scientific reports

OPEN

Cross-modal dual-domain bi-direction feature interaction network for medical imaging semantic segmentation

Tao Zhou[1,3], Qitao Liu[1,3]✉, Ke Song[1,3], Wenwen Chai[1,3], Kaixiong Chen[1,3] & Huiling Lu[2]

In the cross-modal medical image segmentation method, it is easy to ignore the dependence between spatial features and frequency features, and fine-grained frequency features are not fused effectively. To solve the above problems, this paper proposes a cross-modal segmentation network DBW-Net. The main innovation work are as follows: Firstly, a cross-modal dual-domain bi-direction feature interaction segmentation network DBW-Net is designed. there are 3 encoders and 1 decoder, the 3 encoders are used to extract the features of PET/CT, PET and CT respectively. Secondly, a Cross-Modal Feature Extractor "from frequency to spatial " (CMFE(F-> S)) is designed in the encoder. The module converts the spatial map into multiple spectral maps by 2D Discrete Cosine Transform (2D DCT). Multi-frequency cross-dimension attention is used to capture the correlation among multiple spectral maps feature in different dimension. so as to generate a refined frequency attention map. This module uses the refined frequency attention map to enhance modal feature and fuse cross-modal interaction, and completes the recalibration about input feature map. Thirdly, a Cross-modal Feature Coupler "from spatial to frequency" (CMFC(S->F)) is designed in the bottleneck layer. The module maps multi-modal information to the spatial and frequency domain through the spatial-frequency feature extractor, Cross-domain coupled attention is used to fuse the semantic gap between multi-modal fine-grained frequency features and spatial features. Finally, in order to verify the effectiveness of the proposed method, experiments are carried out on the clinical multi-modal lung tumor medical image dataset and the Brats2019 brain tumor public dataset. The experimental results show that for lung tumor segmentation, the Miou, Dice, Voe, Rvd and Recall are increased by 3.02%, 2.32%, 4.66%, 2.63% and 4.16%, respectively. For brain tumor segmentation, the Miou, Dice, Voe, Rvd, Recall are increased by 3.06%, 2.31%, 4.68%, 2.64%, 5.76%, respectively. It shows that the model for complex shape lesion segmentation, has high precision and relatively low redundancy. It significantly improves the segmentation accuracy and robustness of the lesion area, and provides technical support for accurate identification and diagnosis of early lesions.

Keywords Medical image segmentation, U-Net, Discrete cosine transform, Discrete wavelet transform, Cross-modal interaction

Cancer, also known as malignant tumor, is a general term for all kinds of malignant tumors, and is one of the main diseases that endanger human life and health. In this background, the importance of medical image analysis is become more and more significance. It is a key means to realize the early detection of malignant tumors, thus providing the possibility to extend the life of patients[1]. With the continuous development of medical image technology, multi-modal medical images are gradually become research hotspot in the medical image analysis field[2]. According to the principle of imaging, medical images are divided into anatomical images and functional images. Anatomical images, such as high-resolution gray-scale magnetic resonance imaging (MRI) and computed tomography (CT) images, can clearly show the anatomical structure about tissues and organs. Functional images, such as pseudo-color positron emission tomography (PET) and single photon emission computed tomography (SPECT) images, accurately reflect functional information such as metabolic changes.

[1]School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China. [2]School of Medical information & Engineering, Ningxia Medical University, Yinchuan 750004, China. [3]Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China. ✉email: liuqitao@stu.nmu.edu.cn

Positron Emission Tomography/Computed Tomography (PET/CT) is an important multi-modal medical image. It combines the high sensitivity of PET images in tumor detection with the advantages of anatomical details of CT images, it can effectively utilize the complementarity of different modality medical images in lesion information expression, and provide more comprehensive support for medical diagnosis and analysis.

Current multi-modal medical image segmentation methods realize feature complementation based on segmentation tasks by fusion the modal information, and complete the accurate identification of the lesion area. Zhou et al.[3] proposed an interactive segmentation network Guide-YNet, which uses the lesion location information of PET images for feature guided learning, considering the characteristics of lesions in different modal medical images. Fu et al.[4] introduced a framework based on deep learning, which used the multi-modal spatial attention module to automatically learn the spatial regions related to tumors, and then performed segmentation tasks on lesion regions. Kumar et al.[5] to quantify the importance of each modal feature at different spatial locations, using a supervised convolution neural network for feature learning. Fan et al.[6] proposed a Transformer backbone network, dual-flow multi-modal features fusion by using deformable cross-attention fusion module to continuous matching features. Hou et al.[7] introduced a lightweight diffraction module, which successfully avoided the limitation of receptive field in traditional convolution operation, and designed a modal fusion diffraction network MFD-Net for automatic and accurate brain tumor segmentation. Gao et al.[8] proposed a semantic segmentation network based on global features, which can effectively aggregate representative cross-modal features for multi-modal semantic segmentation. The above shows that the combination of multi-modal images can obtain richer disease semantic information.

However, traditional multi-modal segmentation methods mostly focus on the extraction and fusion of spatial features. Lesion areas in medical images often show complex morphological structures and fuzzy boundaries, and it is difficult to capture the global semantic information and the complementary relationship between modals by using spatial domain method. The low frequency information of medical image reflects the general contour and structure of the image, and provides approximate features. High-frequency of medical image contains textures and details, capturing microscopic features. In order to better show the characteristics of frequency domain analysis in medical images, a space-frequency domain comparison experiment of multi-mode lung tumor medical images are done, and the specific results are shown in Fig. 1.

Figure 1 shows the differences of multi-modal lung tumor medical image (CT, PET, PET/CT) in spatial domain and frequency domain. The first column is the original spatial domain image, which can directly present the shape, boundary, density change, etc. of the tumor in the image. The second column is the low-frequency image, which corresponds to the region where the image brightness changes gently, preserving the main background and large-scale structure of the image, such as the contour of the tumor or the tissue distribution. The third column is the high-frequency image, which corresponds to the regions where the image brightness changes drastically, emphasizing the boundary features of the object, such as the boundaries of the tumor and the details of the tiny structures. The fourth column is the Fourier spectral distribution, which reveals the distribution pattern of spatial features in frequency of lung tumors medical.

