MD-TransUNet: TransUNet with Multi-attention and Dilated Convolution for Brain Stroke Lesion Segmentation

Jie Xu[1,2], Jian Wan[1,3], and Xin Zhang[1,2,4(✉)]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, 310018 Zhejiang, China
{211050059,zhangxin}@hdu.edu.cn, wanjian@zust.edu.cn
[2] Key Laboratory of Complex Systems Modeling and Simulation Ministry of Education, Ministry of Education, Beijing, China
[3] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Zhejiang, China
[4] Key Laboratory of Marine Ecosystem Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou, China

Abstract. The accurate segmentation of stroke lesion regions holds immense significance in shaping treatment strategies and rehabilitation protocols. Due to the large difference in the volume of stroke lesion areas and the great similarity between lesion areas and normal tissues, most of the existing methods for lesion segmentation cannot deal with these problems well. This paper proposes a novel network named MD-TransUNet for the segmentation of stroke lesions, whose framework is based on the UNet architecture. To fully obtain deep image features, it uses ResNet50 for downsampling. MD (multi-dilated) module is employed as the skip connection to gain more receptive fields. Different receptive fields can adapt to varying volumes of lesion areas. Then, a feature extraction module with multi-level attention mechanism is designed using ConvLSTM, non-local spatial attention, and channel attention modules to suppress useless information expression in skip connections and upsampling processes while focusing more on effective spatial and channel information in features. The experiments show that our proposed network gets superior performance than benchmark methods and indicates the generalization and effectiveness of the proposed model.

Keywords: Brain Stroke Lesion · Attention · Dilated Convolution

1 Introduction

According to recent global statistics, stroke is ranked second as a cause of mortality and third as a cause of both mortality and disability in 2019. In fact, the

H. Gao et al. (Eds.): CollaborateCom 2023, LNICST 562, pp. 151–170, 2024.
https://doi.org/10.1007/978-3-031-54528-3_9

total number of stroke cases recorded that year was 12.2 million, contributing to a massive 6.55 million deaths worldwide [1]. MRI (Magnetic Resonance Imaging) scans are useful in detecting ischemic strokes in the early stage. Additionally, MRI provides essential functional information, including cerebral blood flow and tissue metabolism, which is beneficial in the early diagnosis and differential diagnosis of strokes. Therefore, segmentation of the lesion areas from the MR images is particularly important.

Initially, before the rise of deep learning technology, Montiel et al. [2] calculated the edge confidence on DWI images and assigned each observation value to the point closest to the gradient direction along the observation point, grouping a set of observation values into different classes. Subsequently, all voxels that had converged to the same point were used to merge and mark the tissue region. Ozertem et al. [3] used kernel annealing to automatically segment brain lesion regions.

In recent years, with the development of deep learning methods, the architectures such as UNet [4], DeepLab3+ [5] and CLCI-Net [6] have been widely used in the field of image segmentation and have achieved great success. UNet uses an encoder-decoder symmetric architecture. This structure facilitates the extraction of more accurate information by fusing high-resolution features with their corresponding upsampled counterparts via skip connections. In the upsampling process, the network has more channels due to the combination of downsampling features, allowing the network to capture more feature information and achieve greater accuracy with less training data. DeepLab3+ incorporates a substantial ensemble of dilated convolution layers within its encoder module to augment the receptive field without compromising information loss. The utilization of dilated convolutions enables each convolutional layer to accumulate information across a wider spatial extent. By applying a specific dilation rate to each convolutional layer, the receptive field expands progressively, facilitating the integration of contextual information from a larger region. Consequently, this design empowers DeepLab3+ to capture long-range dependencies and exploit global contextual cues, leading to a more comprehensive understanding of the input data. Although UNet is efficient, easy to build, and has strong scalability, it has fewer layers in the encoder and cannot obtain deeper features. The deficiency in obtaining deeper features restricts the network's capability to comprehend intricate visual contexts and nuances present within the input data. Pure attention mechanisms and dilated convolution cannot compensate for this deficiency when applied to brain lesion segmentation.

In this paper, our model uses the TransUNet architecture [7] as the backbone. The model combines UNet and vision transformer, integrating both their advantages to extract improved global and local information that includes shallow and deep features. This fusion of methods allows for comprehensive analysis of the input data, empowering the model to effectively extract and incorporate both local and global visual cues. However, brain lesion segmentation is significantly different from brain structure segmentation. There are two key differences. One is that the size of brain lesions is difficult to predict and tends to vary widely. The other is that the proportion of positive and negative samples in the dataset

is extremely unbalanced. In order to enhance the receptive field and facilitate the integration of a greater volume of information into the upsampling process, we introduce the MD (multi-dilated) module within the skip connection pathway. By incorporating the MD module, we address the limitations of conventional skip connections, which may not adequately capture and incorporate expansive contextual information. Through the utilization of dilated convolutions within the MD module, we enable an increased receptive field, allowing for a more comprehensive understanding of the input data. As a result, the expanded receptive field enables the model to effectively exploit long-range dependencies and capture important contextual cues, enriching the feature representation and improving the accuracy of the upsampling process. Then, we integrate a FEM (Feature Extraction Module) in the upsampling process, which consists of two parts: GCCA (Global Context Channel Attention) and ConvLSTM (Long short-term memory). Given that each convolutional layer in the upsampling process combines downsampling features, resulting in an increased number of channels, we employ the GCCA to effectively mitigate redundant channels and prioritize the expression of pertinent spatial information. By leveraging the GCCA, we can enhance the discriminative power of feature representations by selectively attending to informative channels and suppressing irrelevant ones. By using ConvLSTM to learn long-term dependent information, it reduces the negative impact of excessive useless information generated during the encoder-to-decoder process on the prediction results. Our main contributions are summarized as follows:

- In this study, we introduce the MD-TransUNet model, which addresses the challenge of the network's tendency to overlook small lesion features caused by the significant difference in lesion region sizes, and the problem of the abundance of irrelevant information due to reuse features.
- We use the convolutional MD module to expand the receptive field so that the upsampling process can obtain as much downsampled feature information as possible.
- We propose a Feature Extraction module, which makes the feature restoration of the upsampling process more efficient, suppresses useless channels and increases the weight of useful information.
- We evaluate our network on an open-source dataset of stroke lesions. Extensive experiments are conducted to demonstrate the superiority of our method.

