segmentation. Attention U-Net [14] added an attention gate based on U-Net to better fuse features from the encoder and decoder. U-Net++ [15] designed a series of nested, dense skip pathways to reduce the semantic gap between encoder and decoder. U-Net 3+ [16] concatenated feature maps from different layers and created a full-scale connection to incorporate low-level details with high-level semantics. ResU-Net [17] replaced convolution blocks in U-Net with modified residual blocks, using multiple parallel atrous convolutions to extract features from various receptive fields. MSNet [18] used a multi-scale subtraction connection to replace the skip connection in UNet to reduce redundant information generated during the skip connection between the encoder and decoder. Furthermore, M^2SNet [19] refined the network by equipping convolution kernel of various receptive fields at each subtraction block. With the presentation of ViT [20] in 2020, Modules based on the self-attention mechanism and transformer architecture [21] are also continuously introduced into the medical image segmentation network. TransU-Net [6] combined advantages of CNN and transformer, using CNN as an encoder to preliminary extract the local features of the input image, then encoding the global features through a series of transformer encoders. To maximize the global modeling potential of the transformer architecture. MedT [7] used shallow CNN and transformer to extract features from local and global perspectives. SwinU-Net [22] replaced all CNN blocks in traditional U-Net with shifted windows transformer blocks. These segmentation methods based on deep learning have achieved good results.

2.2. Multi-modal brain lesion segmentation

In clinical diagnosis, multimodal images are the most important diagnostic tools. Many methods for brain disease segmentation based on multimodal inputs have been proposed. Zhou et al. [4] combined 2D convolution and 3D convolution to segment stroke by dimensional fusion. Xu et al. [23] proposed a receptive field-based attention mechanism to improve stroke segmentation performance by controlling the size of the receptive field. Clèrigues et al. [24] enhanced the features of multimodal images based on the symmetry of brain hemispheres. Liu et al. [25] used transformer blocks to enhance the combination of context information. de Vries et al. [26] noted asymmetry between infarcted and healthy hemispheres and proposed a spatiotemporal attention mechanism to encode brain features and perceive differences between hemispheres.

However, these methods simply stack multimodal images as input but neglect potential correlations and complementarities between different modalities. Other works begin to research multimodal fusion methods. Aygün et al. [27] used multimodal MRI brain slices to study the effects of fusion at early, middle, and late stages on the results of brain tumor segmentation. Li et al. [28] observed the difference between the modalities, divided multimodal brain MRI into two groups, using two encoders to capture the features of different modality groups, and fusing them through self-attention and cross-attention modules. Zhang et al. [11] proposed mmFormer, which encodes each modality of brain MRI using different transformer encoders and implements another transformer module to fuse the features. Zhu et al. [29] and Marinov et al. [30] assigned different sub-tasks to different modalities of medical images to strengthen the network's learning ability from

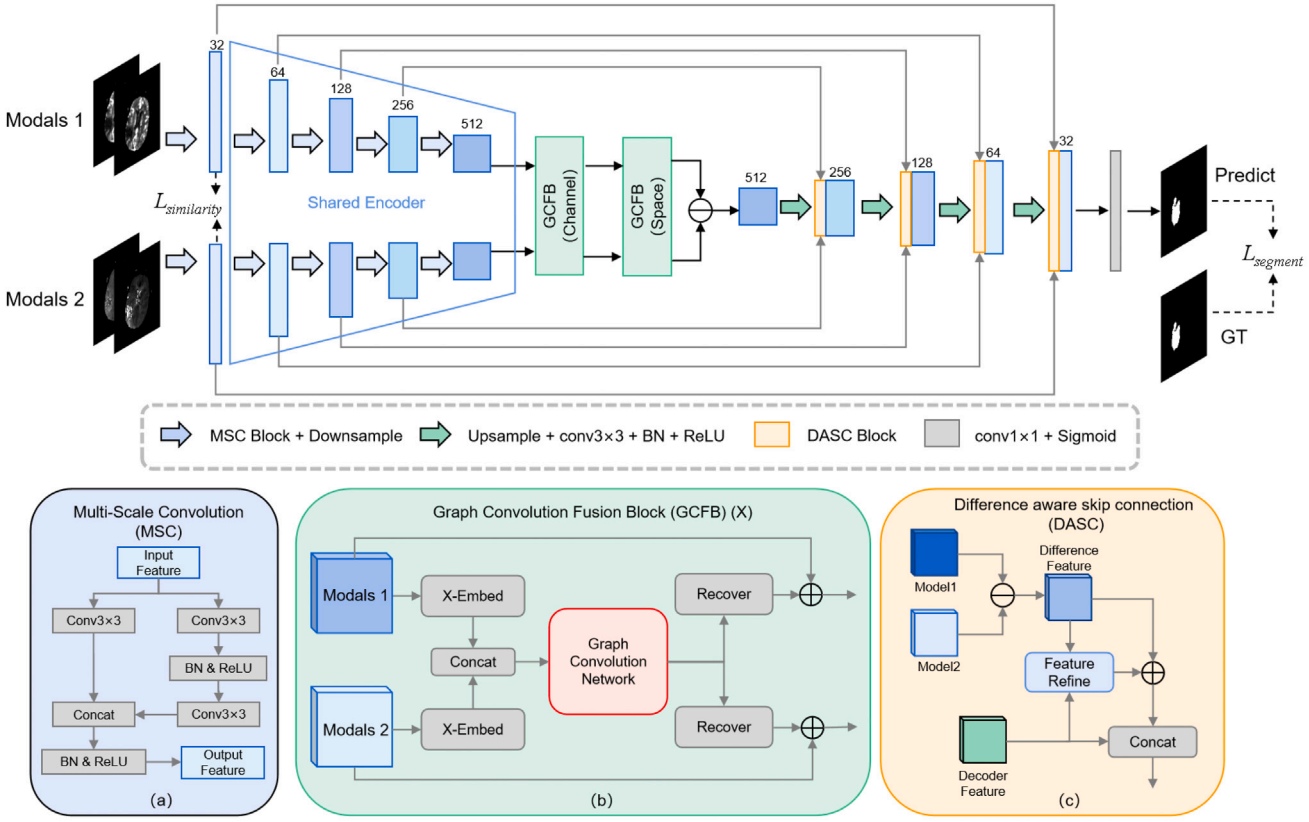


Fig. 3. The main frame of the MDANet. (a) Multi-Scale Convolution (MSC). (b) Graph Convolution Fusion Block (GCFB). (c) Difference Aware Skip Connection.

each modality. For brain CT perfusion maps, Chen et al. [9] used different encoders for each modal input, concatenated and fused the multimodal features in each layer of skip connections. Shi et al. [31] separated the four modalities into blood parameter maps and time parameter maps based on the imaging principle of the perfusion maps, and fused multimodal features through a well-designed group attention block. Kumar et al. [32] used multimodal images to strip skull areas in the original CT image and proposed a patch-based asymmetric U-Net framework to deal with the class imbalance in stroke segmentation.