As can be seen from Fig. 1, the boundary, texture and background of medical images have clear distribution characteristics in different frequency components of the frequency domain, which makes frequency domain analysis more advantageous in dealing with complex lesion structures. Therefore, for multi-modal medical images, how to effectively utilize the dependence between multi-modal spatial features and frequency domain

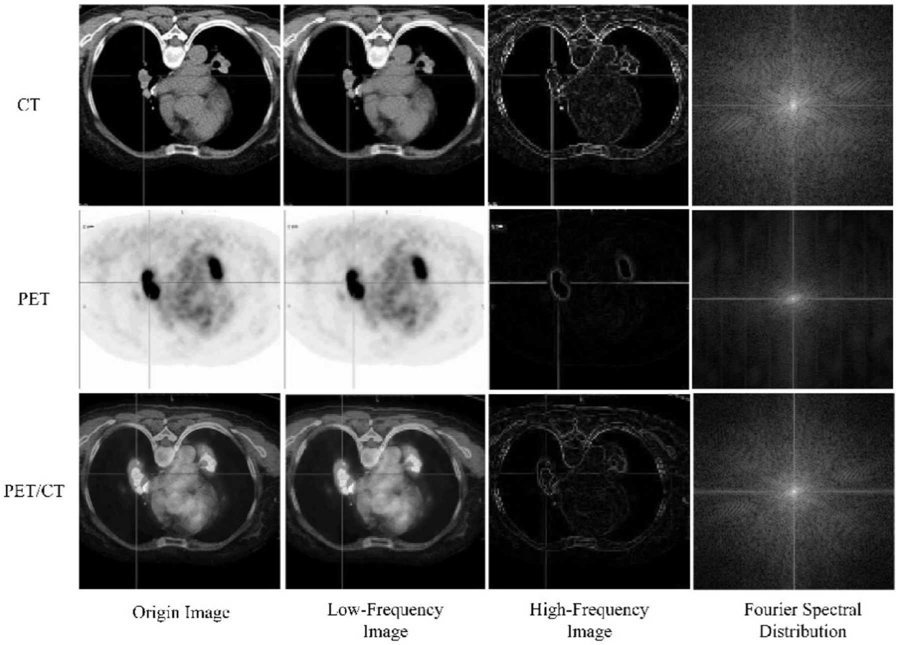

Fig. 1. Comparison map of multi-modal lung tumor medical images in space-frequency domain.

information, and realize the efficient fusion of fine-grained frequency domain features for information complementarity, improving the feature extraction ability of the network. this paper proposes DBW-Net, a cross-modal dual-domain bi-direction feature interaction semantic segmentation network. The main contributions are as follows:

1. From the perspective of frequency domain and space domain, a cross-modal dual-domain bi-directional feature interaction semantic segmentation network DBW-Net is proposed, which is "from frequency domain to space domain" and "from space domain to frequency domain". There are 3 encoders and 1 decoder, the 3 encoders are used to extract the features of PET/CT, PET and CT respectively. The network makes full use of the modal information of PET and CT branches to supplement the information of the main branch of PET/CT, enhances the feature expression ability, and realizes more accurate multi-modal medical image segmentation.
2. A cross-modal feature extractor "from frequency domain to spatial domain" (CMFE(F-> S)) is designed in the encoder. The module effectively utilizes the dependence of multi-modal spatial features on frequency domain information. The refined frequency-domain attention maps are used to apply complementary to the spatial input feature map, thereby enhancing the expressive abilities of the learned features and accurately locating the object of interest, resulting in better performance with fewer layers.
3. Aiming at the problem of effective fusion of multi-modal spatial domain features and fine-grained frequency domain features, a cross-modal feature coupler "from spatial to frequency" (CMFC(S-> F)) is designed in the bottleneck layer. By effectively bridging the semantic gap between multi-modal image fine-grained frequency domain features and spatial domain features, this module generates more powerful hybrid domain features. Through fully integrated multi-modal cross-domain information, significantly improve model accuracy and reliability in tumor segmentation task.

Related work

Single-modal semantic segmentation method refers to the use of deep learning methods to segment tissues or lesions by relying on a certain type of medical image data (such as CT or MRI). Ronneberger et al.[9] proposed a U-shaped structure network to extract medical image features. Zhang et al.[10] proposed a lightweight Residual transformer UNet (MRC-TransUNet) method for medical image segmentation tasks. Yuan et al.[11] proposed a multi-scale context fusion method based on U-shape encoder-decoder, which enriched the extracted semantic features and detailed information and improved the segmentation ability of the lesion area. Zhu et al.[12] proposed a Self-Regularized UNet method to reduce feature redundancy through feature distillation and improve the network's ability to extract effective features. Zhang et al.[13] proposed a meningioma feature extraction model that combines radiomics and deep learning methods to improve the accuracy of brain tumor grading.

Multi-modal semantic segmentation refers to the technology of combining data from different medical imaging modalities (such as CT and PET) for semantic segmentation. Multi-modal medical image semantic segmentation method complements and enhances the lesion information of different modalities by fully interacting with different modal images, so as to achieve the purpose of accurate segmentation lesions. Zhou et al.[14] proposed a Transformer segmentation model for cross-modal, cross-scale and cross-dimensional PET/CT images, which captured the effective information of each dimension of cross-modal images from the perspective of modality, scale and dimension. Lin et al.[15] proposed a clinical knowledge-driven brain tumor segmentation model, which designed a modal-related cross-attention module and a Trans&CNN feature calibration module to perform modal feature enhancement and cross-modal interactive fusion. Xu et al.[16] designed a cross-modal feature fusion strategy of dual-space graphs in the bottleneck layer of the U-Net framework network. From the perspective of global and semantic integration of multi-modal deep features, effectively obtain rich context information, and alleviate the feature differences between different modal. Diao et al.[17] proposed a framework called MIFPN, which enhances model robustness by integrating modal invariance and modal specific information. Wang et al.[18] proposed an improved network structure MPA-Net based on multi-modal parallel attention mechanism, which effectively fused high-dimensional features and low-dimensional features of images for segmentation tasks. However, the existing multi-modal image segmentation networks do not fully consider the information complementarity between different modal images.

Attention mechanism is a method that imitates human visual and perceptual ability, which aims to let the neural network focus on the most relevant part of the data to the task, so as to improve the performance of the model. Wang et al.[19] proposed an efficient Channel Attention module (ECA), which generates channel attention through one-dimensional convolution determined by nonlinear mapping of channel dimensions. Misra et al.[20] proposed a new method with multi-branch structure to generate cross-dimensional attention by capturing the attention weights of three dimensions. Hou et al.[21] proposed a coordinated attention method that embeds location information into channel attention, which encodes the resulting feature map into a pair of direction-aware and location-sensitive attention maps. Zhang et al.[22] proposed ResNeSt architecture, which integrates channel attention represented by multiple paths into a single unified segmentation attention block to improve network segmentation performance. Ouyang et al.[23] proposed a multi-scale attention module (EMA), which encodes global information and interacts across dimensions to capture pixel-level refined features. Yu et al.[24] proposed a multi-dimensional collaborative attention MCA, which can infer channel, height, and width attention simultaneously, and introduces almost no additional overhead. Zhou et al.[25] proposed Cross-modal Feature Fusion (CFF) and Adaptive Attention Fusion (AAF), which is used to fuse and enhance the complementary information of the three modalities. However, the above methods ignore the importance of frequency domain information, especially the dependence of spatial features on frequency domain information in multi-modal medical images. Therefore, this paper proposes a cross-modal feature extractor "from frequency domain to spatial domain" (CMFE(F-> S)) in the encoder, which aims to compensate multi-modal spatial features by refined frequency-domain attention map, highlighting the lesion area to enhance the interest object representation.