2 Related Work

Medical image segmentation can separate the pathological tissue structures or specific human organs that need special attention in an image, which can help doctors make more accurate diagnoses and reduce the proportion of misdiagnoses or missed diagnoses. In recent years, with the development of intelligent diagnosis and online medical technologies, medical image segmentation plays an extremely important role. In general, image segmentation can be divided into semantic segmentation and instance segmentation. Semantic segmentation

is usually used to classify each pixel in an image, resulting in a pixelated set. Instance segmentation is more detailed than semantic segmentation and can distinguish objects of the same category but not belonging to the same entity. Due to the unique characteristics of medical images, it is only necessary to know whether it's tissue that needs attention. Therefore, medical image segmentation belongs to semantic segmentation. Currently, popular medical image segmentation tasks include liver and liver tumor segmentation, brain and brain tumor segmentation, optic disc segmentation, cell segmentation, lung segmentation, and lung nodule segmentation [8].

2.1 Traditional Methods

Early medical image segmentation relies on methods such as edge detection, template matching, and statistical shape models.

- Threshold-based segmentation method [9] divides the greyscale histogram of a medical image into multiple classes by setting one or multiple thresholds to achieve segmentation. The threshold can be set manually or calculated by specific algorithms, such as fixed threshold segmentation, histogram bimodal method, iterative threshold image segmentation, adaptive threshold image segmentation, Otsu's method, mean method, and optimal threshold method. Because of the simplicity of the algorithm, it has been widely used.
- The basic idea of the edge detection segmentation method [10] is to first identify the edge pixels in the image and then connect them to form the target area. Edge pixels refer to the set of pixels where the greyscale of the image undergoes a spatial variation. Conventionally, first-order and second-order derivatives are employed to depict and detect these edges. First-order differential operators, such as the Roberts, Prewitt, and Sobel operators are commonly utilized to capture the primary characteristics of edges. On the other hand, second-order differential operators, such as the Laplace and Kirsh operators are frequently employed to uncover more intricate edge features.
- Clustering-based segmentation method [11] belongs to unsupervised segmentation. It divides the samples in the dataset into several disjoint subsets, each subset is called a "cluster". Specifically, Similar pixels in the dataset are grouped into the same cluster, while dissimilar pixels are assigned to different clusters. By employing this pixel classification technique, the objective is to unveil the intrinsic characteristics and underlying patterns inherent in the image.
- The method based on deformable model is an improvement of edge detection algorithms [12]. The method is based on the boundaries of the object is considering the shape, smoothness, and external forces that act on the segmented object, which are all factors that positively influence the results of the segmentation. Then, closed curves and shapes in the image are used to determine the boundary of the segmented object, which can be continuously segmented through different sections.

- Region-based segmentation method [13] includes threshold method, region growing method, region separation and merging method, and clustering segmentation method, which mainly uses the local spatial information of the image to connect and combine pixels with similar properties to obtain the final target area.
- The registration-based segmentation method [14] is mainly divided into the atlas-based segmentation algorithm and the joint segmentation and registration algorithm. The atlas-based algorithm is to align a pre-segmented image with a target image and utilizes spatial transformation parameters to deform the pre-segmented image to match the target image. Joint segmentation is to incorporate image-related structural information on the basis of registration, constructs a joint energy function, and obtains the segmentation result by minimizing the function.

These methods are basically applied directly to the information from the pixel images and to the structural similarity to achieve segmentation. Traditional segmentation methods are often susceptible to various factors, including image clarity and variations in the size of the lesion areas. With the development of deep learning in recent years, medical images no longer require manually crafted features, and convolutional neural networks can achieve hierarchical feature representation of images well and are less sensitive to factors such as clarity.

2.2 Deep Learning-Based Methods

There are two directions, supervised learning and weak supervised learning, for deep learning based medical image segmentation.

The primary motivation behind adopting weakly supervised learning for segmentation tasks in medical imaging stems from the inherent challenges associated with the manual labeling process. Labeling medical images requires the expertise and involvement of qualified professionals, which can be both time-consuming and resource-intensive. Generative confrontation networks have good applications in many fields such as image restoration, which has attracted the attention of researchers. Guibas et al. [15] proposed a pipeline that composed of GAN and cGAN for a segmentation task. They input random variables into the GAN to generate labeled images of retinal blood vessels and then put the generated image into a conditional GAN to generate real retinal fundus images. Finally, the discriminator judges the similarity between the generated image and the real image. Through continuous iteration, segmentation results are achieved. The model proposed by Chen et al. [16] is adapted to CT and MRI images. They use GAN to transform the labeled original image into the required image, which combines the features of CT and MRI and enables the network to have a good performance on images from both modalities. Then, the transformed image and the original image share the same segmentation network, and the loss function generated by both modalities is used to update the segmentation model. As a result, only CT labels are needed to train a network for MRI segmentation.