2.3. Graph based modeling

Graph convolution was introduced by Bruna et al. [33] for the first time. Graph Convolutional Network (GCN) is mainly aimed at relation reasoning based on the nodes and edges in the classic graph data structure. Chen et al. [34] proposed a graph-based global reasoning module for image processing tasks, which maps image features from coordinate space to interaction space and processes global information through GCN. Based on the idea of converting feature maps into graphs, Lu et al. [35] applied graph convolutional networks to image semantic segmentation for the first time. Li et al. [36] introduced the graph reasoning into the original images of various scales to realize a light-weight segmentation based on GCN. Some studies [37,38] used graph convolution to establish the relationship between different images and improve the robustness of the network for segmenting targets of different sizes and textures. In the medical image segmentation tasks, Zhu et al. [29] implemented GCN to fuse semantic and edge features encoded from different modalities. These methods demonstrated the effectiveness of the graph reasoning modules.

3. Method

3.1. Overview

The main architecture of the proposed MDANet is shown in Fig. 3. Our network consists of a difference aware network with a newly designed skip connection (DASC) and a two-stage graph convolution fusion block (GCFB) for the interaction of multimodal features. The input multimodal medical images are categorized into two groups based on the properties of the lesions. Specifically, the grouping criteria are determined by whether the intensity of the lesion area is higher or lower than the surrounding pixels in each modality. The DASC at each layer extracts difference feature maps between features of different modality groups and refines them by interacting with difference feature maps from deeper semantic features. GCFB serves as the bottleneck between the encoder and decoder to fuse features and model correlations between modalities based on graph convolution. GCFB consists of two sub-blocks: GCFB(channel) and GCFB(space), which enables the module to encode the feature from global and local perspectives. The final segmentation result is output by the decoder.

3.2. Difference aware module

To fully leverage complementary information of multimodal inputs, we adopt a dual-path encoder. This architecture enables the network to better encode the identical features from each modality. The input multimodal images are divided into two groups based on their visual characters. Considering the differing gray distributions in different modal images may cause the network to fail to decouple the multimodal feature input. We first encode each group of inputs with independent parameters, and supervise the background area's features of different modal images as similar as possible. More details of this supervision will be described in Section 3.4. After the initial encoding stage, the

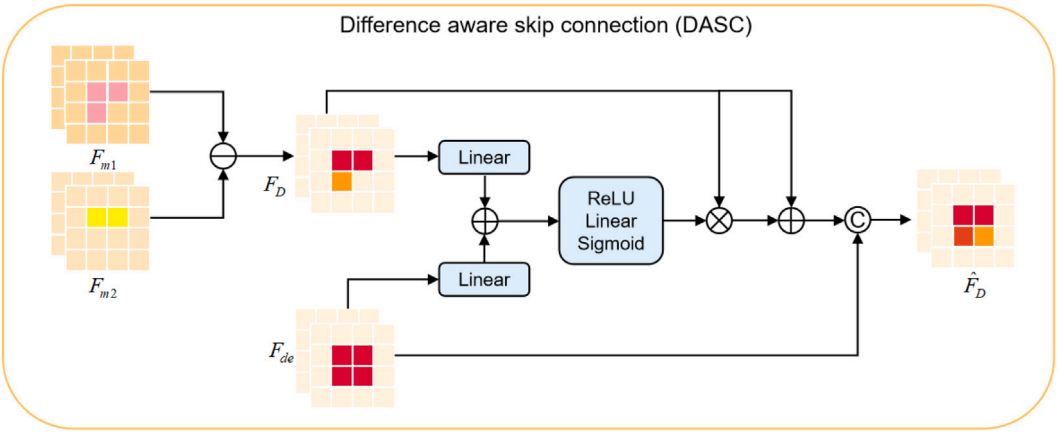


Fig. 4. The architecture of the Difference Aware Skip Connection (DASC).

features of different modalities are mapped to the same semantic space through a parameter-sharing encoder for subsequent interactive processing.

Lesions in medical images are diverse in shape and size. To better deal with this situation, we introduce a multi-scale convolution (MSC) encoder block to replace the original block in the U-Net encoder. The structure of MSC is shown in Fig. 3(a), We use different numbers of 3×3 convolution layers to encode features under different receptive fields, and finally concatenate them to obtain the multi-scale feature. The newly designed block can capture features of different reception fields with fewer parameters, which enables the network to recognize the potential lesion regions more effectively.

For the skip connection between the encoder and decoder, we propose Difference aware skip connection (DASC). As shown in Fig. 4, compared with the standard encoder–decoder architecture that directly passes through skip connections. The transferred features between the encoder and decoder are the difference feature maps F_D based on the mathematical difference between the two modality groups as:

$$F_D = F_{m1} - F_{m2} \tag{1}$$

where F_{m1}, F_{m2} refers to the features of different modality groups. The difference feature map F_D enhances the mismatch features between modalities and weakens the features of irrelevant areas. To further improve the module's awareness of potential lesion areas, we refine the difference feature map through the guidance of deeper semantic features. In DASC, we embed difference feature maps from the encoder and the decoder into an interaction space. Then we fuse the features and generate an attention map, the process can be expressed as:

$$W = \sigma(\theta_1(\theta_2(F_D) + \theta_3(F_{de}))) \tag{2}$$

where F_{de} refers to the difference feature map from the decoder. $\sigma()$ represents the sigmoid operator and $\theta_i()$ represents point-wise convolution operator. We use the weight map to update the difference features and obtain the refined feature through a residual connection with the original difference features. The whole process can be described as:

$$\hat{F_D} = F_D + W \times F_D \tag{3}$$

3.3. Graph convolution fusion block

Global information reasoning plays a crucial role in medical image segmentation. The transformer architecture is used widely in various computer vision tasks due to its global modeling capabilities. However, this architecture does not always perform well on many small datasets due to its lack of inductive bias, which always happens in medical image tasks [5,19]. In recent years, the combination of GCN and computer vision has been demonstrated to have good feasibility in terms of global reasoning [29,34]. In this work, we design GCFB to model and fuse features based on graph convolution.

In GCFB, the feature map $F_i \in R^{C\times H\times W}$ from coordinate space will be embedded into a graph $G_i \in R^{N\times E}$ in the interact space at first, where i denote to the modal groups, N represents the number of nodes in the graph and E represents the embedding dimension for each node. By concatenating the graphs generated by the feature map of each modality group, we can obtain a larger graph G_{concat} with nodes from all modalities in an interaction space.

When converting 3D feature maps into 2D graphs, taking inspiration from existing methods that incorporate attention from both channel and space to enhance feature encoding ability [39,40]. We design two different embedding strategies: channel embedding and space embedding.

Channel embedding: As shown in Fig. 5. Given the feature map $F_i \in R^{C\times H\times W}$, we utilize two learnable projection layers to reduce the dimension to obtain $F_\theta \in R^{N\times H\times W}$ and $F_\varphi \in R^{E\times H\times W}$, where N is the number of nodes and E is the feature length of each node. Subsequently, we flatten the feature maps and do matrix multiplication of F_θ and F_φ to get the graph $G_i \in R^{N\times E}$. The process can be expressed as:

$$G_i = \theta(F_i) \otimes (\varphi(F_i))^T \tag{4}$$

where $\theta(\cdot)$, $\varphi(\cdot)$ denotes the learnable projection layer, the feature is obtained by fusing features in the coordinate space from feature maps through dense sampling, which has a robust representation of global information. After updating the graph and getting graph $\hat{G}$, we can recover the feature from interaction space to coordinate space using:

$$\hat{F_i} = \hat{G_i} \times F_\theta \tag{5}$$

which has been proven to be effective in [34].