Frequency analysis in an image is a method to transform an image from the “spatial domain” to the “frequency domain”, and to analyze the image features in the frequency domain. At this point, the pixels in the traditional image no longer represent the spatial position of the image, but the frequency characteristics of the image brightness change. Common methods for frequency analysis include discrete cosine transform (DCT) and discrete wavelet transform (DWT). Among them, DCT is widely used in digital image and video processing because of its excellent energy concentration. In particular, DCT can represent most of the energy of the signal with fewer coefficients, thus achieving efficient compression while ensuring high quality. This property is very suitable for representing channel information in scalar form in the channel attention mechanism. As a powerful image transformation tool, DWT has shown unique advantages in deep learning, especially in processing low-resolution images and detecting small objects. DWT can provide multi-scale decomposition ability and effectively enhance the feature representation ability of the network. For the above reasons, recently, the introduction of frequency analysis into CNN-based medical image segmentation tasks has become a research hotspot. Huang et al.[26]proposed a frequency domain learning method called FDAM, which uses intra-class frequency relationships to retain valuable frequency features for medical image segmentation and classification tasks. Azad et al.[27] proposed a frequency recalibrated U-Net (FRCU-Net) for medical image segmentation to segment skin lesions. Tang et al.[28] proposed tKFC-Net to introduce frequency representation into convolutional neural networks. The proposed model combines frequency and spatial domain feature representations together for medical image segmentation tasks. Mu et al.[29] proposed a frequency-domain attention-guided Cascade U-Net (FACU-Net) for spatially tunable segmentation of vascular systems. Chen et al.[30] introduced Spatial Frequency Enhanced Network (SFE-Net), which promotes the consistency of features in the nucleus by fusing spatial frequency features to achieve accurate nucleus segmentation. However, the above methods cannot effectively solve the problem of effective fusion of multi-modal spatial domain features and fine-grained frequency domain features in multi-modal segmentation tasks. Therefore, this paper designs a cross-modal feature coupler “from spatial to frequency” (CMFC(S-> F)) in the bottleneck layer, which effectively bridges the semantic gap between multi-modal fine-grained frequency domain features and spatial domain features. This module fully combines multi-modal cross-domain information, enables the model to establish a more effective connection between functional and anatomical information, and can improve the detection and segmentation accuracy of tumor regions.

Method

The structure of DBW-Net, a cross-modal two-domain bidirectional feature interactive semantic segmentation network proposed in this paper, is shown in Fig. 2. Firstly, in order to make full use of the complementary information of different modal images, 3 encoders are designed to extract the image features of PET, PET\CT and CT respectively. Secondly, a Cross-Modal Feature Extractor “from frequency domain to spatial domain " (CMFE(F-> S)) is designed in the encoder. This module uses multi-frequency cross-dimensional attention to capture the correlation among multiple spectral maps in different dimensions and generates a refined frequency domain attention map. Through these refined frequency domain attention maps, the model can further optimize the performance of modal spatial features, enhance the interactive fusion of cross-modal features, and improve the complementarity of multi-modal information in segmentation tasks. Finally, a cross-modal feature coupler CMFC(S-> F) from spatial domain to frequency domain is constructed in the bottleneck layer. The core goal of this module is to supplement the fine-grained frequency domain features of PET and CT through the spatial features of PET/CT in the deep layer of the network, so as to obtain more powerful hybrid domain features. These hybrid domain features are able to contain more comprehensive lesion and tissue information, helping the model to identify tumors and other lesion regions more accurately. In general, by introducing the bidirectional feature interaction mechanism between frequency domain and spatial domain, and the deep fusion of cross-

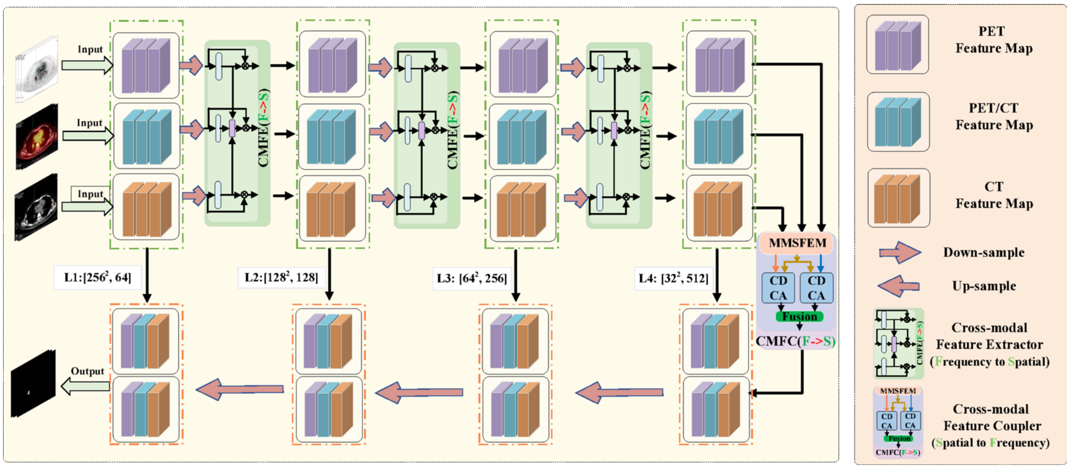

Fig. 2. Cross-modal dual-domain bi-direction feature interaction semantic segmentation network DBW-Net.

modal features, DBW-Net can not only enhance the information expression ability in segmentation tasks, but also improve the accuracy and robustness of the model in processing multi-modal medical images.

