scientific reports

OPEN

SFMANet: A Spatial-Frequency multi-scale attention network for stroke lesion segmentation

Hualing Li✉, Jianqi Wu, Yonglai Zhang & Lei Wang

In neuroimaging analysis, accurately delineating stroke lesion areas is crucial for assessing rehabilitation outcomes. However, the lesion areas typically exhibit irregular shapes and unclear boundaries, and the signal intensity of the lesion may closely resemble that of the surrounding healthy brain tissue. This makes it difficult to distinguish lesions from normal tissues, thereby increasing the complexity of the lesion segmentation task. To address these challenges, we propose a novel method called the Spatial-Frequency Multi-Scale Attention Network (SFMANet). Based on the UNet architecture, SFMANet incorporates Spatial-Frequency Gating Units (SFGU) and Dual-axis Multi-scale Attention Units (DMAU) to tackle the segmentation difficulties posed by irregular lesion shapes and blurred boundaries. SFGU enhances feature representation through gating mechanisms and effectively uses redundant information, while DMAU improves the positioning accuracy of image edges by integrating multi-scale context information and better allocates the weights of global and local information to strengthen the interaction between features. Additionally, we introduce an Information Enhancement Module (IEM) to reduce information loss during deep network propagation and establish long-range dependencies. We performed extensive experiments on the ISLES 2022 and ATLAS datasets and compared our model's performance with that of existing methods. The experimental results demonstrate that SFMANet effectively captures the edge details of stroke lesions and outperforms other methods in lesion segmentation tasks.

Medical image segmentation constitutes a particularly challenging research area in computer vision. The objective of this task is to identify and extract relevant regions of medical images, providing a reliable foundation for clinical diagnosis, pathological research, and assisting clinicians in making accurate decisions. It has numerous applications in areas such as disease diagnosis, clinical decision support, and pathological analysis.

In recent decades, significant progress has been made in studying brain injuries and exploring brain anatomy using magnetic resonance imaging (MRI). These advances have generated large volumes of high-quality medical image data. However, for clinicians, analyzing these extensive and intricate MRI datasets and manually extracting critical information has become a burdensome task. Manual analysis is time-consuming and prone to errors, often influenced by variations in experience between different clinicians or among individual practitioners.

With the advancement of deep learning technologies, neural network-based methods have become the dominant approach for medical image segmentation. These methods are primarily based on convolutional neural networks (CNNs). Long et al.[1] introduced the fully convolutional network (FCN), replacing fully connected layers with convolutional layers, which significantly enhanced the computational efficiency of image segmentation. Building upon the FCN, Ronneberger et al.[2] developed U-Net, which consists of an encoder-decoder architecture. The encoder's shallow and deep networks extract simple and abstract features from the image, respectively, while skip connections between the encoder and decoder facilitate the capture of contextual information. Since then, U-Net and its various extensions have substantially advanced medical image segmentation.

Recent advancements in medical image segmentation have shown significant progress, particularly in leveraging the strong generalization ability of convolutional kernels to extract high-dimensional features, which is crucial for visual tasks. However, there are still some disadvantages of stacked convolutional layers and downsampled layers: (1) as the number of convolutional layers increases, the model parameters increase significantly; (2) The small perceptual domain of the convolution operation cannot establish the long-distance dependence between pixels. This limitation may lead to incomplete or incorrect segmentation of the margins of the stroke lesion area[3]; (3) Continuous downsampling and convolution operations in the encoding phase will lead to a large loss of high-level semantic information[4].

School of Software, North University of China, Taiyuan, Shanxi, China. ✉email: lihualing750108@163.com

In order to solve the problem of insufficient long-distance dependence of convolutional neural networks, In order to solve the problem of insufficient long-distance dependence of convolutional neural networks, researchers have introduced Transformer[5] technology. Transformers can theoretically capture the dependencies between any two locations and have powerful global information extraction capabilities. However, Transformer cannot learn local information effectively and cannot naturally process multi-scale information through a hierarchical structure like CNNs. Therefore, transformers need additional mechanisms to introduce locality and multiscale when processing images, such as introducing local window attention (such as Swin Transformer[6]) or combining with CNNs. For example, TransUNet[7] absorbs the advantages of ViT[8] and UNet, and mixes the Transformer structure that emphasizes global information with CNN for hybrid encoding, which improves the segmentation effect of UNet. In later work, researchers proposed the use of different types of transformers for medical image segmentation. For example, EG-TransUNet[9] uses a self-attention-based Transformer in the encoder and decoder stages to improve the discrimination ability at the spatial detail and semantic position levels. DS-TransUNet[10] uses a dual-scale coding mechanism to extract coarse-grained and fine-grained feature representations at different semantic scales. UNETR[11] the Transformer is used as an encoder to learn the sequence representation of the input and effectively capture the global multi-scale information. SSCFormer[12] takes the hybrid structure of ConvNet-Transformer as the framework and co-captures the features of the encoder within and between scales through the intra-scale ResInception and the inter-scale Transformer bridge. Although these studies use Transformer and CNN to learn multi-scale information, in multi-scale tasks, if the shape of the lesion area is irregular or the boundary is blurred, the mixed model may have the problem of small target failure to detect or edge localization due to improper allocation of local and global information weights. In addition, as the number of network layers increases, the model is also more susceptible to a large number of redundant features.

In this study, we propose a novel multi-scale network based on UNet and attention mechanism, named it as a spatial-frequency multi-scale attention network, to solve the problems mentioned above. Initially, UNet's encoder used pooling operations to reduce the size of the feature map. Although computational efficiency is improved, it leads to the loss of feature information and the difficulty of recovering the edges and details of the image. To alleviate this problem, we have designed an information enhancement module. The module performs feature enhancement by combining multi-scale depth separable convolutions and establishes long-distance dependencies using axial attention[13]. Subsequently, to make effective use of the redundancy characteristics caused by the increase of network layers, a spatial-frequency gating unit is introduced in this paper. To solve the problem of boundary blurring caused by the fact that the stroke area and the surrounding brain tissue may have similar signal representations, this unit refines the redundant features of the spatial and frequency domains and enhances the feature representation of the lesion region through the gating mechanism. In addition, in order to solve the problem of uneven distribution and different sizes of lesions, we combine the attention mechanism with biaxial multi-scale feature extraction to better allocate the weight of global and local information, enhance feature expression, and better locate the stroke lesion area. We evaluated the network on an open-source stroke lesion segmentation dataset and achieved competitive segmentation performance.

The contributions of this work are summarized as follows:

·We propose a Spatial-Frequency Gating Unit, which can effectively refine the feature redundancy in the spatial and frequency domains and enhance the refinement features.

·We introduce the Dual-axis Multi-scale Attention Unit (DMAU), which can better allocate the weights of global and local information, enhance feature expression, and be used to locate stroke lesions.

·We design an Information Enhancement Module (IEM) that mitigates information attenuation during deep propagation, facilitating the establishment of long-range dependencies.

·We present SFMANet, built upon the proposed SFGU and DMAU, which demonstrates excellent segmentation performance. This network design is highly significant for advancing the application of medical imaging from the laboratory to clinical settings.

Related work

In this section, the network structure and multi-scale feature fusion for medical image segmentation are briefly introduced, and we also review some attention mechanisms related to our work.

Architecture of medical image segmentation

Deep learning technologies have demonstrated outstanding performance in semantic segmentation and are widely employed in medical image analysis and diagnostic support. Applications include the segmentation of coronary artery trunks[14]lung lesions associated with COVID-19[15], prostate tissues[16]and brain tumors[17]. U-Net, a widely used architecture in medical image segmentation, combines low-dimensional and high-dimensional feature information through its unique skip connection mechanism between the encoder and decoder, enabling high-quality segmentation even with relatively small datasets. In recent years, researchers have made various enhancements to the original U-Net architecture to further improve segmentation accuracy[18–26]. Notable advancements include UNET + +[27], which introduces dense skip connections to strengthen feature transfer; R2U-Net[28]which integrates the advantages of ResNet and U-Net; and KiU-Net[29]which proposes a novel structure that utilizes both incomplete and super-complete features to better segment small anatomical structures.

Additionally, Transformer-based frameworks have gained prominence in medical image segmentation. Labbihi, Ismay, et al.[30] proposed a 3D medical image segmentation model that combines CNNs with Transformers, incorporating a CNN encoder and a frequency transformer branch to reduce model parameters, computation time, and reliance on large datasets. Cao et al.[31] introduced SwinUNet, based on the Swin Transformer, which reduces computational resource demands by calculating self-attention within a limited window. Jiang et al.[32] proposed a gated axial attention ResNeSt[33] network for polyp segmentation, utilizing a global-local training strategy, with ResNeSt as the backbone, a parallel decoder to aggregate features, and gated axial attention to adapt to small datasets for effective polyp segmentation.

Multi-scale feature fusion

The detailed texture and contextual information captured in multi-scale features are essential for enhancing model performance and robustness. Reza Azad et al.[34] proposed a multi-scale encoder-decoder architecture based on an efficient variant of the Transformer block. By incorporating patch blocks of various scales into the Transformer block, a feature map with multi-scale representations is generated. Wang et al.[35] introduced MIFNet, which leverages multi-scale input and feature fusion to automatically extract and combine features from different input scales, significantly improving cardiac magnetic resonance image segmentation. Xu et al.[36] presented EC-CaM-UNet, a U-Net-based model with collaborative attention-guided multi-scale feature fusion, enhanced by convolution. The model employs a multi-dimensional collaborative attention module to estimate local and global self-attention, which is then deeply fused with the multi-scale feature map produced by the multi-scale module, enhancing the prominence of relevant multi-scale features while suppressing irrelevant ones. Yan et al.[37]. proposed MSLF-Net, a supervised segmentation method utilizing multi-scale and multi-level feature fusion. This method employs a cross-layer structure to improve feature fusion and merges low-level and high-level features of the same category based on category supervision to prevent feature contamination. Yu et al.[38] introduced a novel Transformer-CNN hybrid network, which achieves more accurate segmentation by fully fusing semantic features of various scales generated by a pyramid decoder for each segmentation category. Fu et al.[39] proposed HmsU-Net, a hybrid multi-scale segmentation network that addresses inconsistent feature learning between CNNs and Transformers at the same stage by employing cross-axis attention. Abdelrahman et al.[40] proposed the UNETR++network. By introducing the EPA block, spatial and channel attention is applied in two branches to effectively capture rich spatial and channel features. Huang et al.[41] proposed an enhanced Transformer context bridging module, which models the long-range dependencies and local context of the multi-scale features generated by the hierarchical Transformer encoder through the use of the enhanced Transformer module.