Space embedding: As shown in Fig. 6. Given the feature map $F_i \in R^{N\times H\times W}$, referring to the tokenization operation in ViT [20], we divide it into N patches, where $N = HW$ refers to the number of patches, each patch can be considered as a node in a graph. We also add a learnable tensor as position embedding to strengthen local information for each node. For space embedding, features can be converted from the interaction space back to the coordinate space by rearranging the patches.

By combining channel embedding and space embedding, the network can better represent information from global and local contexts. We can update and aggregate features by learning the internal features of each node and the relationships with other nodes based on graph convolution. The formula of graph convolution can be expressed as:

$$\hat{G} = ((I - A)G_{concat})W \tag{6}$$

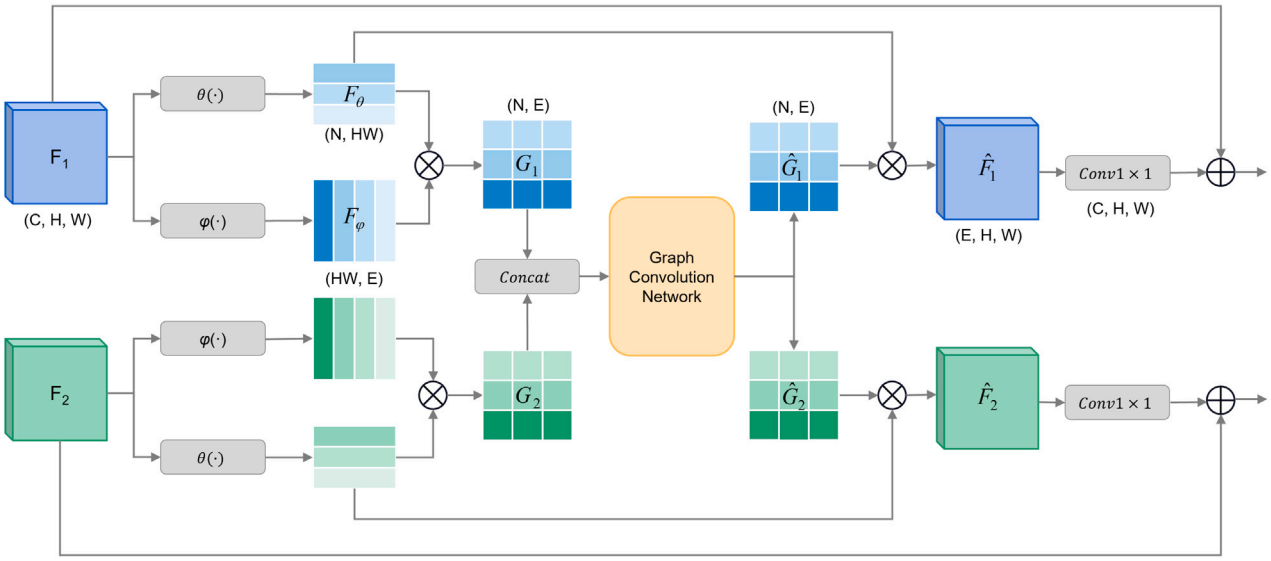


Fig. 5. GCFB with channel embedding.

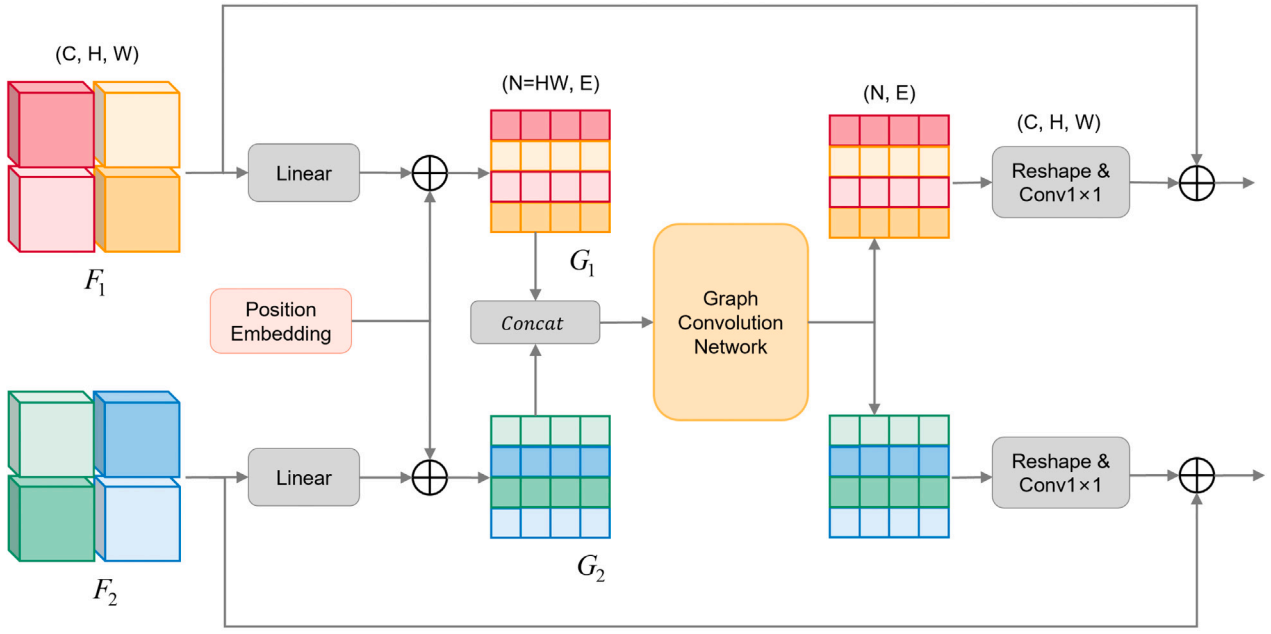


Fig. 6. GCFB with space embedding.

where I represents the identity matrix, $A \in R^{2N\times 2N}$ represents the adjacency matrix of this graph and $W \in R^{E\times E}$ represents the weight matrix, both A and W are learnable parameters. The adjacency matrix A serves the purpose of aggregating features of different nodes along the node dimension. It learns the edge weights, or relationships in other words between each node. By adjusting weights in the adjacency matrix, the network can assign importance to each edge, aggregating the features of different nodes more effectively. The weight matrix W aggregates the features of each node along the feature dimension of the graph and further learns the deep semantic feature of each node. The graph convolution is implemented by two 1D convolutions as shown in Fig. 7. Or expressed as:

$$GCN(G) = ReLU(Conv((G - Conv(G))^T)^T) \tag{7}$$

where $G = G_{concat}$ denotes the input graph.

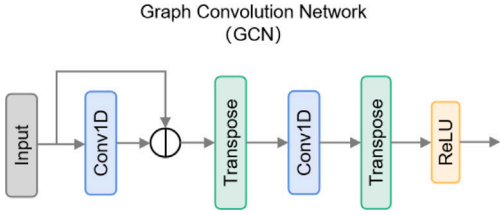


Fig. 7. Structure of graph convolution network (GCN).