Supervised learning is a more mainstream direction. Supervised learning is to use labeled training data to learn a model, and then use this model to predict new samples. In essence, the goal of supervised learning is to construct a mapping from input to output, which is represented by a model. The current mainstream segmentation model usually adopts the encoder-decoder structure. The encoder extracts the features in the image, and the decoder restores the extracted features to the original image size and classifies the output results, typically U-Net, FCN, DeepLab etc.

In 2015, Long et al. proposed methods that can classify images at the pixel level [17]. FCN is a classification of image pixel level. It is different from CNN. CNN is fixed by using a fully connected layer in the convolutional layer. The length of the feature vector is classified and the FCN can accept images of any resolution. After downsampling, the deconvolution layer is used to restore the low-resolution features to the original resolution. In this way, each Pixel classification, while preserving the spatial information. In 2017, Yang et al. [6] proposed a segmentation method that fully exploits cross-scale information, which uses UNet as a framework and reuses features in the encoder process. The subsequent convolution layers include the features of the previous layer, and an improved ASPP is used to expand the receptive field.

In 2021, Gu et al. [18] proposed to utilize multiple attentionas to boost the feature representation in CNN, which is also a modification of the UNet framework. They use ratio attention modules and channel attention modules to suppress the expression of irrelevant information and focus more on useful information. Bao et al. [19] proposed a segmentation method based on mirror difference perception. They used the differential feature augmentation (DFA) module and the mirror position-based difference augmentation (MDA) module to compare and enhance the differences between the original image and the horizontally flipped image. The segmentation accuracy is improved by learning the feature differences between normal and diseased regions. Yu et al. [20] proposed a Fourier-based adaptive normalisation (FAN) and a domain classifier with a gradient reversal layer to reduce the domain shifting problem caused by the different sites where the MR images were acquired. This approach can improve the robustness and prediction accuracy of the model.

However, the existing relatively lightweight models have shortcomings in feature acquisition and cannot extract deeper features. In addition, when using residual connections, a lot of useless information is often added due to the superposition of a large number of channels. Furthermore, they ignore the negative impact on the network caused by the large difference in the area of lesion regions. Using a small convolution kernel for small volume images can only capture local features, so some information may be lost when processing global features. Using a large convolution kernel will ignore the feature information of small area lesions because it will capture broader features such as noise and background. Some models rely on the symmetry of the brain to discriminate the lesion area without considering into account the non-pathological asymmetry, which can lead to misclassification of the lesion area [21–27]. The model we proposed solves

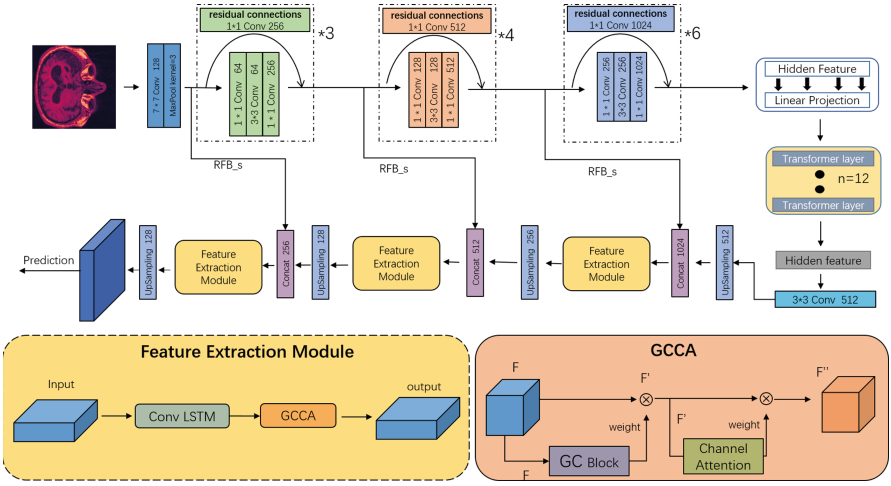

Fig. 1. Overview of our model framework (top). The data flow of the attention module of each layer of the decoder (bottom left). The detailed data flow of GCCA (bottom right). The network captures deep feature maps from the input data by ResNet50. Subsequently, these deep feature maps are fed into a transformer module, where self-attention weights are learned. Following the transformer module, the obtained feature maps undergo decoding through four decoder blocks. Black arrows indicate the direction of data flow.

the above problems through the MD module and feature refinement modules. Through sufficient experiments, the effectiveness of the two modules has been proved and better accuracy has been achieved.

Since there are relatively many labeled images in the ATLAS2.0 dataset (Anatomical Tracings of Lesions After Stroke), our model belongs to the branch of supervised learning. Using the encoder-decoder architecture idea, combined with the advantages of transformer and ResNet50, compared with recent models, the effect significantly increased.

3 Methods

In this section, we first introduce the overall architecture of the model in Sect. 3.1. Then, we introduce the implementation details of the vision transformer in Sect. 3.2. Next, we introduce the implementation details of MD module in Sect. 3.3 and the details of the FEM (feature extraction module) in Sect. 3.4.

3.1 Network Architecture

The network consists of four parts, including downsampling, vision transformer, upsampling and skip connection. We extract deep features from images by downsampling. Then the vision transformer effectively captures global contextual

information and establishes dependencies between various image regions. Upsampling decodes deep features into the final image. And skip connection combines both shallow and deep features to enhance feature information. The general framework of our model is illustrated in Fig. 1.

We adopt a strategy to improve the input representation for the network by expanding the single-channel MRI image to a three-channel input. First is the Encoder, using the first four layers of ResNet50 as our model's downsampling layers for feature extraction. Additionally, we make necessary adjustments to the output channels of the first convolution block to facilitate the integration of ConvLSTM modules. This modification enables seamless compatibility between the encoder and the subsequent ConvLSTM layers, promoting efficient information flow and enhancing the overall performance of the model, so that the final feature dimension is 14*14*1024. During downsampling, we save the features of the first three downsampling convolutional layers for skip connections. The obtained deep features are then input into the Vision Transformer, and the sequence outputs are reconstructed in dimension after the encoder and reshaped to the original input size. Finally, the decoder uses the MD module to process the downsampling features and concatenate them with each layer.