Cross-modal feature extractor from frequency to spatial (CMFE(F-> S))

Motivation: Since the boundary, texture and background of multi-modal medical images are preserved in different frequency components, the multi-frequency information is crucial for building an effective multi-modal medical image segmentation model. Recent studies on neural network demonstrate that the frequency domain attention (FCANet)[31] are very important for improving model performance. In this method, multiple frequency components are introduced to calculate the attention among channels to reduce the information loss caused by the average pooling of traditional channel attention (SE). By calculating the weight information generated by the correlation of multiple frequency components among channels, the recalibration of the input feature maps is completed, and the compensation is realized from "frequency domain information" to "spatial domain feature map". However, it only calculates the frequency domain attention in the channel direction, ignoring the correlation information of other dimensions. This inevitably leads to the information loss. At the same time, for multi-modal medical images, how to effectively use the dependence of multi-modal spatial domain features on frequency domain information, that is very important in clinic.

Therefore, based on the above reasons, this paper proposes a cross-modal feature extractor CMFE(F-> S) "from frequency domain to spatial domain", its structure is shown in Fig. 3A, which aims to make full use of the global information of each dimension in cross-modal multiple spectral maps. The method models the interdependencies among the channels, heights and widths of the input feature maps for each modality, and calibrates the feature maps to improve the representation ability. This module consists of two main steps: In the first step, the refined frequency domain attention map is captured by the multi-frequency cross-dimensional attention module. The second step is the interactive fusion stage. For the auxiliary branches of PET and CT, the generated refined frequency domain attention map is used to recalibrate the input feature map. For the main branch of PET/CT, the refined frequency domain attention map of PET, CT and PET/CT is fused, the cross-modal refined frequency-domain attention maps of each modality are fully interacted, and the input feature map of the PET/CT branch is recalibrated.

Multi-frequency cross-dimensional attention module (MFCDA)

The proposed multi-frequency cross-dimensional attention module is shown in Fig. 3B. For a given input feature map $X_i^S \in R^{C_S \times H_S \times W_S}$(Here i represents the output of the ith layer encoder, i = 1,2,3; S denotes which modal branch, S = PET, PET/CT, CT.),the spatial domain feature map is converted into k frequency domain feature maps by 2D DCT(Here, the number of frequency components k value is selected as 4, and the k value is selected as described in the ablation experiment.). The formula for performing the 2D DCT transform is stated as follows:

$$D_{h,w}^{u_k,v_k} = \cos\left(\frac{\pi h}{H}\left(u_k + \frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(v_k + \frac{1}{2}\right)\right) \tag{1}$$

$$X_i^{S,k} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_i^S \times D_{h,w}^{u_k,v_k} \tag{2}$$

2D DCT basic image is defined as $D_{h,w}^{u_k,v_k}$, with top-k selection strategy[31]. $X_i^{S,k}$ represents the frequency domain feature map with frequency component (u_k, v_k) selected by 2D DCT transformation. In particular,

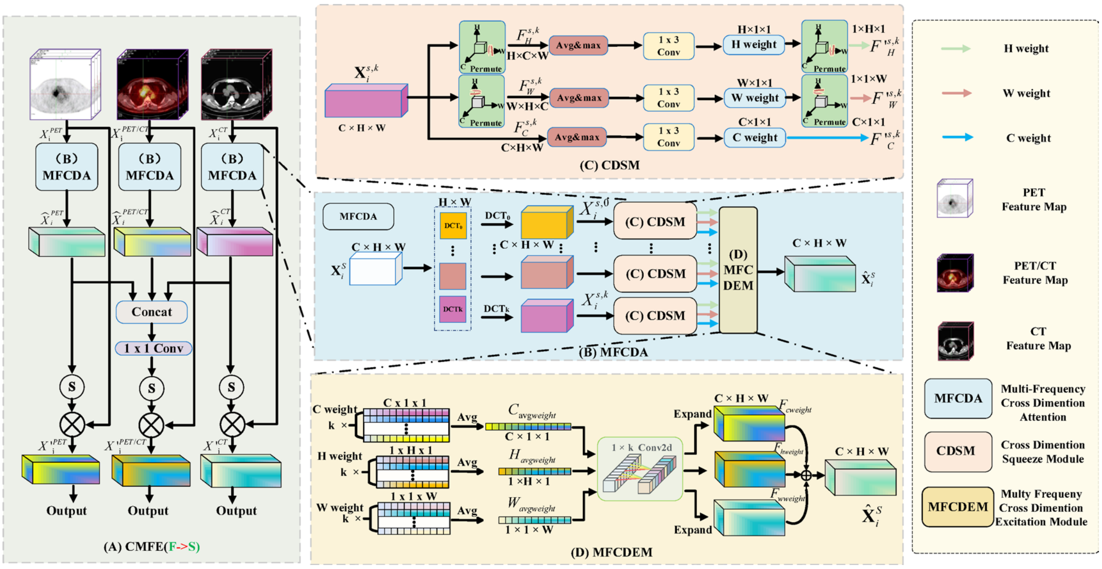

Fig. 3. Cross-modal feature extractor from frequency to spatial (CMFE(F-> S)).

when the frequency component is selected as (0,0), the feature map $X_i^{S,0}$ is a special frequency domain feature map, and its essence is a spatial domain feature map. The formula $X_i^{S,0}$ is stated as follows:

$$X_i^{S,0} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_i^S \times D_{h,w}^{u_k,v_k} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_i^S \times \cos\left(\tfrac{\pi h}{H}\left(0+\tfrac{1}{2}\right)\right)\cos\left(\tfrac{\pi w}{W}\left(0+\tfrac{1}{2}\right)\right) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_i^S \quad (3)$$

After that, each frequency domain feature map $X_i^{S,k}$ is processed by the Cross-dimensional squeeze module (CDSM) to generate k groups of cross-dimensional attention weight vectors (Each group contains three dimensions C, H and W). Then, the above k groups of attention weight vectors are processed by the Multi-Frequency Cross-dimensional Excitation Module (MFCDEM) to generate the selective attention map $\widehat{X}_i^S$ with refined frequency domain features. Finally, these attention maps can be applied complementary to input feature maps for enhance the interested object representation. The specific process is as follows:

Cross-dimensional squeeze module (CDSM): The cross-dimensional squeeze module consists of three parallel branches as shown in Fig. 3C. The top two branches are responsible for capturing feature interdependencies in spatial dimensions H and W, respectively, while the last branch is used to capture inter-channel interactions. Specifically, let $X_i^{S,k} \in R^{C\times H\times W}$ be the input feature map of the CDSM module. In the top branch, $X_i^{S,k}$ is first rotated 90° counterclockwise along the W axis, and the resulting rotated feature map is denoted as $F_H^{S,k} \in R^{H\times C\times W}$. To model the long-range dependency between the channel dimension and the spatial dimension W, $F_H^{S,k}$ is average pooling along the H dimension as well as Max pooling, and the resulting aggregated feature vector is described as $\widehat{F}_H^{S,k}$. $\widehat{F}_H^{S,k}$ is then processed by 1 × 3 Conv to capture the interaction between features in the H dimension. The resulting H direction feature weights are denoted as $\widehat{F\prime}_H^{S,k} \in R^{H\times 1\times 1}$.