After applying two GCFB sub-modules, we subtract the interacted multimodal feature maps to get the difference feature map as the input for the decoder.

3.4. Loss function

We hope the network can focus on the differences between multi-modal inputs, but the overall gray distribution of different modalities may cause interference during the direct subtraction process. To mitigate this interference, we design a masked similarity loss to supervise the features before they enter the parameter-sharing encoder. Specifically, for the feature maps output from the initial independent encode layer, we mask the lesion area in the original brain slice. The feature map is transformed into a vector through global average pooling. By supervising the cosine similarity between the processed vectors, features of non-lesion regions of different modalities are forced to close to each other. The cosine similarity loss can be expressed as:

$$L_{similarity} = 1 - \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{8}$$

where A and B represent the feature vectors of different modalities. This loss will be used as part of the overall loss of the network and provides stability and convenience during training. The overall loss of the network can be expressed as:

$$L = L_{segment} + \lambda L_{similarity} \tag{9}$$

where λ is a hyperparameters that control the trade-off between two losses. The segment loss function is a combination of traditional cross entropy loss function and dice loss function:

$$L_{segment} = L_{dice} + L_{ce} \tag{10}$$

$$L_{dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{11}$$

$$L_{ce} = -\sum_{i=1}^{N} y^i \log x^i \tag{12}$$

where X, x represents the predict results and Y, y represents the ground truth.

4. Experiment results and analysis

4.1. Datasets

We evaluate MDANet on Ischemic Stroke Lesion Segmentation (ISLES) 2018 dataset [41,42] and 2022 dataset [43].

ISLES2018: ISLES2018: This dataset includes 94 cases of CT perfusion maps from 63 acute stroke patients, each case consists of four modalities, namely CBF, CBV, MTT, and TMax.

ISLES2022: This dataset consists of 250 multimodal MRI scans. Each scan includes diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) and fluid attenuated inversion recovery (FLAIR) sequences. The DWI and ADC images are aligned, but FLAIR images require additional registration processing. Therefore, we only use DWI and ADC as input in this paper.

4.2. Implementation details

We use 5-fold cross-validation to evaluate the effectiveness of all models. The network is implemented on NVIDIA GeForce RTX 2080Ti GPU using PyTorch [44] framework. For the ISLES2018 dataset, we allocated CBF and CBV into one group (the pixel intensity of the lesion area is lower than the surrounding value). MTT and TMax (the pixel intensity of the lesion area is higher than the surrounding value) to another. For the ISLES2022 dataset, we use the DWI and ADC images as two group inputs, respectively. We also implement data augmentation operations for both datasets, including random flipping, random rotation (−90°~90°), and random scaling (0.8~1.2) to improve the robustness of the model.

All models are trained with a batch size of 16. The Adam optimizer is selected with an initial learning rate of 10^{-3}, β_1 of 0.9 and β_2 of 0.999. We evaluate the trained networks using the Dice Score (DSC), precision and recall rate. These indicators can be calculated as follows:

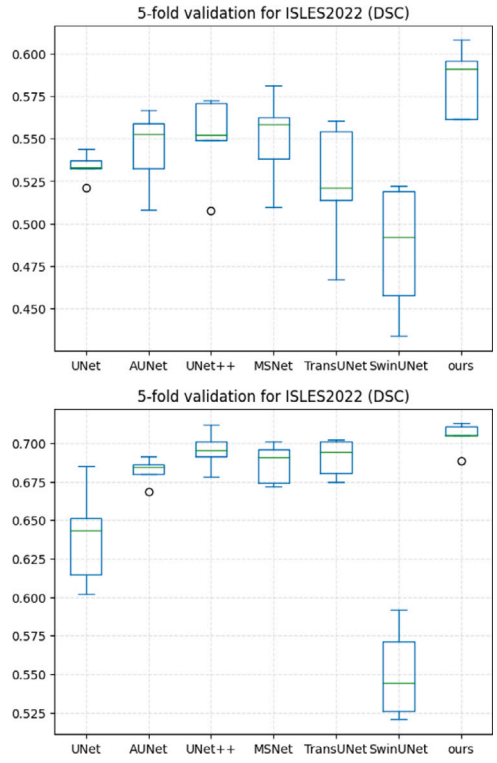


Fig. 8. Boxplot of 5-fold cross-validation dice score.

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F1\ Score = \frac{Precision \times Recall \times 2}{Precision + Recall} \tag{16}$$

where TP represents the prediction of true positives, FP represents the prediction of false positives and FN represents the prediction of false negatives on a single pixel.

4.3. Quantitative results

Tables 1 and 2 show the comparison experiment of MDANet with other classic and effective segment methods on ISLES2018 and 2022 datasets, including classical convolution-based method of U-Net [13], Attention U-Net [14], U-Net++ [15], MSNet [18] and advanced transformer-based methods of TransUNet [6], SwinUNet [22]. All experimental results are based on five-fold cross-validation and performed under the same configuration and computer environment. For the ISLES2018 datasets, we additionally compare several state-of-the-art stroke segmentation methods based on CT perfusion images, including OctopusNet [9], the method proposed by Abulnaga [45], Pool-UNet [25] and Perf-UNet [26].

The experimental result of MDANet is presented at the bottom of the tables, and the best result is highlighted in bold. We can observe that MDANet outperforms other advanced methods on both datasets. Specifically, MDANet achieved a Dice Score of 58.34% on ISLES2018, which was 3.90% higher than the suboptimal result in the same experimental environment obtained by U-Net++ [15], and was 0.42% higher than the suboptimal result of all compared methods. MDANet also achieved 70.44% on ISLES2022, which is 0.87% higher than the suboptimal

Table 1
Comparison of different models on the ISLES2018 dataset. Scores are obtained by 5-fold cross-validation. The best results are in boldface.

Method	DSC (%)	Precision (%)	Recall (%)	F1 Score	Param (M)	FLOPs (G)
U-Net [13]	53.34 ± 0.85	57.09 ± 3.99	56.90 ± 3.40	0.5680	30.0	10.81
Attention U-Net [14]	54.37 ± 2.36	54.61 ± 4.32	62.41 ± 2.23	0.5817	29.9	9.92
U-Net++ [15]	55.04 ± 2.61	55.96 ± 5.03	62.47 ± 3.06	0.5886	35.0	26.69
MSNet [18]	54.99 ± 2.72	55.23 ± 4.43	63.36 ± 1.96	0.5891	20.3	6.96
TransUNet [6]	52.34 ± 3.73	53.42 ± 6.64	59.05 ± 5.11	0.5571	175.3	19.19
SwinUnet [22]	48.50 ± 3.85	51.42 ± 5.01	55.25 ± 8.27	0.5275	106.0	6.14
OctopusNet [9]	57.92	–	–	–	–	–
Abulnaga et.al [45]	54.00	–	–	–	–	–
Pool-UNet [25]	56.04	**67.82**	56.54	0.6167	–	–
Perf-UNet [26]	56.40	56.50	**64.40**	0.6019	–	–
MDANet	**58.34 ± 2.11**	60.22 ± 3.69	64.03 ± 2.89	**0.6195**	25.3	11.22

Table 2
Comparison of different models on the ISLES2022 dataset. Scores are obtained by 5-fold cross-validation. The best results are in boldface.