The process of concatenation facilitates the integration of multiscale information, enhancing the model's ability to capture fine-grained details and contextual cues. The output of the FEM is used as the input to the next layer of the decoder. The prediction result is obtained using softmax.

3.2 Vision Transformer

Through analysis of the dataset, we find that stroke lesion areas are generally located in three areas of the brain, namely gray matter, white matter, and lateral ventricles. However, there is a lack of labels for segmenting these three areas, so we need to use self-attention mechanism to focus on the recognition of these three locations. Therefore, we utilize the self-attention feature of the Vision Transformer to learn self-attention on the deep features extracted by ResNet50. In the original vision transformer network [28], the input is a three-channel image, which is segmented into patches of a certain size and flattened. Taking ViT-H/14 as an example, the 224*224*3 input image is convolved with a kernel to obtain a 14*14*768 feature map, which then is flattened into a 196*768 matrix. In this paper, we only use the linear projection layer and the Transformer encoder layer of ViT, which consists of four encoder blocks. The 14*14*1024 deep features extracted by ResNet50 are used as input for patch embedding. Then, a class token and learnable position embedding are added, and the input is passed to the transformer encoder. Finally, the class token is removed and the flattened two-dimensional matrix is reshaped back to a 14*14*1024 feature matrix for the upsampling stage.

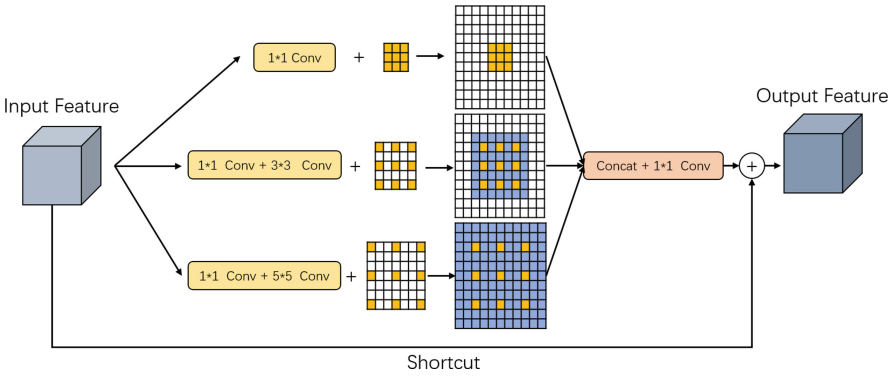

Fig. 2. The structure of MD module. The input is fed into 4 branches. The original input is added to the three branches to obtain the data after expanding the receptive field.

3.3 MD Skip Connection

One of the advantages of the UNet framework is the use of the skip structure which fuses downsampled deep features into upsampled shallow features. This structure can provide the basis for object category recognition and accurate segmentation and positioning. ResNet50 can obtain the deep features of the picture. But the deepness of the network, information will be lost and lesions in the small area cannot be accurately segmented.

In our approach, we incorporate dilated convolutions in the skip structure to expand the receptive field. This method makes up for the loss of information to a certain extent. However, a significant challenge arises in accurately segmenting stroke lesion areas due to the substantial variation of their sizes. When employing a single dilated convolution kernel, smaller lesions often go unnoticed and are consequently ignored, leading to inadequate extraction of crucial information. Taking these issues and inspired by RFBNet [29], we design a multi-dilated skip connection. As shown in Fig. 2, this module is a multi-branched convolutional layer, which comprises three individual branches with dilation rates 1, 3, and 5. Respectively, the MD module effectively combines and fuses diverse features. Incorporating dilated convolutions with varying dilation rates enables the model to capture a broader range of contextual information and significantly expand the receptive field. This strategy effectively mitigates the limitations imposed by a single dilation rate, allowing for a more comprehensive extraction of relevant features and contextual cues.

3.4 Feature Extraction Module

During the upsampling process, since each layer needs to concatenate the features from the downsampling, it doubles the number of channels, resulting in a doubling of irrelevant information in the feature matrix, and increasing the

difficulty of lesion segmentation. The overall framework of the Feature Extraction Module is illustrated in Fig. 3. To address this issue, we use GCCA module. With its channel attention and spatial attention to emphasize channels and regions which contain useful information, while ignoring irrelevant areas. In addition, we employ ConvLSTM to retain essential data in the sequence and discard irrelevant data.

GCCA. In neural networks, the attention mechanism is usually an additional neural network that can selectively choose certain parts of the input or assign different weights to different parts of the input. The attention mechanism can filter out important information from a large amount of information. We design the GCCA module which incorporate both spatial and channel attention mechanisms. The input features undergo spatial attention processing to capture their relative importance, followed by channel attention processing to learn the importance of each channel. We implement the simplified non-local attention mechanism (GC: Global Context block) [30] for the spatial attention component:

$$F'' = M_c\left(M_n(F) \otimes F\right) \otimes M_n(F) \otimes F,$$

where the F and F'' respectively represent the input and output of the module. They have the same dimension. $M_n(F) \otimes F$ represents the operation in the first half, which is the GC block. And $M_c\left(M_n(F) \otimes F\right)$ represents the channel attention.