Attention mechanism

Building on the success of attention mechanisms in natural language processing (NLP), researchers have increasingly applied this concept to computer vision tasks. In semantic segmentation, numerous influential works have incorporated attention modules. For instance, Hu et al.[42] introduced the squeeze-excitation (SE) block, which adaptively recalibrates feature responses across channels by explicitly modeling interdependencies between them. Woo et al.[43] proposed a simple yet effective attention module for feed-forward convolutional neural networks, based on the SE block, which sequentially generates attention maps along two independent dimensions—channel and spatial. These attention maps are then multiplied by the input feature maps for adaptive feature refinement. Ouyang et al.[44] developed a new, efficient multi-scale attention module that reshapes part of the channels into the batch dimension and groups the channel dimension into multiple sub-features, thereby ensuring that spatial semantic features are well-distributed within each feature group.

Method

Overall architecture

The overall architecture of the Spatial-Frequency Multi-scale Attention Network for stroke lesion area segmentation is shown in Fig. 1. This network architecture is built on the classic U-Net framework and combines the Spatial-Frequency Gating Unit (SFGU), the Dual-axis Multi-scale Attention Unit (DMAU), and the Information Enhancement Module (IEM). The IEM enhances the features before downsampling through multi-scale depthwise separable convolutions to mitigate the loss of semantic information caused by the downsampling operation and uses axial attention to model the long-range dependencies. The SFGU employs the spatial gating unit (SGU) to enhance the spatially representative features and suppress the redundant features in the spatial dimension and uses the frequency gating unit (FGU) to enhance the representative features in the frequency domain and suppress the redundant features in the frequency domain dimension. DMAU enhances the features processed by the SFGU through the cascaded extraction of multi-scale features in the horizontal and vertical directions and adopts the attention mechanism to better allocate the weights of global and local information. In the subsequent sections, we will describe in detail the implementation process of SFMANet and its constituent modules.

Spatial-Frequency Multi-scale attention network

The backbone network of SFMANet is UNet. The given input image X has a size of $H \times W \times C$, where H is the width of the input image, W is the height of the input image, and C is the number of channels of the input image. SFMANet feeds the features $x_{64}, x_{128}, x_{256}, x_{512}$ obtained from each layer of the encoder into the information enhancement module. First, multi-scale depthwise separable convolutions are used for feature enhancement. Then, axial attention is applied to modeling the long-range dependencies. After that, the features are downsampled through max pooling operations.In the skip connections, the features $x_{64}, x_{128}, x_{256}, x_{512}$ obtained from the encoder are input into the spatial-frequency gating unit (SFGU) to refine the redundant features. Firstly, depthwise separable convolutions are used to initialize the attention mechanism, providing the basic features for the subsequent removal of spatial redundant features. Then, the obtained basic feature maps are respectively input into the SGU unit and the FGU unit for processing.The SGU unit uses a separation and reconstruction operation. Group normalization is utilized to normalize the input features. Subsequently, by calculating the normalized weights and applying the Sigmoid function, a reweighted tensor is obtained. Through the gating mechanism, the SGU can divide the input features into informative and non-informative

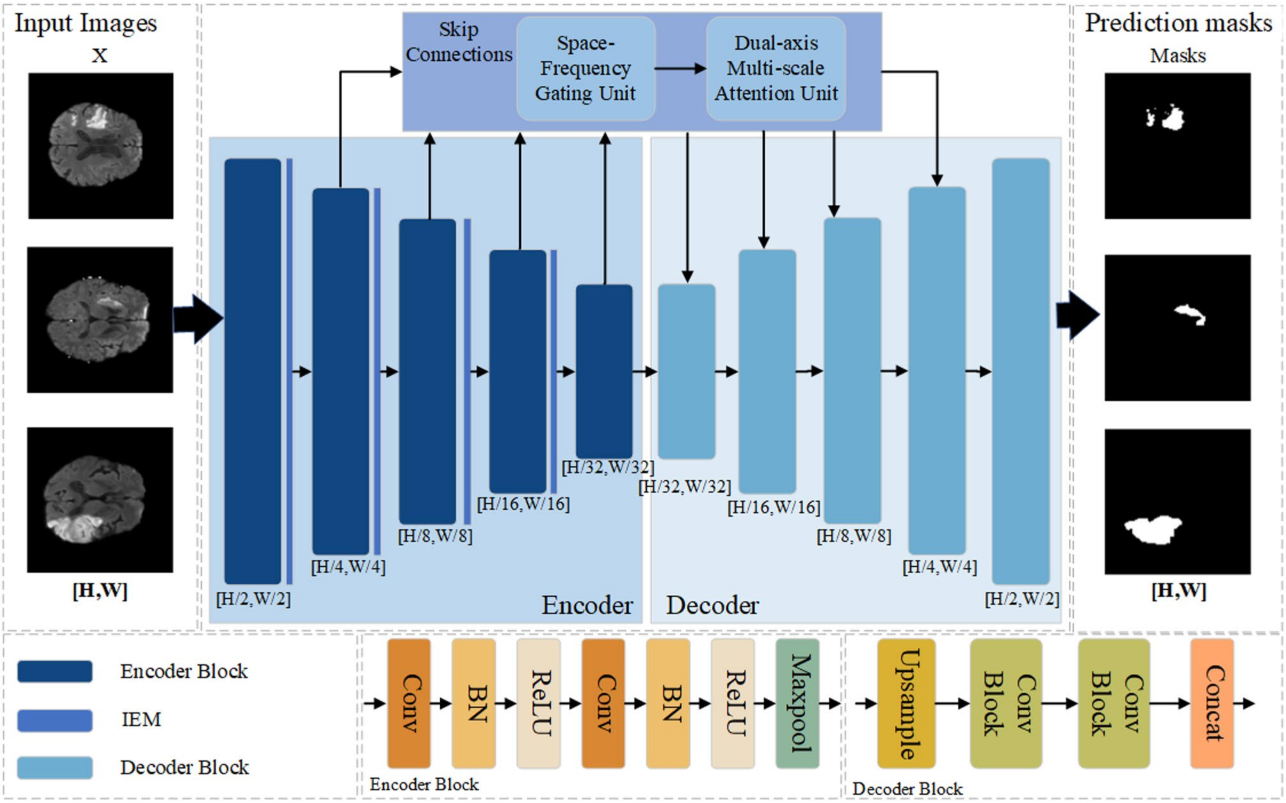

Fig. 1. Overview of our proposed medical image segmentation method.

parts according to the reweighted tensor, and reconstruct the features by cross-adding, finally obtaining the refined feature map.The FGU unit divides the input feature map into high-frequency and low-frequency parts and respectively transforms and fuses these two parts to enhance the model's ability to capture features.After that, the feature maps calculated by the SGU unit and the FGU unit are respectively input into the DMAU unit. Convolution kernels with sizes of 7×7、 11×11 and 21×21 are used to extract multi-scale features from the features. Then, these multi-scale features processed by the DMAU unit are added to the initial attention features, and then channel mixing is carried out through a pointwise convolution layer to generate the weights of global and local information. The fused attention features are multiplied by the original input features to obtain the feature map after weight adjustment.Finally, the feature map generated by the skip connection is concatenated with the feature map generated by the previous layer of the decoder, and upsampling is carried out again until the final prediction output is obtained. The algorithm process of SFMANet is summarized as Algorithm 1.

Spatial-Frequency gating unit

Spatial gating unit

To leverage spatially redundant information in features, we introduced a spatial gating unit, inspired by Li et al.[45]which employs a separation and fusion operation. The structure is illustrated in Fig. 2. The purpose of the separation operation is to differentiate feature maps with rich spatial information from those with less spatial information. We utilize the scaling factors and biases in the Spatial Group Normalization (SGN) layer to assess the information content of different feature maps. Specifically, given a feature map $X \in \mathbb{R}^{N\times C\times H\times W}$ where the batch axis is denoted as N, the channel axis as C, and the spatial height and width axes as H and W, we first standardize the input feature X by subtracting the mean μ and dividing by the standard deviation σ .

$$\mathrm{X_{out}} = SGN\left(\mathbf{X}, G_c, G_s, \gamma\ , \beta\ \epsilon\right) = \left[\gamma\ \cdot\ \left(\frac{x_{icghw} - \mu_{\ igs}}{\sqrt{\sigma\ _{igs}^{2} + \epsilon}}\right) + \beta\right]_{i=1}^{N}{}_{，c=1，g=1，s=1，h=1，w=1}^{\quad C \quad G \quad G \quad H \quad W} \tag{1}$$

Among them, X is the input feature map, G_c and G_s are the number of channel groups and the number of space groups respectively, γ and β are learnable scaling and offset parameters, ϵ is a small positive number used to prevent division by zero, x_{icghw} is the element in the i-th sample, the g-th channel group, and the s-th space group, $\mu\ _{igs}$ is the mean of the i-th sample, the g-th channel group, and the s-th space group, $\sigma\ _{igs}^{2}$ is the variance of the i-th sample, the g-th channel group, and the s-th space group, $\mu\ _{igs}$ and $\sigma\ _{igs}^{2}$ the calculation formulas of are as follows:

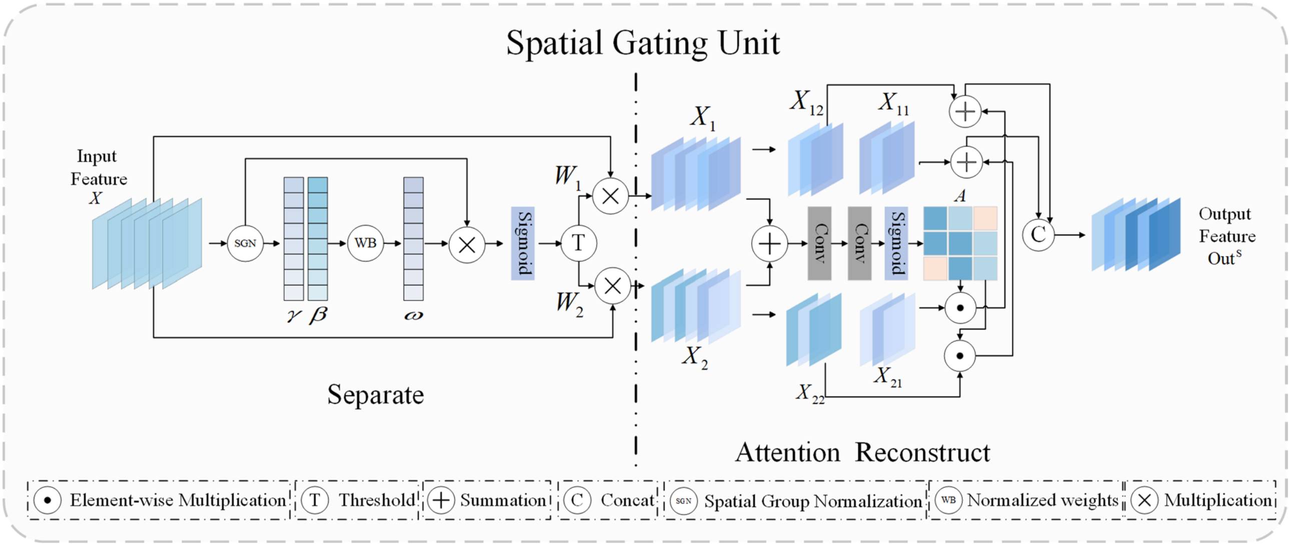

Fig. 2. The structure of SGU.

$$\mu_{igs} = \frac{1}{\frac{C}{G_c}\cdot\frac{H\cdot W}{G_s}} \sum_{c'=1}^{\frac{C}{G_c}} \sum_{h'=1}^{\frac{H}{G_s}} \sum_{w'=1}^{\frac{W}{G_s}} x_{ic'gh'w'} \tag{2}$$

$$\sigma^{2}_{igs} = \frac{1}{\frac{C}{G_c}\cdot\frac{H\cdot W}{G_s}} \sum_{c'=1}^{\frac{C}{G_c}} \sum_{h'=1}^{\frac{H}{G_s}} \sum_{w'=1}^{\frac{W}{G_s}} \left(x_{ic'gh'w'} - \mu_{igs}\right)^{2} + \epsilon \tag{3}$$

c' represents the channel index inside the channel group, which is used to traverse the internal channels of the channel group w' and h' represents the position index within the space group, which is used to traverse the positions within each space group.

We further adjust each feature using the trainable parameters $\gamma \in R^{C}$ and $\beta \in R^{C}$ in the SGN layer. The parameter γ controls the normalized features, allowing the network to amplify or attenuate the values of the feature maps, thereby adjusting the importance of each feature. Meanwhile, β provides flexibility for the network to adjust the features after normalization, enabling it to recover some characteristics of the original data distribution. The normalized correlation weights are represented by $W_{\gamma\beta} \in R^{C}$, which indicate the relative importance of different feature maps.

$$W_{\gamma\beta} = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^{C}\gamma_j} + \frac{\beta_i}{\sum_{j=1}^{C}\beta_j}, i, j = 1,2,\cdots,C \tag{4}$$

The feature weights reweighted by $W_{\gamma\beta}$ are mapped to the range $(0,1)$ through a sigmoid function and controlled by a threshold. If the weight after Sigmoid mapping exceeds the threshold (set to 0.5 in this experiment), it is classified as an informative feature with a weight W_1. If the weight is below the threshold, it is considered a weak information bit with a weight W_2. The entire process of obtaining W can be expressed as follows:

$$W = Gate\left(Sigmoid\left(W_{\gamma\beta}\left(\mathrm{SGN}\left(\mathbf{X}, G_c, G_s, \gamma, \beta, \epsilon\right)\right)\right)\right) \tag{5}$$

Finally, we multiply feature X by W_1 and W_2 to obtain weighted features X_1 and X_2, where X_1 is the spatial content that is informative and expressive, and X_2 is the one that has little information and is considered redundant.

To reduce spatial redundancy and better utilize the features in X_2, we propose the Attention Reconstruct operation. This operation adds X_1 and X_2 elementwise, then reduces the feature dimension by half using a convolution layer. The resulting feature map is subsequently convolved again to produce a single-channel map that represents attention distribution. The Sigmoid function is applied to map this distribution to the range $(0,1)$, which indicates the weight of the features at each position.

$$A = \sigma\left(\mathrm{Conv}_2\left(\mathrm{Conv}_1\left(X_1 + X_2\right)\right)\right) \tag{6}$$

Finally, we perform channel segmentation $\mathrm{Split}(\cdot)$ on X_1 and X_2 respectively and then cross-multiply the attention weights to emphasize the information-rich areas. Finally, we restore the original feature dimension through the splicing operation to achieve effective integration of information.

$$Out_{final}^{S} = Cat\left(\text{Split}\left(X_1\right) \odot \ A, \text{Split}\left(X_2\right) \odot \ A\right) \tag{7}$$

After SGU processes the input feature X, we not only separate the features with large information content from the features with small information content but also reconstruct them to enhance the representative features and suppress the redundant features in the spatial dimension, thereby further optimizing the feature representation while maintaining the integrity of the information.

Frequency gating unit

To effectively leverage the information from different frequency components of the features, we introduce the Frequency Gating Unit (FGU), which employs a split-rearrangement-transformation fusion-rearrangement strategy. The structure of the FGU is illustrated in Fig. 3. FGU enhances the model's performance by splitting the input feature map into high-frequency and low-frequency components, which are then separately transformed, rearranged, and fused. For a given feature map $X \in \mathbb{R}^{N\times C\times H\times W}$, we first divide the map into two parts: a high-frequency component U and a low-frequency component L. The high-frequency component contains αC channels, and the low-frequency component contains $C - \alpha C$ channels, where $\alpha \in (0{,}1)$ determines the proportion of channels allocated to the high-frequency part.

$$U = X_{1:\alpha C}, L = X_{\alpha C+1:C} \tag{8}$$

Next, two independent 1×1 convolutional layers are applied to compress U and L with a compression ratio of $1/s$, where s is the compression factor (set to 2 in the experiment). This step aims to reduce the number of channels and, consequently, decrease the computational complexity in subsequent processing stages.

$$U' = \text{Conv}_{1\times 1}\left(U, \frac{\alpha C}{s}\right), L' = \text{Conv}_{1\times 1}\left(L, \frac{(1-\alpha)C}{s}\right) \tag{9}$$

Then, to break the local correlation between feature maps and promote the fusion of features, we use channel rearrangement technology to obtain U'' and L'' by rearranging the compressed high-frequency features U' and low-frequency features L'.

$$U'' = Perm\left(U'\right), L'' = Perm\left(L'\right) \tag{10}$$

$Perm$ represents the rearrangement operation, which is completed by randomly sorting the indexes of the channel dimensions of U' and L'.

The high-frequency features U'', obtained after segmentation, compression, and rearrangement, are input into the upper-layer transformation component. Since the high-frequency part typically contains detailed information from the image, we use group convolution[46] in the upper layer to capture local features more effectively, and pointwise convolution[47] to preserve the overall structural information. This operation can be expressed as follows:

$$Y_1 = \text{GC}\left(U'', G\right) + \text{PC}_1\left(U''\right) \tag{11}$$

Among these, $\text{GC}(\cdot)$ denotes the group convolution operation, where G represents the number of groups (set to 4 in the experiment), and $\text{PC}(\cdot)$ represents the pointwise convolution operation. The low-frequency part typically contains large-scale structural and color information of the image. Therefore, in the lower layer, we only use pointwise convolution to preserve the overall structural information and subsequently concatenate it with the original low-frequency feature map. This approach maintains the original structural information while introducing new transformation features. This operation can be expressed as follows:

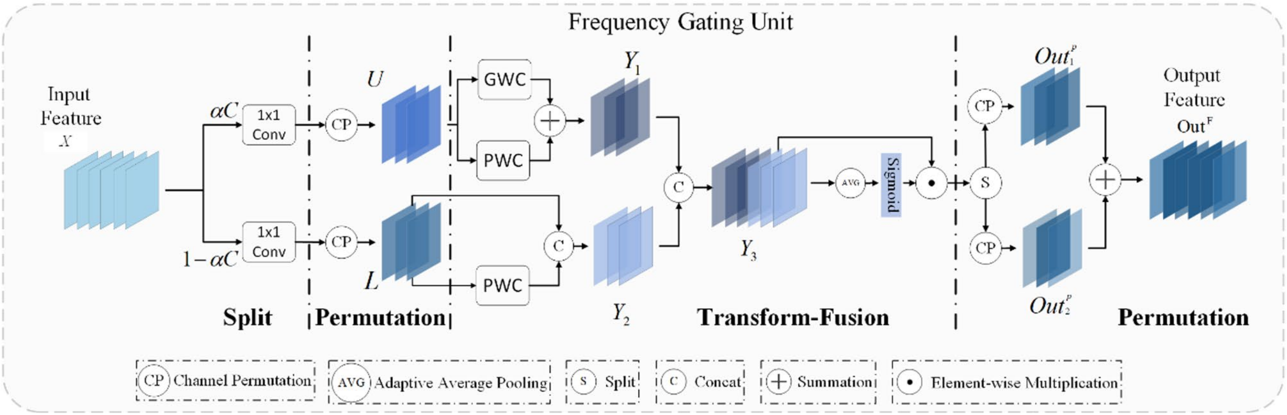

Fig. 3. Structure of FGU.