Method	DSC (%)	Precision (%)	Recall (%)	F1 Score	Param	FLOPs (G)
U-Net [13]	63.93 ± 3.26	74.55 ± 1.98	62.11 ± 5.07	0.6764	30.0	2.69
Attention U-Net [14]	68.20 ± 0.86	72.95 ± 2.47	70.65 ± 2.01	0.7173	29.9	2.47
U-Net++ [15]	69.57 ± 1.24	74.94 ± 1.19	70.91 ± 1.53	0.7286	35.0	6.66
MSNet [18]	68.66 ± 1.30	74.59 ± 2.65	69.42 ± 2.69	0.7184	20.3	1.73
TransUNet [6]	69.75 ± 1.23	75.03 ± 2.11	69.80 ± 1.90	0.7228	175.3	4.61
Swin-Unet [22]	55.08 ± 3.02	56.36 ± 4.00	65.40 ± 3.69	0.6039	105.4	6.10
MDANet	**70.44 ± 0.97**	**75.30 ± 2.71**	**72.22 ± 1.63**	**0.7368**	25.3	3.57

result of TransUNet [6]. Fig. 8 compares the scores of different methods on the dice coefficient under cross-validation more intuitively through the box plot.

In terms of precision and recall, MDANet achieves the highest scores of 75.30% and 72.22% on the ISLES2022 datasets. It outperforms the second-ranked method by 0.27% and 1.31% respectively. On the ISLES2018 datasets, although MDANet does not achieve the optimal precision and recall rate, Pool-Unet [25] achieved a precision rate of 67.82%, but a recall rate of 56.54%. And Perf-UNet [26] achieved a recall rate of 64.40%, but an accuracy rate of 56.50%. These results show a trade-off of these two indicators due to the calculation methods. In contrast, MDANet achieved 60.22% and 64.03% in precision and recall rates. It is a relatively balanced and high-level result, which can also be concluded by comparing the F1 scores. According to the quantitative results, our proposed MDANet improves the performance of stroke segmentation with a small number of parameters and a tiny cost of computational complexity.

4.4. Qualitative results

Figs. 9 and 10 shows some example of segmentation results from different methods on ISLES2018 and ISLES2022. The first column shows one of the representative modalities in the multi-modal inputs, and the second column shows the ground truth of the corresponding inputs, the last column is the segmentation result of the MDANet.

To enhance the visualization of segmentation results, we employ different colors to represent the predictions of true positive (TP), false positive (FP), and false negative (FN). For the vast majority of cases, relying on the feature representation capabilities of deep learning, most methods can effectively segment lesions. However, when paying more attention to the details, it can be observed that the proposed MDANet has better performance. Notably, MDANet demonstrates better performance by significantly reducing the number of false negative predictions (yellow colored region). This improvement is also reflected in the higher recall rate. We believe that this point is of great significance in clinical diagnosis, because the missed diagnosis of lesion areas can lead to patients missing the optimal treatment window and causing more potential harm to their health.

4.5. Ablation study

4.5.1. Components ablation

To further verify the effectiveness of the main components in the MDANet, an ablation study is conducted on the ISELS2018 datasets. U-Net serves as the baseline in this experiment. By adding or replacing proposed components sequentially to the original network, we mainly test the following components:

DASC: Replacing the single encoder with a dual encoder that shares parameters except for the first stage, and using difference aware skip connection (DASC) to replace the original concatenate-based skip connection.

MSC: Replacing the normal convolution layer with the proposed multi-scale convolution layer.

GCFB: Introducing the graph convolution fusion block into the network as a neck between encoder and decoder.

SL: Adding a similarity loss function as a soft supervision for feature aligning between different modalities.

Table 3 presents an overview of the objective performance of different components and their contributions to the final segmentation performance of the network. Each module brings improvement to the segmentation performance. Specifically, the DASC module has the most noticeable effect on improving the dice score. The GCFB module contributes the most to enhancing the precision rate. This is likely because GCFB enables the network to have a better global modeling ability through global reasoning based on graph convolution. By dividing the feature map into different nodes along the channel and space, GCFB allows the model to perform global reasoning and build global relationships between local features, thus improving the network's global modeling ability. Furthermore, both DASC and MSC contribute significantly to increasing the recall rate. DASC can effectively capture mismatches between different modalities and aid in the determination of lesion regions. On the other hand, MSC provides a broader receptive field for the network. These contributions further help MDANet concentrate on lesion areas and suppress false negative predictions.

4.5.2. Skip connection ablation

In MDANet, we use feature subtraction as the pipeline for multi-modal difference awareness and utilizing the different maps through skip connection.

To further investigate the effectiveness of the subtraction operation in the skip connection, a comparative experiment is conducted to

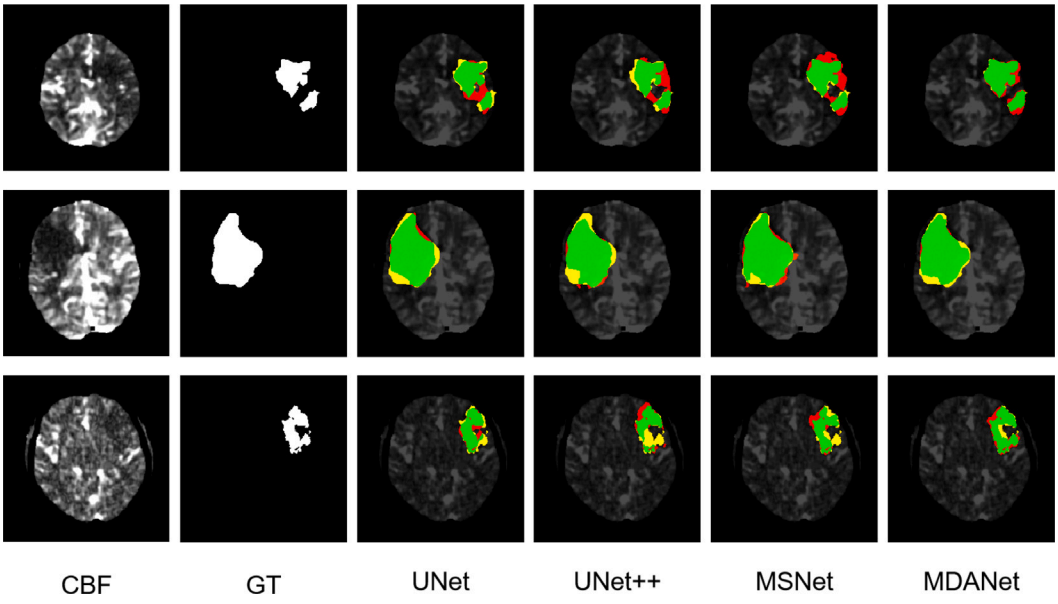


Fig. 9. Qualitative results of stroke segmentation on ISLES2018 datasets, with the corresponding true positive predicts (green), false positive predicts (red) and false negative predicts (yellow).