Compared with non-local attention module, GC block eliminates unnecessary calculation and greatly reduces the calculation cost of parameters. At the same time they are almost identical in their ability to obtain a position-independent global context. GC block performs global attention pooling on the input, uses 1*1 convolution and softmax functions to obtain attention weights, and multiplies them with the flattened features to obtain global contextual features. Then the 1*1 convolution is deployed to integrate the extracted features with the original ones, thereby acquiring comprehensive global contextual information:

$$M_n(F) = F + W_2 * W_1 * \sum_{n=1}^{N_p} \frac{\exp\left(W_c * F_n\right)}{\sum_{j=1}^{N_p} \exp\left(W_c * F_j\right)} * F_n,$$

where F represents the C*H*W input. $\frac{\exp(W_c * F_n)}{\sum_{j=1}^{N_p} \exp(W_c * F_j)}$ represents the weight of global attention pooling. N_p represents the number of positions in the F. n and j respectively represent the index of the position, enumerating all possible positions. W_2 and W_1 represent the linear transformation matrix respectively.

Then the channel attention is performed by using both max pooling and average pooling on the input. The average pooling and max pooling are used to aggregate the spatial information of the feature map. And we do not perform spatial compression when aggregating spatial information to avoid losing some spatial information during compression and restoration. Then, the module uses share MLP module to integrate information from both feature maps. After the

sigmoid activation function, the weights of channel attention are obtained by addition:

$$M_C(F) = \sigma\left(W_1 * W_0 * (F_{avg}) + W_1 * W_0 * (F_{\max})\right),$$

where F_{avg} represents the average pooling. $F_{\max}$ represents the maximum pooling. W_1 and W_0 represent the linear transformation matrix (1*1 Conv). σ represents the activation function of the sigmoid.

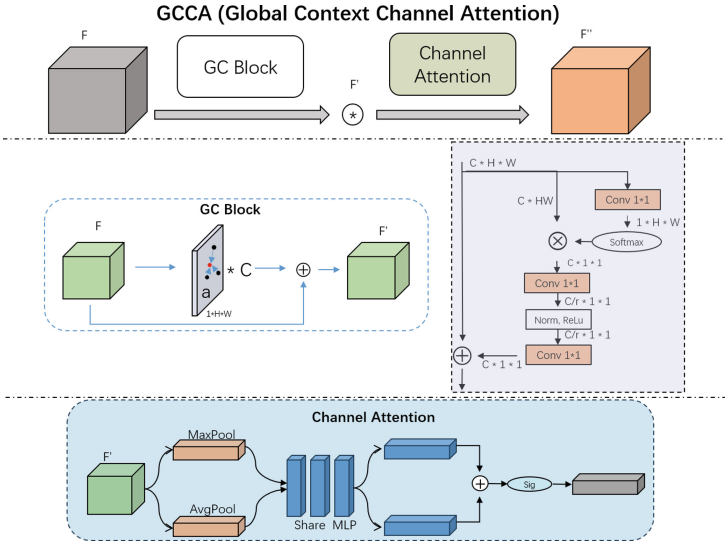

Fig. 3. The structure of GCCA. On the top is an overview of GCCA, which is composed of simplified non-local attention and channel attention. In the middle is the detailed data flow of the GC block, and it represents the detailed implementation process of the GC (Global Context) block. The bottom one shows the detailed data flow of channel attention.

ConvLSTM. LSTM is good at handling temporal information and can also handle spatial information. However, for three-dimensional images, each pixel in the image has strong correlations with the surroundings and contains extremely rich spatial information, which cannot be captured by traditional LSTM. Therefore, we use ConvLSTM [31], which differs from traditional LSTM by using 3D tensors instead of 2D inputs. The difference between traditional LSTM is that it changes from 2D input to 3D tensor.

ConvLSTM is implemented using four components. One is the forget gate, which reads the previous input and current input, applies a sigmoid nonlinearity, outputs a vector used to decide what to forget from the cell state, and finally multiplies with the cell state. The second component is the input gate, which

determines the new information to be stored in the cell state. The third component is the cell state, which is the result of adding the output of the forget and input gates, equivalent to filtering and storing spatial information. The final output gate controls the visibility of the state value at time t.

3.5 Loss Function

During the training process, our model takes in a pair of brain MRI images and a black-and-white image indicating the lesion region, which is used to compare with the predicted lesion region by the model. Our loss function consists of the Dice coefficient loss and the Focal loss.

Dice Loss. This is a measurement function that calculates the similarity of two sets. It is the ratio of the intersection of the two sets to the total number of the two sets. Here we calculate their ratio by the number of pixels. It can be defined as:

$$L_{\text{dice}} = 1 - \frac{2 * \sum_i^N P_i * G_i}{\sum_i^N P_i + \sum_i^N G_i},$$

where $\sum_i^N P_i * G_i$ represents the intersection of the predicted area and the ground truth. i represents the position of the pixel. And $\sum_i^N P_i + \sum_i^N G_i$ represents the sum of pixels in the predicted area and the ground truth area.

Focal Loss. Focal loss is a loss function that solves the unbalanced classification of samples. It focuses on adding loss weights to the losses corresponding to samples according to the difficulty of sample resolution, adding smaller weights to samples that are easy to distinguish, and adding greater weight to samples that are not easy to distinguish. Samples add larger weights, defined as:

$$L_{\text{focal loss}} = -\gamma_t * (1 - P_t)^{\alpha} * \log(p_t),$$

where $\alpha = 2$. P_t is the probability of making a correct prediction for the class t. Our goal is to segment the lesion area. This is a dichotomous task. Therefore $t \in (0, 1)$. $\gamma_{t=0} = 1$ and $\gamma_{t=1} = 50$.

4 Experimental Results

In this section, we first introduce the dataset and evaluation metrics for stroke lesion segmentation. Then we perform qualitative and quantitative experiments on the effectiveness of our proposed model.