Finally, $\widehat{F\prime}_H^{S,k}$ is rotated 90° clockwise along the W axis, and the resulting aggregated feature vector $F\prime_H^{S,k} \in R^{1\times H\times 1}$ is the output of the top branch. The specific calculation process is as follows:

$$\hat{F}_H^{S,k} = Avg_H\left(PM_w^{-1}\left(X_i^{S,k}\right)\right) \oplus Max_H\left(PM_w^{-1}\left(X_i^{S,k}\right)\right) \quad (4)$$

$$F'^{\,s,k}_H = PM_w\left(Conv_{1\times 3}\left(\widehat{F}_H^{S,k}\right)\right) \quad (5)$$

Among them, PM_w operation means 90° clockwise Permute rotation along W axis, and PM_w^{-1} operation means 90° anticlockwise Permute rotation along W axis. Avg_H represents the average pooling along the H dimension direction, and Max_H represents the maximum pooling along the H dimension direction.

Similarly, the aggregated feature vector $F\prime_W^{S,k}$ for the middle branch and the aggregated vector $F\prime_C^{S,k}$ for the bottom branch can be obtained. This process can be summarized as the following equation:

$$\hat{F}_W^{S,k} = Avg_W\left(PM_H^{-1}\left(X_i^{S,k}\right)\right) \oplus Max_W\left(PM_H^{-1}\left(X_i^{S,k}\right)\right) \quad (6)$$

$$F'^{\,s,k}_W = PM_H\left(Conv_{1\times 3}\left(\widehat{F}_W^{S,k}\right)\right) \quad (7)$$

$$F'^{\,s,k}_C = Conv_{1\times 3}\left(\widehat{F}_C^{S,k}\right) \quad (8)$$

Multi-frequency Cross-dimensional excitation module (MFCDEM): As shown in Fig. 3D, the Multi-frequency cross-dimensional Excitation Module (MFCDEM) aims to fully interact the cross-dimensional effective information among multiple frequency feature maps. Firstly, the module calculates the mean value of multiple groups of frequency domain cross-dimensional weight vectors according to the dimension category, and obtains the multi-frequency cross-weight vectors $C_{avgweight}, H_{avgweight}, W_{avgweight}$.Then, the above multi-frequency cross-dimensional weight vector is processed by 1×k convolution, and the dimensional information among multi-frequency feature maps is fully interacted to obtain the cross-frequency cross-dimensional weight vector $C\prime_{avgweight}, H\prime_{avgweight}, W\prime_{avgweight}$. After that, the cross-frequency and cross-dimension weight vectors are expanded to obtain the attention weight feature maps $F_{c\_weight}, F_{h\_weight}, F_{w\_weight}$. Finally, the above three attention weight feature maps are added to obtain the refined frequency domain attention map $\widehat{X}_i^S$. The above process can be summarized as the following equation.

$$\begin{cases} C_{avgweight} = \frac{1}{k}\sum{}_{i=1}^{k} F'^{\,s,i}_C \\ H_{avgweight} = \frac{1}{k}\sum{}_{i=1}^{k} F'^{\,s,i}_C \\ W_{avgweight} = \frac{1}{k}\sum{}_{i=1}^{k} F'^{\,s,i}_W \end{cases} \quad (9)$$

$$\begin{aligned}\hat{X}_i^S = {} & Expand_{C\times H\times W}\left(Conv_{1\times 3}\left(C_{avgweight}\right)\right) \oplus Expand_{C\times H\times W}\left(Conv_{1\times 3}\left(C_{avgweight}\right)\right) \\ & \oplus Expand_{C\times H\times W}\left(Conv_{1\times 3}\left(C_{avgweight}\right)\right)\end{aligned} \quad (10)$$

Interactive fusion stage

The output of the three modal feature maps processed by MFCDA module is denoted as $\widehat{X}_i^{PET}, \widehat{X}_i^{PET/CT}, \widehat{X}_i^{CT}$ respectively. For the main branch of PET/CT, the weight feature map F_{Concat} obtained by Concat aggregation of refined frequency domain attention map $\widehat{X}_i^{PET}$, $\widehat{X}_i^{PET/CT}$, $\widehat{X}_i^{CT}$. Then F_{Concat} interactively calculates the weight information of each modality through 1×1Conv processing, and then activates it by the Sigmoid function. Finally, the activated refined frequency domain weight feature map is multiplied with the input feature map $\widehat{X}_i^{PET}$ of the PET/CT branch to obtain the recalibration of the feature map $\widehat{X}_i^{PETCT}$. For the PET and CT auxiliary branches, the refined frequency domain attention maps $\widehat{X}_i^{PET}$ and $\widehat{X}_i^{CT}$ are activated by Sigmoid, and then multiplied with the input feature maps X_i^{PET} and X_i^{CT} of the PET and CT branches, respectively, to obtain the recalibration of the PET and CT branch feature maps. Through the above methods, not only the feature extraction ability of PET and CT modal is enhanced, but also the information of the other two modal is aggregated for the main branch of PET/CT, and the dependence of multi-modal spatial domain features on frequency domain information is effectively used to improve the extraction ability of the network.

Cross-modal feature coupler from spatial to frequency CMFC (S-> F)

Considering that the low-frequency information of PET images shows the area with slow change of gray value, which corresponds to the macroscopic structural characteristics of the lesion area with abnormal vigorous glucose metabolism and the surrounding tissue background. The high frequency information of CT image is manifested as the area with rapid change of gray value, which reflects the detail characteristics of the boundaries and textures of tissues, organs and lesions. PET/CT images combine the advantages of PET and CT, and directly extracting features in the spatial domain can retain its comprehensiveness and global complementarity. Therefore, in order to make full use of the multi-modal cross-domain high-level semantic information, this paper designs a cross-modal feature coupler "from spatial domain to frequency domain" CMFC(S-> F) in the bottleneck layer. This module aims to supplement the information of the fine-grained frequency domain features of PET and CT through the spatial features of PET/CT, to obtain more powerful hybrid domain features. As shown in Fig. 4A, the multi-modal features extracted from the last layer of the encoder are mapped to the spatial and frequency domains using the Multi-modal Space-frequency Feature Extraction module (MMSFEM). Then, in order to fit the semantic gap between frequency domain and spatial domain features and promote the combination of features from different representation domains, this paper designed cross-domain Coupled Attention (CDFA). CDFA is used to align, select and combine fine-grained high-frequency features, spatial features and fine-grained low-frequency features to obtain hybrid domain features. Finally, the deep cross-domain semantic information of multi-modal features is fully fused. By combining the cross-domain information of the three modalities in the above way, the model can establish a more effective connection between functional and anatomical information, which can improve the detection and segmentation accuracy of tumor regions.