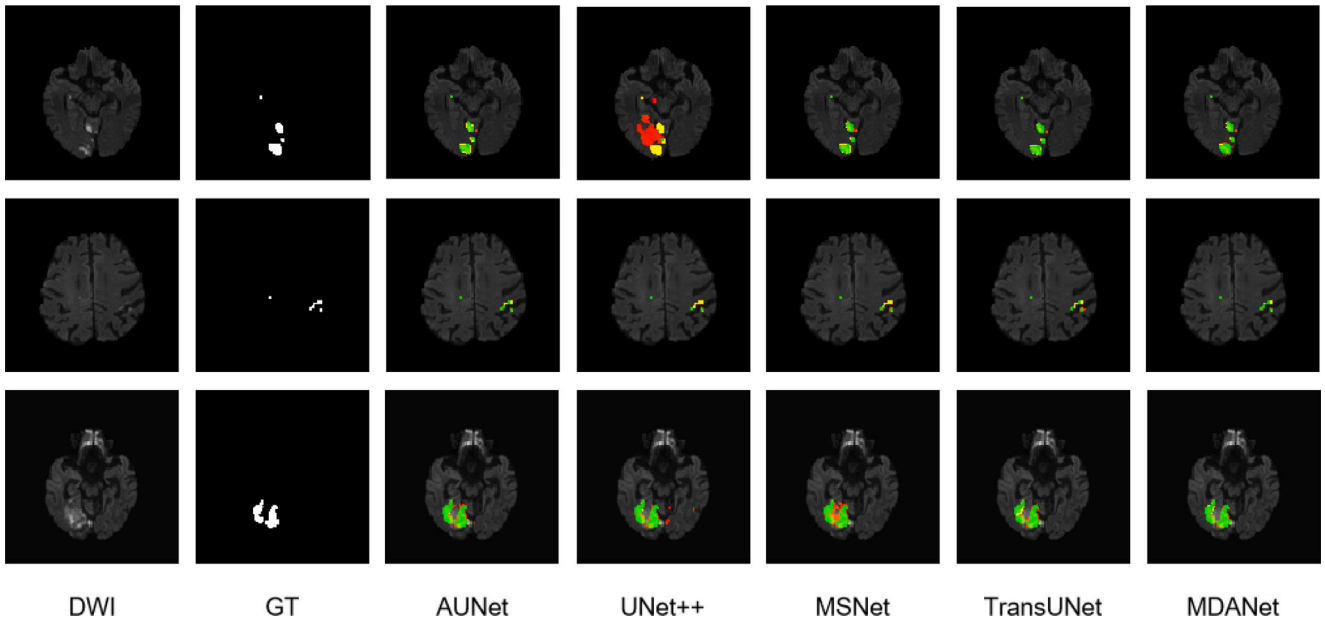


Fig. 10. Qualitative results of stroke segmentation on ISLES2022 datasets, with the corresponding true positive predicts (green), false positive predicts (red) and false negative predicts (yellow).

Table 3
Ablation study on the ISLES2018 dataset. Scores are obtained by 5-fold cross-validation. The best results are in boldface.

DASC	MSC	GCFB	SL	DSC (%)	Precision (%)	Recall (%)	F1 score
				53.34 ± 0.85	57.09 ± 3.99	56.90 ± 3.40	0.5680
✓				55.31 ± 3.36	57.53 ± 4.81	59.89 ± 3.64	0.5869
✓	✓			56.64 ± 1.99	57.39 ± 2.83	62.57 ± 5.40	0.5987
✓	✓	✓		57.50 ± 2.32	60.19 ± 4.77	59.54 ± 2.96	0.5986
✓	✓	✓	✓	**58.34 ± 2.11**	**60.22 ± 3.69**	**64.03 ± 2.89**	**0.6195**

Table 4
Comparison of different skip connection strategies on the ISLES2018 dataset. Scores are obtained by 5-fold cross-validation. The best results are in boldface.

Method	DSC (%)	Precision (%)	Recall (%)
Add	56.22 ± 1.27	58.32 ± 2.71	60.06 ± 0.50
Concat	57.72 ± 2.90	60.01 ± 2.27	59.71 ± 5.87
Subtract	**58.34 ± 2.11**	**60.22 ± 3.69**	**64.03 ± 2.89**

Table 5
Comparison of global fusion and modeling strategies on the ISLES2018 dataset. Scores are obtained by 5-fold cross-validation. The best results are in boldface.

Method	DSC (%)	Param (M)	FLOPs(G)
Baseline	56.32 ± 1.87	22.5	10.88
Non-local	56.98 ± 1.60	27.6	11.22
Transformer	57.14 ± 1.48	61.7	15.83
ours	**58.34 ± 2.11**	25.3	11.22

compare different fusion strategies. The result is shown in Table 4, in which the Subtract indicates the operation implemented in the proposed MDANet. Add means add feature maps of different modalities. Concat means concatenating feature maps of various modalities and applying another 1×1 convolution to adjust the number of channels. When using the addition strategy, there is a sharp drop in the model's performance. One possible explanation for this is that due to the opposite values in pixels of the lesion area across different modalities, the activation value of the lesion area is weakened after addition. On the other hand, the concatenation strategy has a relatively small impact on the model's performance. However, it requires additional computing resources to reduce the dimension of the feature map. Therefore, the subtraction operation in the skip connection is beneficial for differences awareness between multimodal features. It improves the segmentation performance by effectively capturing and emphasizing the discrepancies in the lesion area.

4.5.3. Global modeling ablation

The graph convolution fusion block (GCFB) in MDANet plays a crucial role in fusing multi-modal features and global modeling. The encoding of global features is vital for medical image segmentation. Apart from the global reasoning based on graph convolution employed in MDANet. the transformer architecture [20] and non-local block [46] also can model the global features of images.

In this section, we conduct a comparison of the impact of different advanced global modeling methods on multi-modal image fusion and the final segmentation performance of our model. Since we implemented two layers of GCFB(channel) and GCFB(space) in the proposed method, we keep the same number of layers for the non-local and transformer blocks. In the first layer, feature maps of different modality groups perform self-attention operations. In this process, the query, key and value vectors are derived from the feature maps themselves. In the second layer, key–value pair vectors and query vectors are obtained from different modal groups, and the interaction between modalities is achieved through cross-attention. Finally, we subtract the updated features to acquire the difference feature map as module output. The result of different global modeling methods is shown in Table 5. Our graph convolution fusion block achieves the best performance in the comparative experiments. Furthermore, compared with the transformer-based global fusion method, the graph convolution-based method has fewer parameters, making it more efficient in terms of computational resources and memory requirements.

4.6. Difference aware analysis

To better illustrate the effect of feature map subtraction in DASC, we visualize the attention maps of feature output from DASC in the third layer of MDANet. We compare these attention maps with the output from the standard attention gate. The result is shown in Fig. 11.