4.1 Dataset and Experiment Settings

We evaluate our model on the ATLAS-V2.0 dataset [32]. The dataset has 401 MRI case images, which contain a large number of images of non-lesional areas. In order to reduce the number of images of non-lesional areas, we select the MRI slices in the middle part and finally obtain more than 70,000 pairs of images. We select more than 40,000 images as training sets, and about 15,000 images are used as test sets and verification sets. We crop these data to a uniform size of 224*224.

We implement our proposed method and comparison methods with CUDA 11.1 and train them on a single NVIDIA RTX 3090 GPU. We train our model using Adam optimizer with batch size 8 and weight decay $10^{-}8$. The learning rate for our approach is set as 0.0001.

4.2 Evaluation Metrics

Dice score is one of the commonly used indicators for image segmentation. It is a set similarity measurement function that is used to calculate the similarity between two samples. It is defined as follows:

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2 * TP}{2 * TP + FN + FP},$$

where TP represents the number of pixels that are themselves positive samples and are predicted to be positive samples. FN represents the number of pixels whose label is a negative sample and is correctly predicted as a negative sample. FP represents that the label itself is a negative sample but is predicted to be a positive sample The number of pixels. FP represents the number of pixels whose label itself is a positive sample but is predicted to be a negative sample.

Precision is also a commonly used image segmentation index. It considers how many of the predicted positive samples are positive samples themselves. The larger the value, the higher the accuracy of the prediction. It is defined as follows:

$$Precision = \frac{TP}{TP + FP}.$$

Recall is similar to Precision. It considers how many positive samples on the label are correctly predicted. The larger the value, the more comprehensive the prediction. It is defined as follows:

$$Recall = \frac{TP}{TP + FN}.$$

IOU calculates the overlap ratio of the intersection of two sets and their union, and the larger the value, the more consistent the predicted lesion area with the location of the label. It is defined as:

$$IOU = \frac{TP}{TP + FN + FP}.$$

Table 1. Quantitative comparison results on the ATLAS-V2 Dataset of our network compared to the state-of-the-art methods for brain stroke lesion segmentation in terms of Dice, IOU, Precision, RVD, VOE, and Recall. The best result is in bold.

Methods	Dice	Precision	Recall	VOE	RVD	IOU
U-Net	0.7604	0.8743	0.8318	0.6198	–0.7946	0.7490
CA Net	0.6558	0.7929	0.8080	0.6721	**–0.6634**	0.6511
DeepLab3+	0.8152	**0.9793**	0.8179	0.5924	–0.9681	0.8069
TransUnet	0.8581	0.9602	0.8516	0.5709	–0.9000	0.8404
CLCI_Net	0.8062	0.9300	0.8394	0.5969	–0.9102	0.7909
Our model	**0.8681**	0.9692	**0.8669**	**0.5660**	–0.8921	**0.8494**

RVD represents the difference between the volume of the predicted region and the region in the label, which can be defined as:

$$RVD = \frac{V_{seg}}{V_{gt}} - 1.$$

VOE can be called the volume overlap error, which represents the error rate. It can be said to be the opposite of the definition of Dice, which can be defined as:

$$VOE = 1 - \frac{|X \cap Y|}{|X \cup Y|} = 1 - \frac{TP}{TP + FN + FP}.$$

4.3 Comparison with Baselines

We compare our model to five models, including UNet [4] and DeepLab3+ [5], which are classic medical image segmentation models. CA-Net [18], TransUnet [7], and CLCI-Net [6] are proposed in recent years and have relatively good performance in ATLAS segmentation. We demonstrate the superiority of our model by observing several evaluation metrics, including Dice, IOU, RVD, VOR, Recall, and Precision. As shown in Table 1, our proposed model achieves 0.8681, 0.9692, 0.8669, 0.5660, –0.8921, and 0.8494 for Dice, precision, recall, VOE, RVD, and IOU, respectively.

Compare with the compared models, our model's Dice score improves by 0.01 to 0.1, indicating a higher similarity between the predicted labels and Ground Truth. Our model's precision score is 0.01 lower than DeepLab3+ but improves by 0.08 to 0.15 compared to other models. This is mainly due to the inclusion of the FE module, which uses spatial attention to represent channels carrying useless information during the upsampling process with the help of spatial attention. Non-local and ConvLSTM help the network to focus more on the feature representation of the target area, while retaining more effective feature information and avoiding the influence of too many background features on the prediction. Next is the recall metric, which measures how many true positives are predicted as positives. Our model's recall score improves by 0.03 to 0.07