$$Y_2 = Cat\left(\mathrm{PC}_2\left(L''\right), L''\right) \tag{12}$$

Among these, $\mathrm{Cat}(\cdot)$ represents the concatenation operation. The high-frequency feature Y_1 and the low-frequency feature Y_2 are concatenated along the channel dimension to obtain the complete feature map Y_3. Adaptive average pooling is then applied to compress the spatial size of Y_3 to 1×1. The softmax function is used to calculate the importance weight W of each channel, and this weight is element-wise multiplied with the feature map Y_3 to obtain the weighted fused feature map Y_3'. The specific process is as follows:

$$\left\{\begin{array}{c} Y_3 = Cat\left(Y_1, Y_2\right) \\ W = softmax\left(\mathrm{AdaptiveAvgPool}_2\left(Y_3\right)\right) \\ Y_3' = W \odot Y_3 \end{array}\right. \tag{13}$$

Among these, $\odot$ represents element-wise multiplication. Finally, for the weighted fused feature map Y_3', we first perform a channel splitting operation $Split$, then rearrange the channels of the split Out_1^s and Out_2^s. Subsequently, Out_1^p and Out_2^p are added channel by channel for final fusion, which further enhances the interaction between features. The output is the final feature map Out_{final}^C. The specific process is as follows:

$$\left\{\begin{array}{c} Out_1^s, Out_2^s = Split\left(Y_3'\right) \\ Out_1^p = Perm\left(Out_1^s\right), Out_2^p = Perm\left(Out_2^s\right) \\ Out_{final}^C = Out_1^p + Out_2^p \end{array}\right. \tag{14}$$

After FGU processes the input feature X, we effectively utilize high-frequency and low-frequency features, adjust the weights of different channels through the channel attention mechanism, and further enhance the interaction between features through the final fusion step to reduce the redundancy of channel information.

Dual-axis multi-scale attention unit

In image processing, convolution kernels of varying sizes capture features at different scales. Smaller kernels, such as 7×7, are effective for capturing local details, while larger kernels, such as 11×11 and 21×21, capture more global information. By leveraging the attention mechanism to combine features from multiple scales, the model can learn a more comprehensive feature representation. The proposed Dual-axis Multi-scale Attention Unit (DMAU) extracts features through two parallel branches, each calculating multi-scale features along the horizontal and vertical directions, respectively. The DMAU structure is illustrated in Fig. 4. Each branch consists of three 2D convolution kernels of varying sizes, which encode multi-scale contextual information along the horizontal and vertical spatial dimensions.

The DMAU module first uses 7×7, 11×11, and 21×21 convolution kernels to perform multi-scale feature extraction on the features Out_{final}^S and Out_{final}^C obtained by the SGU and FGU modules in the horizontal direction, respectively. The 7×7 convolution kernel is used to extract local features, and the 11×11 and 21×21 convolution kernels are used to extract global information, which can be expressed as:

$$\left\{\begin{array}{l} F_{s_1} = Conv_{1\times 7}^{Row}\left(Out_{final}^S\right) \\ F_{s_2} = Conv_{1\times 11}^{Row}\left(Out_{final}^S\right) \\ F_{s_3} = Conv_{1\times 21}^{Row}\left(Out_{final}^S\right) \end{array}\right. \tag{15}$$

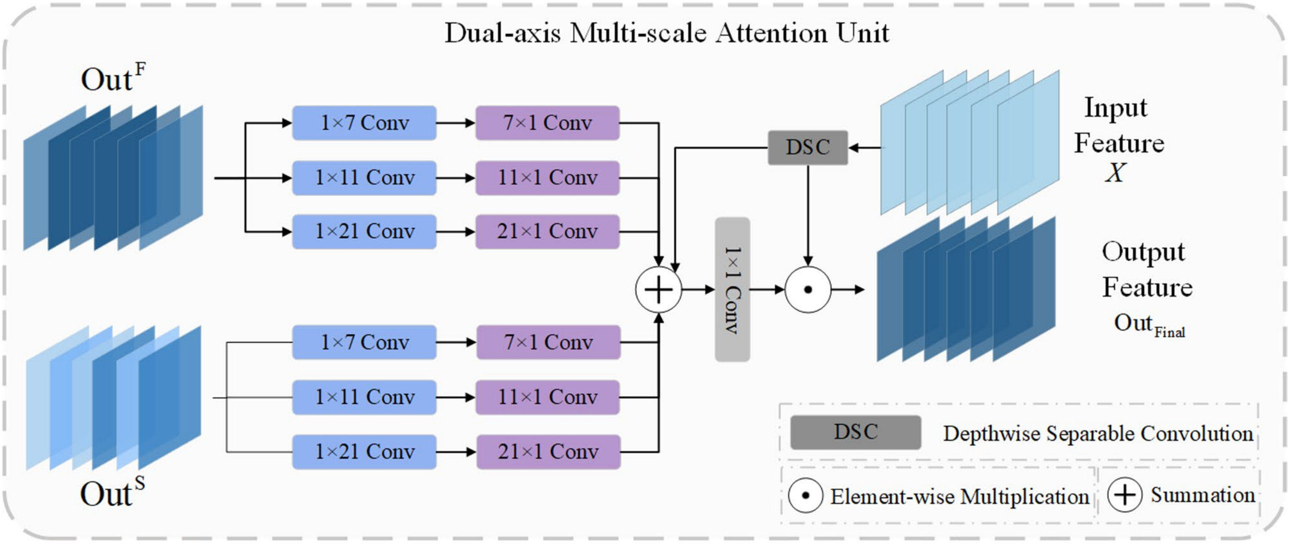

Fig. 4. DMAU structure diagram.

$$\left\{\begin{array}{l} F_{c_1} = Conv_{1\times\,7}^{Row}\left(Out_{final}^{C}\right) \\ F_{c_2} = Conv_{1\times\,11}^{Row}\left(Out_{final}^{S}\right) \\ F_{c_3} = Conv_{1\times\,21}^{Row}\left(Out_{final}^{S}\right) \end{array}\right. \tag{16}$$

$Conv_{\bullet}^{Row}\left(\bullet\right)$ represents convolution along the horizontal direction. After extracting the features in the horizontal direction, multi-scale convolution is performed along the vertical direction to extract the feature information in the vertical direction:

$$\left\{\begin{array}{l} F_{s_1}' = Conv_{7\times\,1}^{Col}\left(F_{s_1}\right) \\ F_{s_2}' = Conv_{11\times\,1}^{Col}\left(F_{s_2}\right) \\ F_{s_3}' = Conv_{21\times\,1}^{Col}\left(F_{s_3}\right) \end{array}\right. \tag{17}$$

$$\left\{\begin{array}{l} F_{s_1}' = Conv_{7\times\,1}^{Col}\left(F_{c_1}\right) \\ F_{c_2}' = Conv_{11\times\,1}^{Col}\left(F_{c_2}\right) \\ F_{c_3}' = Conv_{21\times\,1}^{Col}\left(F_{c_3}\right) \end{array}\right. \tag{18}$$

$Conv_{\bullet}^{Col}\left(\bullet\right)$ represents convolution along the vertical direction. The feature maps of different scales are fused to obtain a more comprehensive feature representation:

$$F = \ X + F_{s_1}' + F_{s_2}' + F_{s_3}' + F_{c_1}' + F_{c_2}' + F_{c_3}' \tag{19}$$

Among them, X is the result of the initial feature map X_{Input} after the depth-wise separable convolution. Use $1\times\ 1$ convolution to mix channels on the feature map F, adjust the channel relationship, and then perform convolution attention calculation with the input feature map X_{Input} to obtain the result:

$$O = Conv_{1\times\,1}\left(F\right)\odot X_{Input} \tag{20}$$

Through multi-scale feature extraction and fusion, the DMAU module is capable of capturing feature information at different scales, further enhancing the feature representation ability. The application of channel mixing and the convolutional attention mechanism further adjusts the relationship between local and global features of the feature map, enhancing the interaction among features.

Information enhancement module

In encoder-decoder architectures such as U-Net, maximum pooling is commonly employed during the downsampling stage to reduce the spatial resolution of feature maps, thereby increasing the receptive field. However, this operation can result in the partial loss of positional information, leading to information attenuation as it propagates through deeper layers, making it challenging to capture long-distance dependencies. To mitigate information loss during deep propagation, establish long-distance dependencies, and minimize the increase in the number of parameters, we designed the Information Enhancement Module (IEM). The IEM integrates deep convolution, multi-scale feature extraction, axial attention, and channel shuffling operations. The structure of the IEM is shown in Fig. 5.

To enhance feature information, we input the given feature map $X\in\ \mathbb{R}^{N\times\ C\times\ H\times\ W}$ into a specialized parallel convolution module. This module consists of three branches, each utilizing depthwise convolution, with kernel sizes of $3\times\ 3$, $5\times\ 5$ and $7\times\ 7$. These branches operate in parallel, extracting features at multiple spatial scales, and the resulting features are subsequently merged to provide a richer feature representation.