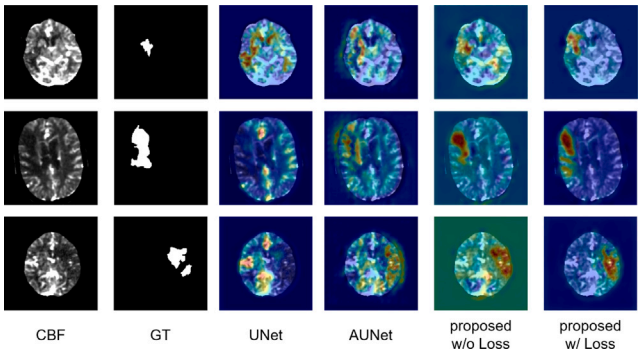


Fig. 11. Examples of the attention maps in baseline and MDANet on ISLES2018 dataset.

Table 6
Experimental results of different weight λ on ISLES2018 dataset.

λ	DSC (%)	Precision (%)	Recall (%)	F1 Score
0	57.50 ± 2.32	60.19 ± 4.77	59.54 ± 2.96	0.5986
0.25	58.19 ± 2.38	**61.67 ± 4.80**	61.16 ± 2.88	0.6142
0.5	58.11 ± 1.87	61.31 ± 2.42	60.80 ± 3.54	0.6095
0.75	58.25 ± 2.21	61.54 ± 3.96	63.20 ± 3.36	**0.6236**
1	**58.34 ± 2.11**	60.22 ± 3.69	**64.03 ± 2.89**	0.6195

The feature maps in the U-Net model without any attention modules are highly relevant to the pixel values in the input image. Consequently, regions with high pixel values in the original image tend to be preserved in the features. However, some of these features have a low correlation with target areas and introduce noise to the deep semantic through direct skip connection.

Attention U-Net utilizes the attention gate to replace the direct concatenation in the skip connection, which enables the network to locate the possible lesion area by reducing the semantic gap between the encoder and decoder. However, Attention U-Net still has deficiencies in accuracy and concentration that suggest the network's attention is not optimal enough.

By introducing the difference aware skip connection (DASC), it is evident that the network can effectively focus more attention on the lesion area. Shows that MDANet can indeed locate the lesion through the difference between various modalities (shown in the 5th column, Fig. 11). However, the process of feature map subtraction can introduce other noise due to variations in pixel's mean and variance across different modal images. Although most of the noise can be suppressed through subsequent attention operations, there may still be residual noise present, which can be manifested as high signals in the non-lesion areas.

The proposed similarity loss further mitigates the impact of residual noise by encouraging the network to align features in the non-lesion areas between modalities and better highlight the lesion area that really needs to be detected (shown in the 6th column, Fig. 11). Cases in ISLES2022 dataset (Fig. 12) also demonstrate that difference feature maps can better locate possible locations of lesions than traditional attention.

We further conduct experiments to evaluate the selection of λ in the total loss calculation. The quantitative results are shown in Table 6. When $\lambda = 0$, the model achieves lower performance on dice score due to lack of alignment of features. In this paper, we use $\lambda = 1$ in all experiments because it achieves the best performance on dice score.

Fig. 13 further explains how difference aware module works. The feature maps from different modal groups exhibit opposite pixel intensities in the lesion area. By subtracting the feature maps, the highlighted areas in the blood parameter features (CBF, CBV) that are not related to the lesion area are suppressed. On the other hand, the highlighted regions of interest in temporal parameter features (MTT, TMax), which

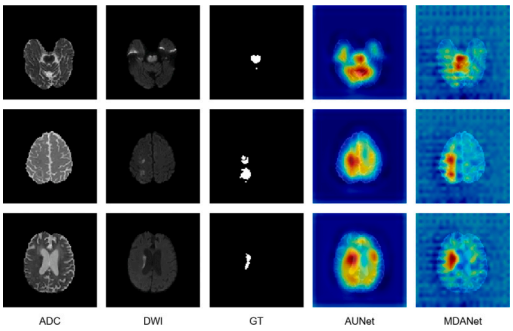


Fig. 12. Examples of the attention maps on ISLES2022 dataset.

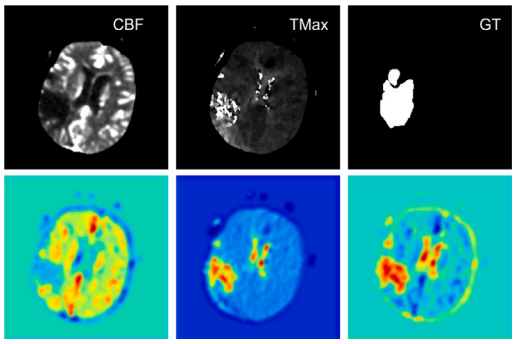


Fig. 13. Through difference aware skip connection, the lesion area is further highlighted, and the unrelated area is suppressed.

represents the possible lesions are further enhanced and supplemented. These help the network reduce the influence of irrelevant information and focus attention on the lesion areas.

5. Conclusion

In this study, we propose MDANet, a novel multimodal brain stroke segmentation framework. Our approach is motivated by observing lesion characteristics of stroke images from different modalities. We design the difference aware module to extract the mismatched features between modalities and transfer the obtained difference feature maps through a difference aware skip connection. We further introduce a similarity loss function to align features in the healthy area to mitigate noise caused by grayscale differences in multimodal images, improving the model's concentration of the potential lesions. For the multimodal feature fusion, we propose a graph-based fusion block. Two distinct graph embedding strategies are developed to build the graph by modeling the features from channel and space perspectives, and the interaction between modalities is realized based on graph convolution. Our MDANet is evaluated on the ISLES2018 and ISLES2022 datasets. The experimental results demonstrate that MDANet surpasses many existing classic methods, and the proposed components have a positive impact on the segmentation outcomes. However, MDANet still has some limitations. Our method lacks supervision of the lesion edges, which may lead to the model being insensitive to the boundary of the lesion area. The introduction of dual-encoder and graph convolution module may also bring additional computational complexity to affect the efficiency of the network. In the future work, we aim to solve the above problems by further exploring the utilization of features from different modalities, investigating and developing more effective and efficient fusion strategies for multimodal inputs. Additionally, we plan to extend the proposed method to tackle other multimodal medical segmentation tasks.

CRediT authorship contribution statement

Kezhi Zhang: Writing – original draft, Validation, Software, Methodology, Conceptualization. **Yu Zhu:** Writing – review & editing, Supervision, Project administration. **Hangyu Li:** Visualization, Formal analysis. **Zeyan Zeng:** Investigation, Data curation. **Yatong Liu:** Formal analysis, Data curation. **Yuhao Zhang:** Investigation.

Declaration of competing interest

All 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

Data will be made available on request.

Acknowledgment

This work was supported by the Exploratory Device R&D Projects of National Clinical Research Center for Interventional Medicine, China (NO. 2021-002) and the Key Clinical Research Projects of National Clinical Research Center for Interventional Medicine, China (NO. 2019-003).

References

[1] Valery L. Feigin, Benjamin A. Stark, Catherine Owens Johnson, Gregory A. Roth, Catherine Bisignano, Gdiom Gebreheat Abady, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Foad Abd-Allah, Vida Abedi, et al., Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019, Lancet Neurol. 20 (10) (2021) 795–820.