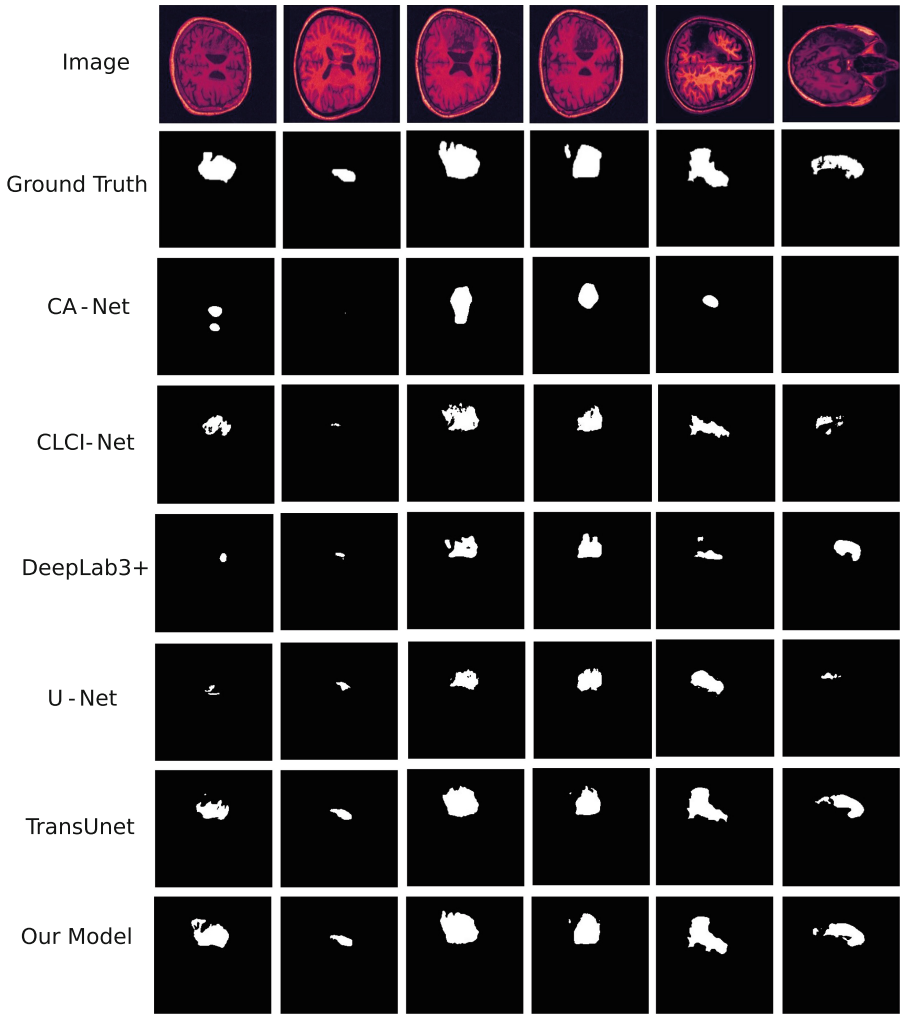

Fig. 4. Comparisons of our method, U-Net, DeepLabv3+, CLCI-Net, CA-Net and Trans-Unet on six different patients.

compared to other models, indicating more accurate predictions of lesion areas. The VOE metric measures the opposite effect of Dice, and our model is 0.02 to 0.11 lower than other models, indicating a lower error rate. RVD and IOU are metrics that measure the difference in volume between the predicted region and the Ground Truth. The improvement in these metrics is mainly due to the fact that the MD module expands the sensory field to adequately extract feature information through multi-scale cavity convolution.

In general, our model achieves higher accuracy in stroke lesion segmentation after using dilated pooling convolution and feature enhancement modules, with significant improvements in region size and overlap.

4.4 Ablation Analysis

We conduct ablation experiments on the ATLAS-V2.0 dataset to evaluate the effectiveness of our proposed module:

- Baseline: Change the convolution output channel of the first layer of ResNet downsampling in TransUNet to 128 as a benchmark for comparison.
- B + MD module: Add MD module on the basis of Baseline.
- B + Feature Extraction Module: Add a feature extraction module composed of GCCA and ConvLSTM on the basis of Baseline.
- B + Feature Extraction Module + MD module: Filling module (our module).

Component Analysis of the MD Module. As Shown in Table 2, we incorporate an MD module that combines features with different receptive fields through dilated convolutions at different rates. This novel approach enables the network to effectively handle lesion regions characterized by diverse sizes and shapes encountered within the dataset. Compare to the baseline model, the addition of the MD module lead to an increase in the Dice coefficient, Precision, and Recall, indicating improved performance. As a result, the incorporation of the MD module enhances the prediction of lesion regions with improved accuracy and precision. The results demonstrate that this dynamic fusion process enhances the model's ability to capture both local details and global context, resulting in improved performance for the semantic segmentation task. The MD module is an effective approach to improve the model's performance in predicting lesion regions.

Table 2. Results of component ablation studies of MD-TransUNet. The best result is in bold.

Methods	Dice	Precision	Recall	VOE	RVD	IOU
Base-model	0.8529	0.9681	0.8575	**0.5736**	–0.8978	0.8347
B + FEM	0.8621	0.9640	**0.8683**	0.5689	**–0.8774**	0.8441
B + MD	0.8554	0.9560	0.8659	0.5723	–0.8803	0.8369
B + FEM + MD(our model)	**0.8681**	**0.9692**	0.8669	0.5660	–0.8921	**0.8494**

Component Analysis of the FE Module. To address the loss of features associated with the downsampling process, we incorporate a feature extraction module into our model to repair and compensate for the lost features. Our model showed improvements in various evaluation metrics compared to the baseline model and achieved comparable results to models that contained only the MD

module. Our model exhibits superior performance across all evaluation metrics compared to models that featured only a single module. This indicates that the MD module improves the network's ability to integrate complete feature information, and the feature extraction module retains more useful features. These modules complement each other effectively. Consequently, our model exhibits a considerable advantage over other models.

4.5 Visual Comparison

To further validate the effectiveness of our MD-TransUnet network, we select six groups of stroke MRI images and their predicted results from various models in Fig. 4. It is clear from the images that our model's predictions are superior to those of other models. As shown in the evaluation metrics such as the Dice coefficient in the table above, our model can achieve more accurate predictions in small lesion areas compared to U-Net, CLCI-Net, and other models, demonstrating the effectiveness of incorporating the MD module. In predicting large lesion areas, our model's predictions are more detailed, and the predicted shapes are extremely similar to the Ground Truth, indicating that the GCCA module can indeed enrich the details of the predicted results. These prove the effectiveness of our model.