Multi-modal space-frequency feature extraction module (MMSFEM)

The module consists of three parallel branches, which are used to extract low-frequency information of PET image features, spatial information of PET/CT image features and high-frequency of CT image respectively. The multi-modal features extracted by the last encoder are mapped into spatial and frequency domains, as shown in Fig. 4B. Specific operations are as follows: For the high-level semantic feature Y^{PET} of the PET branch, the Haar wavelet transform is used for down-sampling to convert Y^{PET} into four frequency domain components, one of which is the low frequency component Y_A^{PET}, and the other three are the horizontal, vertical and diagonal high frequency components $Y_H^{PET}, Y_V^{PET}, Y_D^{PET}$. Then, the feature representation of the low-frequency information is learned by 1×1Conv, and the fine-grained low-frequency feature $\widehat{Y}_L^{PET}$ is obtained.

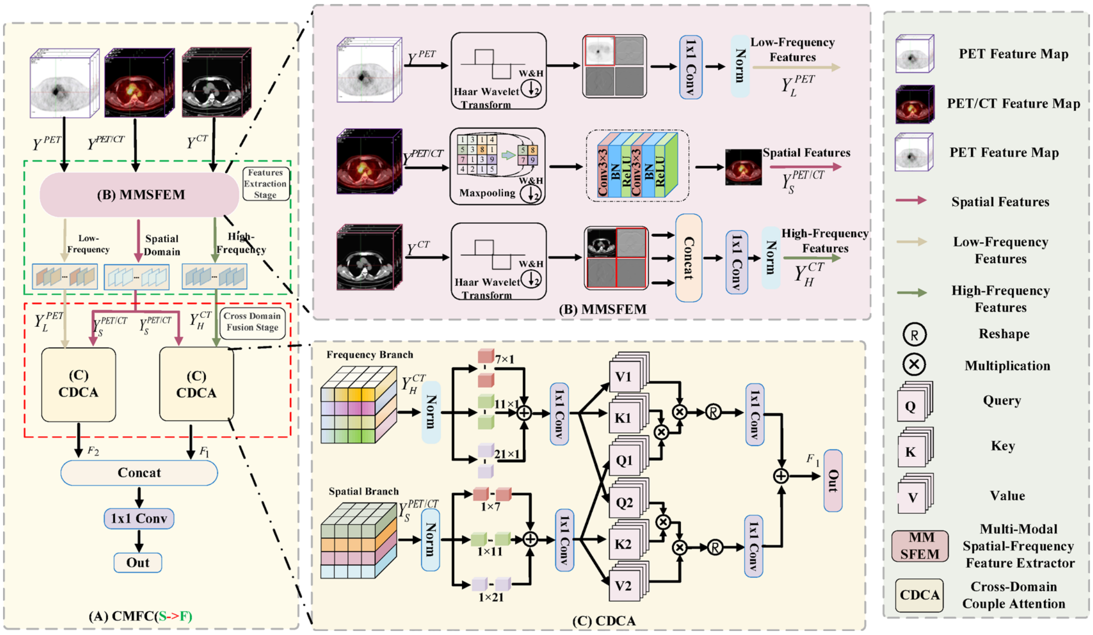

Fig. 4. Cross-modal feature coupler from spatial to frequency (CMFC(S-> F)).

For the high-level semantic feature $Y^{PET/CT}$ of PET/CT branch, Max pooling is used for downsampling, and then the spatial feature $\widehat{Y}_S^{PET/CT}$ is obtained by twice 3×3 Conv, BN and Relu. For the high-level semantic feature Y^{CT} of CT branch, the Haar wavelet transform is also used for downsampling to convert Y^{CT} into 4 frequency domain components, and the low frequency component Y_A^{CT} and 3 different directions of high frequency components Y_H^{CT}, Y_V^{CT} , Y_D^{CT} are obtained. After that, the high frequency components of the 3 directions are aggregated by Concat. Finally, the 1 × 1 Conv is used to perform feature representation learning of high-frequency information, and the fine-grained high-frequency feature $\widehat{Y}_H^{CT}$ is obtained.

Taking the input of CT branch $Y^{CT} \in R^{C\times H\times W}$ as an example, the Haar wavelet transform is performed to transform each channel feature $Y_C^{CT} \in R^{H\times W}$ into a low frequency component and 3 high frequency components in different directions are shown in Formula (11) and (12):

$$\left\{\begin{array}{c} A_c\left(i,j\right)=\frac{Y_C^{CT}(i,2j-1)+Y_C^{CT}(i,2j)}{2} \\ D_c\left(i,j\right)=Y_C^{CT}\left(2i-1,j\right)-Y_C^{CT}\left(2i,j\right) \end{array}\right. \tag{11}$$

$$\left\{\begin{array}{c} A_c=AA_c\left(i,j\right)=\frac{A_c(i,2j-1)+A_c(i,2j)}{2} \\ H_c=AD_c\left(i,j\right)=\frac{D_c(i,2j-1)+D_c(i,2j)}{2} \\ V_c=DA_c\left(i,j\right)=A_c\left(i,2j-1\right)-A_c\left(i,2j-1\right) \\ D_c=DD_c\left(i,j\right)=D_c\left(i,2j-1\right)-D_c\left(i,2j\right) \end{array}\right. \tag{12}$$

Where $A_c\left(i,j\right)$ represents the low-frequency approximation coefficient of channel C, $D_c\left(i,j\right)$ represents the high-frequency detail coefficient of channel C. The range of *i* is $i \in \ (1,H)$, and the range of j is $j \in \ (1,W/2)$. Then, the Haar wavelet transform is applied to the approximation coefficient and detail coefficient of each column, where A_c is the single-channel approximation coefficient, H_c is the single-channel horizontal detail coefficient, V_c is the single-channel vertical detail coefficient, D_c is the single-channel diagonal detail coefficient. The range of i is $i \in \ (1,H/2)$, and the range of j is $j \in \ (1,W/2)$. Finally, the 3 high-frequency components are concatenated and then point convolution is performed for low-dimensional mapping to obtain the final fine-grained high-frequency features. The specific formula is shown in Formula (13):

$$\hat{Y}_H^{CT}=f_{WH}\left(Y^{CT}\right)=\left(\rho\left(\delta_{1\times1}\left(Concat\left(Y_H^{CT},Y_V^{CT},Y_D^{CT}\right)\right)\right)\right) \tag{13}$$

Similarly, for the PET branch, the mathematical description of the fine-grained low-frequency characteristics is shown in Formula (14):

$$\hat{Y}_L^{PET}=f_{WL}\left(Y^{PET}\right)=\left(\rho\left(\delta_{1\times1}\left(Y_A^{PET}\right)\right)\right) \tag{14}$$

Where $\widehat{Y}_H^{CT} \in R^{C\times H\times W}$ represents fine-grained high-frequency features, $\widehat{Y}_L^{PET} \in R^{C\times H\times W}$ represents fine-grained low-frequency features, ρ (●)represents batch normalization calculation, and $\delta_{k\times k}$ represents the size of convolution kernel k×k.