In each branch, a depthwise convolution is first performed, in which each input channel is convolved separately without considering the information of other channels. The advantage of this is that while reducing the number of parameters, the spatial relationship of the input data is retained:

$$\left\{\begin{array}{l} \mathrm{F}_3 = X * \mathrm{W}_{\mathrm{d}3} \\ \mathrm{F}_5 = X * \mathrm{W}_{\mathrm{d}5} \\ \mathrm{F}_7 = X * \mathrm{W}_{\mathrm{d}7} \end{array}\right. \tag{21}$$

$\mathrm{W}_{\mathrm{d}3}$, $\mathrm{W}_{\mathrm{d}5}$, $\mathrm{W}_{\mathrm{d}7}$ are the depth convolution weights of $3\times\ 3$, $5\times\ 5$, and $7\times\ 7$ convolution kernels respectively and $*$ represents multiplication operation.

After each branch performs the convolution operation, three feature maps of different scales, F_3, F_5, and F_7, are obtained. To integrate this multi-scale information, we residually connect these feature maps with the original feature map. This ensures that features from different scales complement each other, resulting in a richer and more comprehensive feature representation. The final enhanced feature map, F, is then obtained. The steps are as follows:

$$F = Cat\left(X, \mathrm{F}_3, \mathrm{F}_5, \mathrm{F}_7\right) \tag{22}$$

To fully leverage the information from different channels in the feature map F, we rearrange the channels and mix their information, ensuring that the data is more evenly distributed across the feature map. The rearranged features are then input into the column-axis attention module, where attention weights between Query, Key, and Value are computed. These weights are applied to the aggregate value vector to generate the feature map $\mathrm{F}_{\mathrm{col}}$. Subsequently, $\mathrm{F}_{\mathrm{col}}$ is passed through the row-axis attention to obtain the feature map F_{Axial}, which

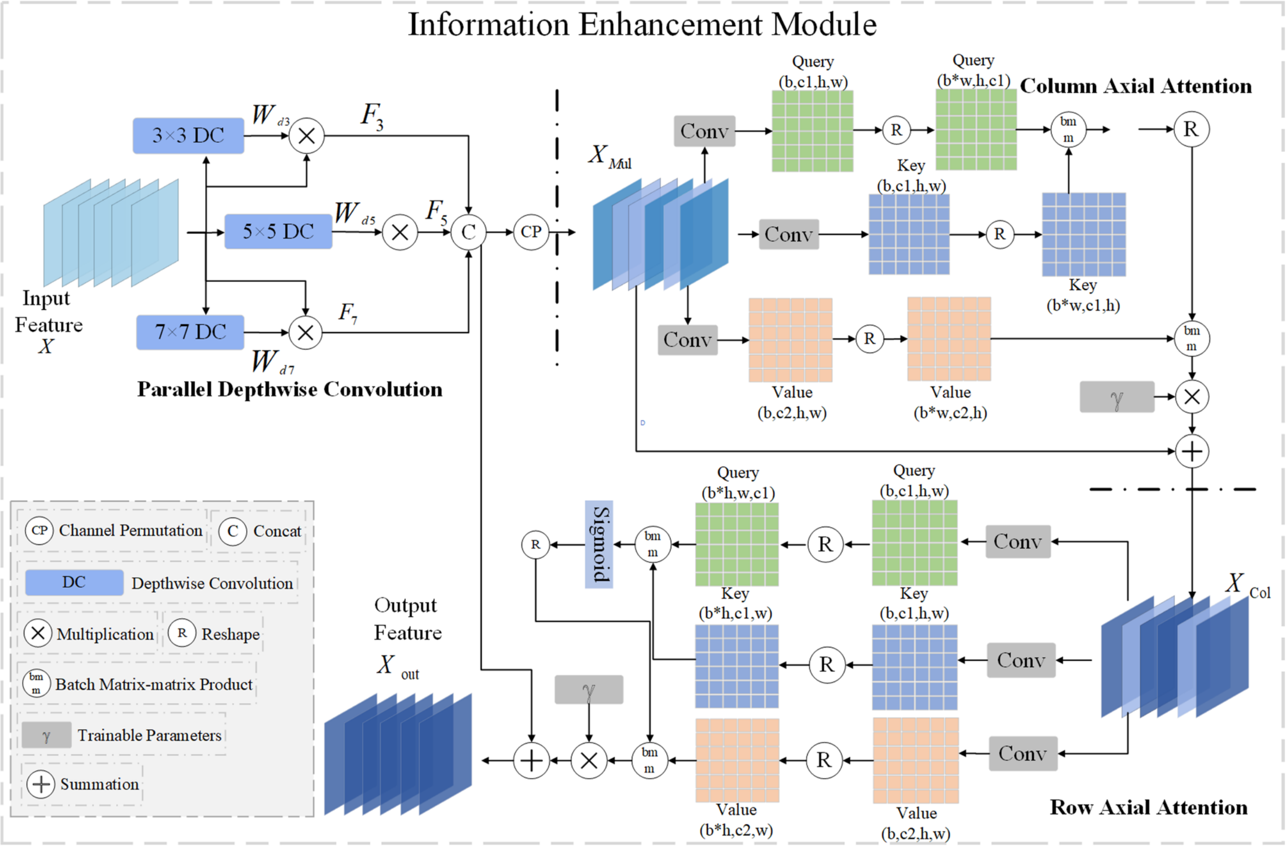

Fig. 5. IEM structure diagram.

undergoes the same processing steps. The axial attention mechanism helps establish long-distance dependencies by integrating information from each pixel position through multiple attention calculations, allowing the model to capture long-range dependencies across the entire feature map. The overall process is as follows:

$$\mathrm{F}_{Axial} = RAA\left(CAA\left(Perm\left(F, G\right)\right)\right) \tag{23}$$

$Perm(\cdot)$ represents channel rearrangement, $CAA(\cdot)$ represents Column Axial Attention, $RAA(\cdot)$ represents Row Axial Attention, and G represents the number of groups (set to 4 in the experiment). Finally, F_{Axial} is added to F to obtain the final information-enhanced feature map F_{final}

Experiments

Datasets

We conducted extensive experiments on two ischemic stroke datasets using single-modality MRI images to evaluate the effectiveness and reliability of our approach.The ATLAS dataset[48] comprises 239 T1-weighted brain images with corresponding lesion annotations provided by neurologists. The Ischemic Stroke Lesion Segmentation (ISLES 2022) challenge[49] includes 400 brain MRI images, with 250 annotated images released as training data and 150 unannotated images provided as test data. The ISLES 2022 challenge focuses on acute and subacute lesion segmentation, while the ATLAS dataset is derived from post-stroke MRI data and addresses the subacute and chronic stages of stroke. Clinically, these two datasets cover all stages of stroke, enabling a comprehensive evaluation of the generalization ability of the proposed method.

Implementation details

We implement SFMANet using PyTorch 2.0.1, training the model on a deep learning server equipped with an Nvidia GeForce RTX 4090. The Adam optimizer is employed to adjust network parameters, with k1 set to 0.9 and k2 set to 0.999. The ReduceLROnPlateau learning rate scheduler is applied during training, with the maximum Dice coefficient used as the performance metric. If the Dice coefficient on the training set does not improve for four consecutive epochs, the learning rate is reduced. The initial learning rate is set to 1e-4, and the network is trained for 50 epochs with a batch size of 16. The original size of the ATLAS dataset is $233 \times 197 \times 189$, while the ISLES 2022 dataset includes images of varying sizes, such as $115 \times 115 \times 25$, $112 \times 112 \times 73$, and $281 \times 352 \times 352$. For the 2D processing, the two datasets are sliced along the X-axis, and all resulting slices are resized to 256×256 resolution. Except for the SAN-Net and MDA-Net experimental data, which are the experimental data of the original paper, all other experimental data are the original data of this study.

During the training process, we randomly partitioned the two stroke lesion segmentation datasets at a ratio of 8:1:1. The training set accounted for 80% of the entire dataset, the validation set made up 10%, and the test set also constituted 10%.

Evaluation metrics

Medical image segmentation usually employs a variety of metrics to measure the performance of segmentation models, including the Dice Similarity Coefficient (DSC), Precision, Recall, F1 Score, and the Mean Intersection over Union (MIoU).

The DSC represents the degree of overlap between the prediction result $\mathrm{R_{pred}}$ and the ground truth $\mathrm{R_{gt}}$:

$$DSC = \frac{2\left|\mathrm{R_{pred}} \cap \mathrm{R_{gt}}\right|}{\left|\mathrm{R_{pred}}\right| + \left|\mathrm{R_{gt}}\right|} \tag{24}$$

Precision represents the proportion of samples that are truly positive among the samples predicted as positive by the model:

$$Precision = \frac{\mathrm{R_{gt}} \cap \mathrm{R_{pred}}}{\mathrm{R_{pred}}} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{25}$$

Recall represents the proportion of samples that are actually positive and are correctly predicted as positive by the model among all the samples that are actually positive:

$$Recall = \frac{\mathrm{R_{gt}} \cap \mathrm{R_{pred}}}{\mathrm{R_{gt}}} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{26}$$

The F1 Score represents the balance and comprehensive performance of the model between Precision and Recall:

$$\mathrm{F_1} = 2 \times \frac{\mathrm{Precision \times Recall}}{\mathrm{Precision + Recall}} \tag{27}$$

The MIoU (Mean Intersection over Union) is used to measure the degree of overlap between the predicted segmentation results and the ground truth labels:

$$MIoU = \frac{1}{\mathrm{k+1}} \sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{FN+FP+TP}} \tag{28}$$

Among them, the TP prediction value represents the prediction of true positives, the FP prediction value represents the prediction of false positives, and the FN prediction value represents the prediction of false negatives for a single pixel. The variable k represents the total number of predicted categories. Meanwhile, in order to measure the time complexity and space complexity of the model, in this study, "Parameters" is used to measure the space complexity of the model, and "FLOPs" (floating-point operations) and the number of images that the model can process per second, denoted as $\mathrm{M_{Speed}}$, are used to measure the time complexity of the model.