[2] Mario Mascalchi, Massimo Filippi, Roberto Floris, Claudio Fonda, Roberto Gasparotti, Natale Villari, Diffusion-weighted MR of the brain: Methodology and clinical application, Radiol. Medica 109 (3) (2005) 155—197, URL http://europepmc.org/abstract/MED/15775887.

[3] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[4] Yongjin Zhou, Weijian Huang, Pei Dong, Yong Xia, Shanshan Wang, D-UNet: A dimension-fusion u shape network for chronic stroke lesion segmentation, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (3) (2019) 940–950.

[5] Pooya Ashtari, Diana M. Sima, Lieven De Lathauwer, Dominique Sappey-Marinier, Frederik Maes, Sabine Van Huffel, Factorizer: A scalable interpretable approach to context modeling for medical image segmentation, Med. Image Anal. 84 (2023) 102706.

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, Yuyin Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

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

[8] Yatong Liu, Yu Zhu, Ying Xin, Yanan Zhang, Dawei Yang, Tao Xu, MESTrans: Multi-scale embedding spatial transformer for medical image segmentation, Comput. Methods Programs Biomed. 233 (2023) 107493.

[9] Yu Chen, Jiawei Chen, Dong Wei, Yuexiang Li, Yefeng Zheng, OctopusNet: A deep learning segmentation network for multi-modal medical images, in: Multiscale Multimodal Medical Imaging: First International Workshop, MMMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 1, Springer, 2020, pp. 17–25.

[10] Yuhang Ding, Xin Yu, Yi Yang, RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3975–3984.

[11] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, Yefeng Zheng, mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 107–117.

[12] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[13] Olaf Ronneberger, Philipp Fischer, 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, Springer, 2015, pp. 234–241.

[14] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., Attention u-net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.

[15] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, 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, Springer, 2018, pp. 3–11.

[16] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Jian Wu, Unet3+: A full-scale connected unet for medical image segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 1055–1059.

[17] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, Chen Wu, ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data, ISPRS J. Photogramm. Remote Sens. 162 (2020) 94–114.

[18] Xiaoqi Zhao, Lihe Zhang, Huchuan Lu, Automatic polyp segmentation via multi-scale subtraction network, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 120–130.

[19] Xiaoqi Zhao, Hongpeng Jia, Youwei Pang, Long Lv, Feng Tian, Lihe Zhang, Weibing Sun, Huchuan Lu, M^2 SNet: Multi-scale in multi-scale subtraction network for medical image segmentation, 2023, arXiv preprint arXiv:2303.10894.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[22] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[23] Mou-Cheng Xu, Neil P Oxtoby, Daniel C Alexander, Joseph Jacob, Learning to pay attention to mistakes, 2020, arXiv preprint arXiv:2007.15131.

[24] Albert Clèrigues, Sergi Valverde, Jose Bernal, Jordi Freixenet, Arnau Oliver, Xavier Lladó, Acute and sub-acute stroke lesion segmentation from multimodal MRI, Comput. Methods Programs Biomed. 194 (2020) 105521, http://dx.doi.org/10.1016/j.cmpb.2020.105521, URL https://www.sciencedirect.com/science/article/pii/S0169260719305899.

[25] Ruihua Liu, Wei Pu, Yangyang Zou, Linfeng Jiang, Zhiyong Ye, Pool-unet: Ischemic stroke segmentation from CT perfusion scans using poolformer unet, in: 2022 6th Asian Conference on Artificial Intelligence Technology, ACAIT, IEEE, 2022, pp. 1–6.

[26] Lucas de Vries, Bart J. Emmer, Charles B.L.M. Majoie, Henk A. Marquering, Efstratios Gavves, PerfU-Net: Baseline infarct estimation from CT perfusion source data for acute ischemic stroke, Med. Image Anal. 85 (2023) 102749.

[27] Mehmet Aygün, Yusuf Hüseyin Şahin, Gözde Ünal, Multi modal convolutional neural networks for brain tumor segmentation, 2018, arXiv preprint arXiv:1809.06191.

[28] Xuejian Li, Shiqiang Ma, Jijun Tang, Fei Guo, TranSiam: Fusing multimodal visual features using transformer for medical image segmentation, 2022, arXiv preprint arXiv:2204.12185.

[29] Zhiqin Zhu, Xianyu He, Guanqiu Qi, Yuanyuan Li, Baisen Cong, Yu Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Inf. Fusion 91 (2023) 376–387.

[30] Zdravko Marinov, Simon Reiß, David Kersting, Jens Kleesiek, Rainer Stiefelhagen, Mirror U-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging, 2023, arXiv preprint arXiv:2303.07126.

[31] Tianyu Shi, Huiyan Jiang, Bin Zheng, C2 MA-Net: Cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on CT perfusion scans, IEEE Trans. Biomed. Eng. 69 (1) (2021) 108–118.

[32] Amish Kumar, Palash Ghosal, Soumya Snigdha Kundu, Amritendu Mukherjee, Debashis Nandi, A lightweight asymmetric U-Net framework for acute ischemic stroke lesion segmentation in CT and CTP images, Comput. Methods Programs Biomed. 226 (2022) 107157, http://dx.doi.org/10.1016/j.cmpb.2022.107157, URL https://www.sciencedirect.com/science/article/pii/S0169260722005387.

[33] Joan Bruna, Wojciech Zaremba, Arthur Szlam, Yann LeCun, Spectral networks and locally connected networks on graphs, 2013, arXiv preprint arXiv:1312.6203.

[34] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, Yannis Kalantidis, Graph-based global reasoning networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 433–442.

[35] Yi Lu, Yaran Chen, Dongbin Zhao, Jianxin Chen, Graph-FCN for image semantic segmentation, in: International Symposium on Neural Networks, Springer, 2019, pp. 97–105.

[36] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, Hong Liu, Spatial pyramid based graph reasoning for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8950–8959.

[37] Guo-Sen Xie, Jie Liu, Huan Xiong, Ling Shao, Scale-aware graph neural network for few-shot semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5475–5484.

[38] Yanqi Bao, Kechen Song, Jie Liu, Yanyan Wang, Yunhui Yan, Han Yu, Xingjie Li, Triplet-graph reasoning network for few-shot metal generic surface defect segmentation, IEEE Trans. Instrum. Meas. 70 (2021) 1–11.

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[40] Yongheng Sun, Duwei Dai, Qianni Zhang, Yaqi Wang, Songhua Xu, Chunfeng Lian, MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation, Pattern Recognit. 139 (2023) 109524.

[41] Oskar Maier, Bjoern H Menze, Janina Von der Gablentz, Levin Häni, Mattias P. Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al., ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI, Med. Image Anal. 35 (2017) 250–269.

[42] Michael Kistler, Serena Bonaretti, Marcel Pfahrer, Roman Niklaus, Philippe Büchler, The virtual skeleton database: An open access repository for biomedical research and collaboration, J. Med. Internet Res. 15 (11) (2013) e245.

[43] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al., ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset, Sci. Data 9 (1) (2022) 762.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).

[45] S. Mazdak Abulnaga, Jonathan Rubin, Ischemic stroke lesion segmentation in CT perfusion scans using pyramid pooling and focal loss, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4, Springer, 2019, pp. 352–363.

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.