Cross-domain coupled attention (CDCA)

Frequency domain features and spatial domain features reflect different aspects and properties of images, but due to their semantic differences, semantic alignment is required to ensure consistency and complementarity. To achieve semantic alignment, feature selection of frequency and spatial domain features, Cross-Domain Coupled Attention (CDCA) is designed, as shown in Fig. 4C. The module consists of two parallel branches in frequency domain and spatial domain. Each branch consists of 3 one-dimensional convolution with different kernel sizes. The frequency domain branch encodes multi-scale context information in vertical direction, and then aggregates cross-domain attention features in horizontal direction. The spatial branch encodes multi-scale context information along the horizontal direction, and then fuses features along the vertical cross-domain attention. In the following, the fine-grained high-frequency feature Y_H^{CT} and the spatial feature $Y_S^{PET/CT}$ are taken as examples to explain the working mechanism behind CDCA.

Firstly, the frequency domain features and spatial domain features are mapped to a unified scale using scale mapping, that is, vertical and horizontal bar convolution at different scales are used to process each feature, secondly, they are connected and mapped to a unified scale matrix Q, K, and V using 1 × 1 convolution as input for the next stage. Then two matrices (Q1, K1, V1) and (Q2, K2, V2) are obtained respectively for spatial domain features and fine-grained frequency domain features. Finally, Cross-Domain Coupled Attention is used for feature fusion. Specifically, for the frequency domain branch, the frequency domain features after unified scale are regarded as the key value matrix K1, V1, and the spatial features are regarded as the query matrix Q2. Similarly, for the spatial branch, the unified scale spatial feature is regarded as the key-value matrix K2, V2, and the fine-grained frequency domain feature is regarded as the query matrix Q1, and then the cross-attention calculation is performed. The calculation process is shown in Formula (15), (16), (17) and (18):

$$Q1,K1,V1=f_{ms}\left(Y_H^{CT}\right)=\delta_{1\times1}\left(\sum_{i=0}^{2}Conv1D_i^y\left(Norm\left(Y_H^{CT}\right)\right)\right) \tag{15}$$

$$Q2,K2,V2=f_{ms}\left(Y_S^{PETCT}\right)=\delta_{1\times1}\left(\sum_{i=0}^{2}Conv1D_i^x\left(Norm\left(Y_S^{PETCT}\right)\right)\right) \tag{16}$$

$$Attn\left(Q,K,V\right)=Softmax\left(\frac{QK^{T}}{\sqrt{d}}\right)V \tag{17}$$

$$F_{1}=\delta_{1\times 1}\left(\mathrm{Attn}\left(Q2,K1,V1\right)\right)\oplus\delta_{1\times 1}\left(\mathrm{Attn}\left(Q1,K2,V2\right)\right) \tag{18}$$

Here, $Conv1D_{i}^{y}, Conv1D_{i}^{x}$ represent one-dimensional convolutions along vertical and horizontal directions, $\delta_{\ k\times\ k}$ represent convolutions with kernel size k×k. For the kernel sizes of 1D convolutions, they are set to 1×7, 1×11 and 1×21, respectively. $Norm\left(\bullet\ \right)$ is layer normalization, and $f_{ms}\left(\bullet\ \right)$ represents the multiscale mapping operation. Q, K, V are the matrices after using multi-scale orthogonal convolution operation, F_{1} is the output of CDCA.

Experiments

Dataset and implementation details

To evaluate the method proposed in this paper, the effectiveness of DBW-Net is verified using the clinical multi-modal lung tumor medical image dataset and the Brats2019 public brain tumor dataset[32]. The specific situation is as follows:

Ⅰ: This study collected 90 clinical patients with lung tumors who underwent PET/CT whole-body examinations using private datasets in the nuclear medicine department of a tertiary hospital in Ningxia. Each patient had 500 PET/CT, PET, and CT images. The labels of these datasets are manually defined by clinicians after referring to multi-modal medical images of the lungs. The dimensions of the obtained doctor-labeled CT, PET and PET/CT two-dimensional sections are 356 pixels ×356 pixels. After data augmentation processing such as rotation and mirroring, the image size is adjusted to 256 pixels ×256 pixels through Resize processing. Finally, the sample sizes of the three modal image datasets are 1,770 respectively, among which 1,416 are divided into the training set and 354 into the test set.

Ⅱ: Due to the particularity of the multi-modal lung tumor dataset used in this chapter, there are currently no publicly available datasets on multi-modal lung tumors. To ensure the consistency of the experiment, our method is evaluated on the open dataset of the Multi-modal Brain Tumor Segmentation Challenge (BraTS) in 2019. We processed the BraTS (2019) dataset as follows: In our experiment, data from FLAIR, T1 and T2 modes are selected, with 259 3D image data selected for each mode. For each three-dimensional image, the longitudinal slicing method is adopted, and the slices are made into two-dimensional images and saved in PNG format. Then, through resizing processing, the image size is adjusted to 256 pixels ×256 pixels. Finally, the sample size of the three modal image datasets is 1,745, of which 1,396 are divided into the training set and 349 into the test set.

Experimental environment and parameter setting

Experimental environment: The model in this chapter is implemented based on Pytorch, laboratory hardware environment server Intel(R) Xeon(R) Gold 6154 CPU, memory 256GB, graphics card NVIDIA TITAN V, python3.7, PyTorch1.7.0, CUDA version is 11.1.106. In this chapter, we used the Adam optimizer.

Experimental parameters: In the parameter setting, the training period is 100, the learning rate is 0.0001, and the batch size is 4 to ensure that the model parameters are updated faster and converge at the global optimal point. The binary cross-entropy loss function is used to reflect the gap between the predicted value and the actual value of the model, and Miou is used as a reference to update the weight during the training process.