Evaluate metric comparisons

Comparative experiments were conducted on SFMANet and several other well-known and effective segmentation methods, including the classic convolution-based U-Net, Attention U-Ne[50] U-Net++, SAN-Net[51] MDA-Net[52] U2Net[53] Acc_UNet[54] HmsU-Net[39] MSCA-Net[55] NLIE-UNet[56] as well as advanced transformer-based methods such as TransUNet, SwinUNet, TransFuse[57] and Polyp-pvt[58]. All experiments were performed on the ATLAS and ISLES 2022 datasets, using five-fold cross-validation. The experiments were carried out on the same computational setup to ensure consistent and comparable results (Table 1).

For the ATLAS dataset, we compared several state-of-the-art stroke segmentation methods, including SAN-Net, MDA-Net, HmsU-Net, MSCA-Net and NLIE-UNet, etc.

The experimental results of SFMANet are presented at the bottom of the table, with the best results highlighted in bold. According to Table 2, the Dice score of SFMANet on ATLAS is 0.8365, which is 0.0251 higher than the suboptimal result obtained by Acc_unet under the same experimental environment. The Precision score is 0.1262 higher, and the MIoU score is 0.0433 higher, both of which are higher than the suboptimal results of all the comparison methods. However, the F1 score of SFMANet is lower than that of Acc_unet, and the number of parameters is higher.

As can be seen from Table 2, in terms of Dice, Recall, and MIoU, SFMANet obtained the highest scores on the ISLES 2022 dataset, which were 0.7767, 0.9071, and 0.6911 respectively. It was 0.0128, 0.0634, and 0.1138 higher than the second-ranked method respectively. On the ISLES 2022 dataset, although SFMANet did not achieve the best F1 score and Precision. MDA-Net achieved an F1 score of 0.7368, but its DSC was 0.7160; NLIE-UNet achieved a Precision of 0.8923, but its DSC and MIoU were significantly lower than those of SFMANet. The DSC of Acc_unet was 0.7639, but its F1 score was 0.6675. These results indicate that due to the calculation methods, there is a trade-off between these two metrics. In contrast, the F1 score and DSC of SFMANet were 0.7160 and 0.7767 respectively, which is a relatively balanced and high-level result.

Qualitative analysis

Figures 6 and 7 present examples of segmentation results for different methods on the ATLAS and ISLES 2022 datasets. The first column shows the input image, the second column displays the corresponding ground

Method	DSC	Precision	Recall	F1score	MIoU
Unet	0.6609	0.6242	0.7382	0.6518	0.5632
SAN-Net	0.4301	-	0.4316	0.4301	-
MDA-Net	0.5662	0.6436	0.5945	-	-
Acc_unet	0.8114	0.8031	0.8173	**0.8011**	0.7296
UNet++	0.6567	0.7026	0.6396	0.6437	0.5674
AttentionUNet	0.8100	0.8003	0.8189	0.7994	0.7273
SwinUNet	0.4615	0.5090	0.4749	0.3703	0.4609
TransUNet	0.5927	0.7021	0.5626	0.4945	0.5924
TransFuse	0.6080	0.6111	0.7111	0.6073	0.4978
HmsU-Net	0.7105	0.9016	0.9148	0.7005	0.6373
MSCA-Net	0.6970	0.8879	0.9251	0.6843	0.6209
NLIE-UNet	0.7644	0.9216	0.8947	0.7523	0.6720
Ours	**0.8365**	**0.9293**	**0.9372**	0.7408	**0.7729**

Table 1. Evaluation results of different models in ATLAS segmentation results.

Method	DSC	Precision	Recall	F1score	MIoU
UNet	0.5910	0.6182	0.6653	0.5515	0.4892
MDA-Net	0.7044	0.7530	0.7222	**0.7368**	-
Acc_unet	0.7639	0.8150	0.7379	0.6675	0.6675
UNet++	0.6980	0.7940	0.7028	0.6784	0.5954
U2Net	0.7325	0.7838	0.7062	0.7186	0.6401
AttentionUNet	0.7352	0.8001	0.7110	0.7240	0.6394
SwinUNet	0.5199	0.5949	0.5980	0.5031	0.4150
TransUNet	0.6272	0.7666	0.5842	0.6272	0.5330
TransFuse	0.6872	0.6739	0.7933	0.6868	0.6739
Polyp_PVT	0.7000	0.7054	0.7318	0.6888	0.5802
HmsU-Net	0.6597	0.8037	0.8844	0.6307	0.5679
MSCA-Net	0.7385	0.8633	0.8608	0.7302	0.6343
NLIE-UNet	0.7457	**0.8923**	0.8923	0.7170	0.6487
Ours	**0.7767**	0.8784	**0.9071**	0.7160	**0.6911**

Table 2. Evaluation results of different models in ISLES 2022 segmentation results.

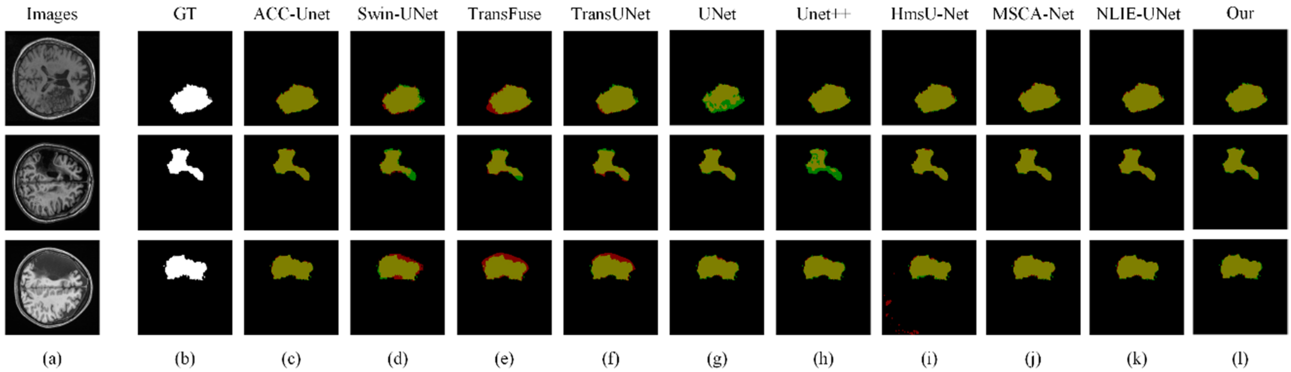

Fig. 6. Test results of the ATLAS dataset obtained using different models (yellow areas indicate true positive areas, red areas indicate true negative areas, and green areas indicate false negative areas.).

truth mask, and the final column depicts the segmentation result of SFMANet. To improve the visualization of segmentation results, yellow is used to highlight true positive areas, red represents true negative areas, and green indicates false negative areas. In most cases, leveraging the feature representation capabilities of deep learning, various methods can effectively segment lesions. However, upon closer inspection, it is evident that the proposed SFMANet outperforms others. Notably, SFMANet demonstrates superior performance by significantly reducing

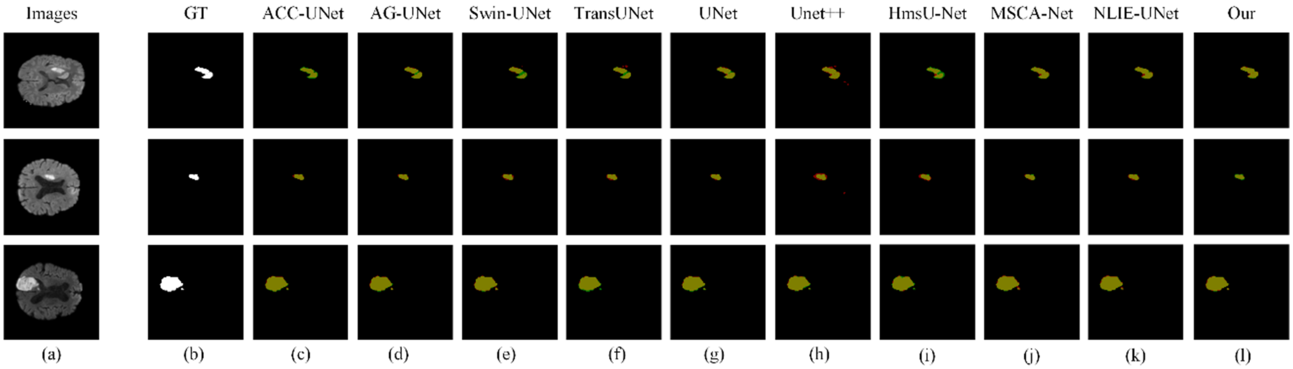

Fig. 7. Test results of the ISLES 2022 dataset obtained using different models (yellow areas indicate true positive areas, red areas indicate true negative areas, and green areas indicate false negative areas.).

UNet	SFGU	IEM	DMAU	DSC	Precision	Recall	F1 score	MIoU
√				0.5910	0.6182	0.6653	0.5515	0.4892
√	√			0.7359	0.8481	0.8903	0.6970	0.6479
√		√		0.7336	0.8637	0.8390	**0.7309**	0.6297
√	√	√		0.7660	0.8511	**0.9077**	0.7008	0.6797
√	√		√	0.7616	0.8620	0.8913	0.6885	0.6752
√	√	√	√	**0.7767**	**0.8784**	0.9071	0.7160	**0.6911**

Table 3. Ablation experiments of SFMANet on the ISLES 2022 dataset.

true negative (red) and false negative (green) areas. This improvement is also reflected in higher DSC, Precision, and Recall values. Such advancements are of considerable clinical importance, as missing lesion areas in a diagnosis could result in missed treatment opportunities, potentially causing greater harm to patient health.