5 Conclusions

This paper proposes a novel deep neural network based on UNet, which is specifically designed to segment brain hemorrhages in stroke patients. To enhance the performance of the model, we introduce the MD module into the skip-connection part, which provides a connection between deep down-sampling and up-sampling features. This approach preserves more information about the texture and edge pertinent to the hemorrhagic region during the convolution process. To mitigate the information loss that arises from the depth of the network during the downsampling process, we employ a feature extraction module consisting of ConvLSTM and GCCA. This module learns long-term dependency information and retains additional information. The GCCA specifically learns multilevel attention weights that help eliminate useless channels and retain more useful information on high channel features in each layer of the up-sampling process. GCCA also reduces the impact of background noise and counteracts other negative factors that could undermine the validity of the spatial dimension information. We further perform comparative experiments with other models in the same field and ablation experiments to test the model's effectiveness.

Our future goal is to adopt more advanced techniques, such as clinician evaluations comparing regions of the left and right brain regions, to achieve greater precision in the segmentation of stroke lesions.

Acknowledgements. This research was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province under No. 2022C03043, Natural Science

Foundation of Zhejiang Province under No. LQ21F020015, and the Open Research Project Fund of Key Laboratory of Marine Ecosystem Dynamics, Ministry of Natural Resources under Grants MED202202.

References

1. Feigin, V.L., 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. The Lancet Neurol. **20**(10), 795–820 (2021)
2. Hevia-Montiel, N., et al.: Robust nonparametric segmentation of infarct lesion from diffusion-weighted MR images. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2102–2105. IEEE (2007)
3. Ozertem, U., Gruber, A., Erdogmus, D.: Automatic brain image segmentation for evaluation of experimental ischemic stroke using gradient vector flow and kernel annealing. In: 2007 International Joint Conference on Neural Networks, pp. 1397–1400. IEEE (2007)
4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
5. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with Atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
6. Yang, H., et al.: CLCI-Net: cross-level fusion and context inference networks for lesion segmentation of chronic stroke. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 266–274. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_30
7. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
8. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: a survey. IET Image Process. **16**(5), 1243–1267 (2022)
9. Johnson, L.A., Pearlman, J.D., Miller, C.A., Young, T.I., Thulborn, K.R.: MR quantification of cerebral ventricular volume using a semiautomated algorithm. Am. J. Neuroradiol. **14**(6), 1373–1378 (1993)
10. Pujar, J.H., Gurjal, P.S., Kunnur, K.S., et al.: Medical image segmentation based on vigorous smoothing and edge detection ideology. Int. J. Electr. Comput. Eng. **4**(8), 1143–1149 (2010)
11. Li, B.N., Chui, C.K., Chang, S., Ong, S.H.: Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. Comput. Biol. Med. **41**(1), 1–10 (2011)
12. Jayadevappa, D., Srinivas Kumar, S., Murty, D.S.: Medical image segmentation algorithms using deformable models: a review. IETE Tech. Rev. **28**(3), 248–255 (2011)
13. Patil, D.D., Deore, S.G.: Medical image segmentation: a review. Int. J. Comput. Sci. Mobile Comput. **2**(1), 22–27 (2013)
14. Hao, L.: Registration-based Segmentation of Medical Images. School of Computing National University of Singapore, Singapore (2006)

15. Guibas, J.T., Virdi, T.S., Li, P.S.: Synthetic medical images from dual generative adversarial networks. arXiv preprint arXiv:1709.01872 (2017)
16. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Trans. Med. Imaging **39**(7), 2494–2505 (2020)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
18. Ran, G., et al.: CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging **40**(2), 699–711 (2020)
19. Bao, Q., Mi, S., Gang, B., Yang, W., Chen, J., Liao, Q.: MDAN: mirror difference aware network for brain stroke lesion segmentation. IEEE J. Biomed. Health Inform. **26**(4), 1628–1639 (2021)
20. Yu, W., Lei, Y., Shan, H.: FAN-Net: fourier-based adaptive normalization for cross-domain stroke lesion segmentation. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
21. He, X., Chen, K., Yang, M.: Semi-automatic segmentation of tissue regions in digital histopathological image. In: Gao, H., Wang, X. (eds.) CollaborateCom 2021. LNICST, vol. 406, pp. 678–696. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92635-9_39
22. Lin, B., Deng, S., Yin, J., Zhang, J., Li, Y., Gao, H.: FocAnnot: patch-wise active learning for intensive cell image segmentation. In: Gao, H., Wang, X., Iqbal, M., Yin, Y., Yin, J., Gu, N. (eds.) CollaborateCom 2020. LNICST, vol. 350, pp. 355–371. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67540-0_21
23. Abdmouleh, N., Echtioui, A., Kallel, F., Hamida, A.B.: Modified u-net architeture based ischemic stroke lesions segmentation. In: 2022 IEEE 21st international Ccnference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), pp. 361–365. IEEE (2022)
24. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med. Image Anal. **87**, 102792 (2023)
25. Liu, L., Huang, C., Cai, C., Zhang, X., Hu, Q.: Multi-task learning improves the brain stoke lesion segmentation. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2385–2389. IEEE (2022)
26. Thiyagarajan, S.K., Murugan, K.: Performance analysis of ischemic stroke lesion segmentation in brain MR images using histogram based filter enhanced FCM. In: 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1343–1348. IEEE (2023)
27. Aboudi, F., Drissi, C., Kraiem, T.: Efficient u-net CNN with data augmentation for MRI ischemic stroke brain segmentation. In: 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), vol. 1, pp. 724–728. IEEE (2022)
28. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
29. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 385–400 (2018)

30. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: GCNet: non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
31. Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
32. Liew, S.-L., et al.: A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci. Data **5**(1), 1–11 (2018)