Ablation experiment

To further validate the effectiveness of the key components in SFMANet, ablation studies were conducted on the ISLES 2022 and ATLAS datasets. U-Net served as the baseline for this experiment. By sequentially adding or replacing components in the original network, we primarily tested the following components: SFGU, IEM, and DMAU. Since the pre-input features of DMAU are the output features of SFGU, the DMAU unit could not be tested in isolation. As shown in Table 3, the DSC score for U-Net is 0.5910, and the addition of each component resulted in a substantial improvement (0.1426–0.1857). When SFGU and IEM components were added separately to the U-Net model, the improvement in DSC was similar, though each had distinct advantages in other metrics. Specifically, the Recall and MIoU scores of U-Net + SFGU were 0.0513 and 0.0182 higher than those of U-Net + IEM, respectively, while Precision and F1 scores were lower. U-Net + IEM, however, achieved the highest F1 score of 0.7309 in the experiment. Combining both SFGU and IEM with U-Net led to significant improvements in DSC, Recall, and MIoU, with the highest F1 score of 0.9077, suggesting that the two modules complement each other. When both SFGU and DMAU were added to U-Net, compared to U-Net + SFGU, all metrics improved (DSC: +0.0257, Precision: +0.0139, Recall: +0.0010, MIoU: +0.0273) except for the F1 score, indicating that DMAU also positively contributes to model performance. Adding all components to the baseline model resulted in substantial improvements across all metrics. While Recall and F1 score did not achieve the best results, other metrics were optimal, indicating that the SFMANet model, with minimal cost on the ISLES 2022 dataset, achieves superior performance.

Similarly, as shown in Table 4, the F1 score of the SFMANet model on the ATLAS dataset is much lower than that of UNet + IEM, but its other indicators are all superior to UNet + IEM. Therefore, SFMANet also has a very good performance improvement on the ATLAS dataset.

Complexity analysis

In this section, we analyzed the computational complexity of SFMANet and summarized the number of parameters (Parameters), FLOPs, and the number of images that the model can process per second $\mathrm{M}_{\mathrm{Speed}}$ f the methods mentioned in Sect. 4.4. We selected ATLAS as the experimental dataset, and the experimental results are shown in Table 5. As can be seen from Table 5 MDA-Net achieved the fastest image calculation speed, but its DSC was only 0.5662, which means the accuracy is too low. Although HmsU-Net obtained the lowest number of parameters, its $\mathrm{M}_{\mathrm{Speed}}$ and FLOPs were inferior to those of MSCA-Net, which were 12.71 and 395.91 respectively. MSCA-Net achieved good results in terms of both $\mathrm{M}_{\mathrm{Speed}}$ and FLOPs, but its DSC was only 0.6970, which is not suitable in the medical field where precise lesion processing is required.Therefore, under

UNet	SFGU	IEM	DMAU	DSC	Precision	Recall	F1 score	MIoU
√				0.6609	0.6242	0.7382	0.6518	0.5632
√	√			0.8147	0.8990	0.9142	0.7323	0.7673
√		√		0.8226	0.9293	0.9352	**0.8126**	0.7409
√	√	√		0.8274	0.9251	0.9401	0.7303	0.7628
√	√		√	0.8313	0.9245	**0.9416**	0.7351	0.7452
√	√	√	√	**0.8365**	**0.9293**	0.9372	0.7408	**0.7729**

Table 4. Ablation experiments of SFMANet on the ATLAS dataset.

Method	DSC	Parameters(M)	FLOPs(G)	M_{Speed}(imgs/s)
Unet	0.6609	31.04	546.46	229.28
SAN-Net	0.4301	132.11	33.61	-
MDA-Net	0.5662	29.00	-	**470.04**
Acc_unet	0.8114	16.77	511.64	192.14
UNet++	0.6567	36.62	1389.24	31.86
AttentionUNet	0.8100	34.88	664.85	153.66
SwinUNet	0.4615	27.17	60.34	170.05
TransUNet	0.5927	105.32	327.14	139.48
TransFuse	0.6080	26.25	317.34	81.64
HmsU-Net	0.6970	**9.00**	48.91	291.24
MSCA-Net	0.6970	27.70	**12.71**	395.91
NLIE-UNet	0.7644	9.95	20.20	50.09
Ours	**0.8365**	36.92	94.77	124.14

Table 5. The computational complexity of different models on the ATLAS dataset.

$\mathrm{SRU_{TH}}$	CRU_{α}	Evaluation indicators				
		DSC	Precision	Recall	F1score	MIoU
0.8	$\alpha = 0.25$	0.7443	0.8486	0.9104	0.6745	0.6589
	$\alpha = 0.50$	0.7560	0.8535	0.9143	0.6862	0.6701
	$\alpha = 0.75$	0.7549	0.8379	0.8912	0.6900	0.6672
0.5	$\alpha = 0.25$	0.7469	0.8560	0.9072	0.6798	0.6611
	$\alpha = 0.50$	**0.7767**	**0.8784**	0.9071	**0.7160**	**0.6911**
	$\alpha = 0.75$	0.7446	0.8418	0.8917	0.6798	0.6583
0.2	$\alpha = 0.25$	0.7285	0.8431	0.9089	0.6669	0.6402
	$\alpha = 0.50$	0.7640	0.8606	**0.9162**	0.6970	0.6787
	$\alpha = 0.75$	0.7514	0.8453	0.8890	0.6922	0.6607

Table 6. Experimental results of selecting different hyperparameters on the ISLES 2022 dataset.

the condition of compromising the FLOPs and $\mathrm{M_{Speed}}$ indicators, SFMANet sacrificed a small amount of the number of parameters and computational speed and achieved good segmentation performance.

Discussion on the setting of hyperparameters

In this section of the research, we analyzed the impact of the threshold $\mathrm{SRU_{TH}}$ of the SGU and the hyperparameter α of the FGU on the performance of the model. The number of groups in the grouped convolution and the compression ratio of the FGU are classic designs and will not be discussed here. The experiments in this section were conducted on the ISLES 2022 dataset, and the hyperparameter settings other than $\mathrm{SRU_{TH}}$ and α were the same as those in the previous experiments. The experimental results are shown in Table 6. The experimental results demonstrate that when both $\mathrm{SRU_{TH}}$ and α are set to 0.5, all the indicators except Recall achieve the optimal values, and the difference from the optimal value of Recall is less than 0.01.

Discussion

In recent years, deep learning algorithms, particularly U-Net and its variations, have made significant advancements in the field of medical image segmentation. This study introduces a spatial-frequency multi-scale attention network (SFMANet) designed for stroke lesion segmentation, addressing the challenges of low segmentation accuracy caused by irregular lesion areas and the similarity in signal strength between the lesion and healthy regions in existing models. In this network, we propose the spatial-frequency gating unit (SFGU), which effectively utilizes the redundant features generated due to the increase in the number of network layers. Through the gating mechanism, it refines the redundant features in the spatial and frequency domains and enhances the feature representation of the lesion area, aiming to solve the problem of blurred boundaries caused by the possible similar signal representations between the stroke area and the surrounding brain tissues. We propose the dual-axis multi-scale attention unit (DMAU), which combines the attention mechanism with dual-axis multi-scale feature extraction to better allocate the weights of global and local information and enhance the feature expression, so as to better locate the stroke lesion area. We design the information enhancement module (IEM). By enhancing the features through multi-scale depthwise separable convolutions, it mitigates the feature loss caused by the pooling operation and uses axial attention to establish long-range dependencies. The performance of our proposed segmentation algorithm is validated using the ATLAS and ISLES 2022 medical image datasets. We compare the segmentation results of other models with those of our method on both datasets. SFMANet demonstrates superior segmentation performance, particularly in handling weak edges in medical images (Figs. 6 and 7). The quantitative results of the experiments, presented in Tables 1 and 2, show that SFMANet remains highly competitive compared to classic and newer methods. Specifically, the DSC for the two datasets reached 0.8365 and 0.7767, respectively; MIoU reached 0.7729 and 0.6911, respectively; Precision reached 0.9293 and 0.8784, respectively; and Recall reached 0.9372 and 0.9071, respectively. Of course, this algorithm still has certain limitations. Although the segmentation results of this model are quite competitive, the model involves a large number of parameters and FLOPs, and the image processing is relatively slow, still requiring a great deal of computational work (Table 5). In addition, the model proposed here is still based on 2D images, and the segmentation of 2D medical images may lead to the loss of some 3D spatial information. In the future, we will explore model compression while retaining competitive prediction results, and expand the model to the 3D space, making full use of the 3D spatial information to segment 3D medical image data.

Conclusion

Edge loss and low segmentation accuracy are major challenges in medical image segmentation. To address these issues, this paper introduces a spatial-frequency multi-scale attention network (SFMANet) for stroke lesion segmentation. First, a traditional convolutional structure is employed as the encoder. Next, the Information Enhancement Module (IEM) is utilized to improve the overall image features during the encoding process. Then, the Spatial-Frequency Gating Unit (SFGU) and Dual-Axis Multi-Scale Attention Unit (DMAU) are applied to refine features and enhance edge detection during the skip connections. Finally, a conventional decoder is used to complete the decoding process. Through multiple experiments on the ATLAS and ISLES 2022 datasets, SFMANet demonstrates superior performance in medical image segmentation tasks.

Data availability

All of the datasets utilized in this paper are publicly available datasets, namely the ATLAS dataset and the ISLES 2022 dataset.The ATLAS dataset is available at: https://doi.org/10.1038/s41597-022-01401-7The ISLES 2022 dat aset is available at: https://doi.org/10.1038/s41597-022-01875-5.

Received: 31 January 2025; Accepted: 3 July 2025
Published online: 08 July 2025

References

1. Shelhamer, E., Long, J. & Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 640–651. https://doi.org/10.1109/tpami.2016.2572683 (2017).
2. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv* abs/1505.04597 (2015).
3. Li, X. et al. Deep learning attention mechanism in medical image analysis: basics and beyonds. *Int. J. Netw. Dyn. Intell.*, 93–116 (2023).
4. Mo, Y., Wu, Y., Yang, X., Liu, F. & Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **493**, 626–646 (2022).
5. Vaswani, A. et al. in *Neural Information Processing Systems.*
6. Liu, Z. et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002 (2021)., 9992–10002 (2021). (2021).
7. Chen, J. et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *ArXiv* abs/2102.04306 (2021).
8. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2020).
9. Pan, S., Liu, X., Xie, N. & Chong, Y. EG-TransUNet: A transformer-based U-Net with enhanced and guided models for biomedical image segmentation. *BMC Bioinformatics* **24**, 85 (2023).
10. Lin, A. J. et al. DS-TransUNet: dual Swin transformer U-Net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **71**, 1–15 (2021).
11. Hatamizadeh, A., Yang, D., Roth, H. R. & Xu, D. U. N. E. T. R. Transformers for 3D Medical Image Segmentation. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1748–1758 (2021)., 1748–1758 (2021). (2022).
12. Xie, Q., Chen, Y., Liu, S. & Lu, X. SSCFormer Revisiting ConvNet-Transformer hybrid framework from Scale-Wise and Spatial-Channel-Aware perspectives for volumetric medical image segmentation. *IEEE J. Biomed. Health Inform.* **28**, 4830–4841 (2024).
13. Ho, J., Kalchbrenner, N., Weissenborn, D. & Salimans, T. Axial Attention in Multidimensional Transformers. *ArXiv* abs/1912.12180 (2019).
14. Weili, J. et al. Ori-Net: Orientation-guided neural network for automated coronary arteries segmentation. *Expert Syst. Appl.* **238**, 121905. https://doi.org/10.1016/j.eswa.2023.121905 (2024).
15. Qiangguo, J. et al. Domain adaptation based self-correction model for COVID-19 infection segmentation in CT images. *Expert Syst. Appl.* **176**, 114848. https://doi.org/10.1016/j.eswa.2021.114848 (2021).
16. Zixuan, W. et al. A two-stage CNN method for MRI image segmentation of prostate with lesion. *Biomed. Signal Process. Control.* **82**, 104610. https://doi.org/10.1016/j.bspc.2023.104610 (2023).

17. Yuqing, Z., Yutong, H. & Jianxin, Z. MAU-Net: mixed attention U-Net for MRI brain tumor segmentation. *Math. Biosci. Eng.* **20**, 20510–20527. https://doi.org/10.3934/mbe.2023907 (2023).
18. Francis Jesmar, P. M. S3AR U-Net A separable squeezed similarity attention-gated residual U-Net for glottis segmentation. *Biomed. Signal Process. Control.* **92**, 106047. https://doi.org/10.1016/j.bspc.2024.106047 (2024).
19. Ibtehaz, N., Rahman, M. S. & MultiResUNet Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks: Official J. Int. Neural Netw. Soc.* **121**, 74–87 (2019).
20. Anand, V., Gupta, S., Koundal, D. & Singh, K. Fusion of U-Net and CNN model for segmentation and classification of skin lesion from dermoscopy images. *Expert Syst. Appl.* **213**, 119230 (2023).
21. Evans Kipkoech, R., Qin, Noor, B., Rehan, R., Muhammad Shehzad, H. & GAIR-U-Net 3D guided attention inception residual u-net for brain tumor segmentation using multimodal MRI images. *J. King Saud Univ. - Comput. Inform. Sci.* **36**, 102086. https://doi.org/10.1016/j.jksuci.2024.102086 (2024).
22. Roba, G., Hoda, B. & Mayada, H. GAU U-Net for multiple sclerosis segmentation. *Alexandria Eng. J.* **73**, 625–634. https://doi.org/10.1016/j.aej.2023.04.069 (2023).
23. Ouyang, J., Liu, S., Peng, H., Garg, H. & Thanh, D. N. H. LEA U-Net: a U-Net-based deep learning framework with local feature enhancement and attention for retinal vessel segmentation. *Complex. Intell. Syst.* **9**, 6753–6766. https://doi.org/10.1007/s40747-023-01095-3 (2023).
24. Hao, D. & Li, H. A graph-based edge attention gate medical image segmentation method. *IET Image Proc.* **17**, 2142–2157 (2023).
25. Kumar, A. et al. CSNet: A new deepnet framework for ischemic stroke lesion segmentation. *Comput. Methods Programs Biomed.* **193**, 105524 (2020).
26. Kumar, A., Ghosal, P., Kundu, S. S., Mukherjee, A. & Nandi, D. A lightweight asymmetric U-Net framework for acute ischemic stroke lesion segmentation in CT and CTP images. *Comput. Methods Programs Biomed.* **226**, 107157 (2022).
27. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *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, S…* 11045, 3–11 (2018).
28. Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M. & Asari, V. K. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. *ArXiv* abs/1802.06955 (2018).
29. Valanarasu, J. M. J., Sindagi, V. A., Hacihaliloglu, I. & Patel, V. M. KiU-Net: Towards Accurate Segmentation of Biomedical Images using Over-complete Representations. *ArXiv* abs/2006.04878 (2020).
30. Labbihi, I. et al. Hybrid 3D medical image segmentation using CNN and frequency transformer fusion. *Arab. J. Sci. Eng.* https://doi.org/10.1007/s13369-024-09602-5 (2024).
31. Cao, H. et al. 205–218 (Springer Nature Switzerland).
32. Jiang, S., Li, J. J. & Hua, Z. GR-Net: gated axial attention ResNest network for polyp segmentation. *Int. J. Imaging Syst. Technol.* **33**, 1531–1548. https://doi.org/10.1002/ima.22887 (2023).
33. Zhang, H. et al. ResNeSt: Split-Attention Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2735–2745 (2020)., 2735–2745 (2020). (2022).
34. Azad, R., Jia, Y., Aghdam, E. K., Cohen-Adad, J. & Merhof, D. Enhancing Medical Image Segmentation with TransCeption: A Multi-Scale Feature Fusion Approach. *ArXiv* abs/2301.10847 (2023).
35. Wang, L. et al. Multiple instances focused Temporal action proposal generation. *Neurocomputing* **538**, 126025 (2023).
36. Xu, Z. et al. Collaborative attention guided Multi-Scale feature fusion network for medical image segmentation. *IEEE Trans. Netw. Sci. Eng.* **11**, 1857–1871 (2024).
37. Yan, H., Xie, J., Zhu, D., Jia, L. & Guo, S. MSLF-Net: A Multi-Scale and Multi-Level Feature Fusion Net for Diabetic Retinopathy Segmentation. *Diagnostics* 12 (2022).
38. Ao, Y. et al. MS-TCNet: an effective Transformer-CNN combined network using multi-scale feature learning for 3D medical image segmentation. *Comput. Biol. Med.* **170**, 108057 (2024).
39. Fu, B. et al. HmsU-Net: A hybrid multi-scale U-net based on a CNN and transformer for medical image segmentation. *Comput. Biol. Med.* **170**, 108013 (2024).
40. Shaker, A. M. et al. UNETR++: Delving into efficient and accurate 3D medical image segmentation. *IEEE Trans. Med. Imaging* **43**, 3377–3390. https://doi.org/10.1109/TMI.2024.3398728 (2024).
41. Huang, X. et al. An effective transformer for 2d medical image segmentation. *IEEE Trans. Med. Imaging.* **42**, 1484–1494 (2022).
42. Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. Squeeze-and-Excitation Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141 (2017)., 7132–7141 (2017). (2018).
43. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. CBAM: Convolutional Block Attention Module. *ArXiv* abs/1807.06521 (2018).
44. Ouyang, D. et al. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. *ICASSP –2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (2023)., 1–5 (2023). (2023).
45. Li, J., Wen, Y., He, L. & SCConv Spatial and Channel Reconstruction Convolution for Feature Redundancy. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6153–6162 (2023)., 6153–6162 (2023). (2023).
46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM.* **60**, 84–90. https://doi.org/10.1145/3065386 (2017).
47. Szegedy, C. et al. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1–9.
48. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data.*
49. Petzsche, M. R. H. et al. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. (2022).
50. Oktay, O. et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018).
51. Yu, W., Huang, Z., Zhang, J. & Shan, H. SAN-Net: learning generalization to unseen sites for stroke lesion segmentation with self-adaptive normalization. *Comput. Biol. Med.* **156**, 106717 (2023).
52. Iqbal, A. & Sharif, M. MDA-Net: multiscale dual attention-based network for breast lesion segmentation using ultrasound images. *J. King Saud University-Computer Inform. Sci.* **34**, 7283–7299 (2022).
53. Qin, X. et al. U2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recogn.* **106**, 107404 (2020).
54. Ibtehaz, N. & Kihara, D. in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 692–702 (Springer).
55. Sun, Y. et al. MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation. *Pattern Recogn.* **139**, 109524 (2023).
56. Wan, L. et al. Dynamic neighbourhood-enhanced UNet with interwoven fusion for medical image segmentation. *The Visual Computer*, 1–19. https://doi.org/10.1007/s00371-025-03832-w (2025).
57. Zhang, Y., Liu, H. & Hu, Q. in *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1*, proceedings, Part I 24. 14–24 (Springer). (2021).
58. Dong, B. et al. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932* (2021).

Author contributions

Methodology, H.L. and J.W.; software, J.W and L.W.; validation, J.W and L.W; resources, H.L. and Y.Z.; funding acquisition, H.L. and Y.Z.; data curation, J.W and L.W.; writing-review and editing, H.L. and J.W. All authors

have read and agreed to the published version of the manuscript.

Funding

This work was supported by Key R&D program of Shanxi Province, No. 202102020101009 and Fundamental Research Program of Shanxi Province, No.202403021211091. This work was also supported by the Natural Science Foundation of Liaoning Province, No. 2022-KF-22-13.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.L.