Evaluation metrics

As shown in Table 1, in order to objectively and comprehensively evaluate the segmentation effect of this network and compare it with other methods, this chapter based on the prediction results of the model, The Mean Intersection over Union (Miou), Dice Similariy Coefficient (Dice), Recall, Volumetric Overlap Error (Voe), Relative Volume Difference (Rvd), Haudorff distance (HD) to evaluate the performance of DBW-Net. To avoid ambiguity, we further standardize the parameter definition: TP (True Positive) : the region correctly segmented as a lesion; TN (True Negative): Region correctly segmented as normal tissue; FP (False Positive) : an area that is actually normal tissue but is incorrectly segmented as a lesion; FN (False Negative) : an area that is actually diseased but incorrectly segmented as normal tissue; P: the target pixel region predicted by the model (prediction); G: the target pixel region in the label value (ground truth); abs means taking the absolute value of the result.

Experimental design and analysis

To evaluate the segmentation model in this paper, a large number of experiments are conducted in this chapter. The experiment is conducted from the following aspects:

Evaluationin index	Definition	Evaluation index	Definition
MIoU	$\mathrm{Miou}=\frac{1}{k+1}\sum\ _{i=0}^{k}\frac{TP}{TP+FN+FP}$	Voe	$\mathrm{Voe}=\mathrm{abs}\left(1-\left\lvert\frac{P\cap\ G}{P\cup\ G}\right\rvert\right)$
Dice	$\mathrm{Dice}=\ \frac{2\times\ TP}{FN+TP+FP+FP}$	Rvd	$\mathrm{Rvd}=\mathrm{abs}\left(\frac{P-G}{G}\right)$
Recall	$\mathrm{Recall}=\ \frac{TP}{TP+FN}$	HD	$\mathrm{HD}\left(\mathrm{G},\mathrm{P}\right)=\max\left\{\max_{g\in\ G}\min_{p\in\ P}?g-p?,\max_{p\in\ P}\min_{g\in\ G}?p-g?\right\}$

Table 1. Evaluation indicators.

(1) Comparative experiments: This group of experiments conducted comparative experiments on multi-modal lung tumors and multi-modal brain tumors respectively. By comparing with single-modal and multi-modal segmentation networks, the effectiveness and advancement of the methods proposed in this chapter are verified.

(2) Ablation Experiment: This group of experiments is divided into two parts. The first part conducts ablation experiments on the method proposed in this paper, verifying the effectiveness and necessity of introducing the proposed method into the medical image segmentation task on two datasets. The second part conducted ablation experiments within the CMFE (F-> S) module and the CMFC(S-> F) module on the multi-modal lung tumor dataset. The selection of the number k of frequency components in the CMFE (F-> S) module of the cross-modal feature extractor is verified, as well as the rationality of using low-frequency information for PET and high-frequency information for CT in the modal feature mapping strategy of the CMFC (S-> F) module.

Comparison experiment

In the comparison experiment in this section, the experimental results as shown in Table 2, seven single-modal networks and four multi-modal networks are set up, and there are eleven groups of comparison experiments in total, among which U-Net, AttUnet, SeResUnet, UTNet, MsTGANet, TransUnet and DconnNet are all single-mode networks. Guide-Ynet, C3TMUNet, MEAUNet and MdCo-Unet are multi-modal segmentation networks that use three types of modal images as input.

For the multi-modal lung tumor dataset, PET/CT images rich in anatomical and metabolic information of the lesion are used as the training samples. The multi-modal network adopts three modal images, namely CT, PET and PET/CT, as the input of the multi-modal segmentation network. As shown in Fig. 5, for the single-modal network, Unet and TransUnet showed incomplete lesion segmentation in the boundary area (as shown in the fourth row, column 5 and column 10). AttUnet, SeResUnet, UTNet, MsTGANet, and DconnNet showed obvious over-segmentation in the lesion area (as shown in columns 6, 7, 8, 9, and 11 of the second row). These phenomena reflect that the traditional single-modal spatial feature extraction method still has obvious shortcomings in tumor boundaries, low-contrast regions and small lesions. In contrast, the multimodal network improves the overall segmentation effect, but Guide-Ynet and MEAUNet are not sensitive to small lesions (as shown in the fifth row, column 12 and column 14), C3TMUNet and MdCo-Unet appear to be not sensitive to the edge of complex lesions. The phenomenon that results in incomplete lesion segmentation (as shown in the third row, column 13, and the fourth row, column 15). As shown in Table 2, compared with the sub-optimal experimental results, the proposed method improves the Miou, Dice, VOE, RVD and Recall results by 0.72%, 0.47%, 1.16%, 0.39% and 0.69%, respectively.

To further verify the generalization ability of the method proposed in this paper on other multi-modal tumor datasets, this paper also conducted comparative experiments on multi-modal brain tumor datasets, in which the single-modal network took Flair images as the network input. The multi-modal network adopts Flair, T1 and T2 modal images as the input of the multi-modal segmentation network. As shown in Fig. 6, for the single-modal networks Unet, AttUnet, SeResUnet, UTNet, and MsTGANet, different degrees of lesion segmentation are incomplete (as shown in columns 5, 6, 7, 8, 9, and 10 of the fourth row). DconnNet suffers from over-segmentation (as shown in the fourth row, column 11). For multimodal networks, the segmentation effect is improved. However, Guide-Ynet and C3TMUNet are not sensitive to the edges of complex lesions, resulting

	Lung tumor DataSet						**Brats2019 DataSet**					
Method	**Miou↑(%)**	**Dice↑(%)**	**VOE↑(%)**	**RVD↑(%)**	**Recall↑(%)**	**HD↓**	**Miou↑(%)**	**Dice↑(%)**	**VOE↑(%)**	**RVD↑(%)**	**Recall↑(%)**	**HD↓**
Unet[9]	76.96	86.04	82.99	83.08	87.20	1.87	67.45	77.25	63.27	71.73	73.78	4.48
AttUnet[34]	78.22	87.09	84.29	84.28	87.59	1.85	**65.70**	**74.78**	**58.49**	**71.32**	**72.18**	**4.54**
SeResUnet[36]	77.06	85.93	81.65	**82.35**	86.04	1.86	66.73	77.01	65.65	73.26	74.72	4.53
UTNet[38]	**75.14**	**84.53**	**81.21**	82.61	**84.99**	**1.94**	72.84	82.83	77.92	75.86	84.09	4.33
MsTGANet[35]	77.98	86.78	83.48	83.79	86.92	1.86	77.55	86.25	82.05	84.34	84.35	3.94
TransUnet[37]	77.70	86.45	82.99	84.65	85.45	1.86	67.58	78.82	73.26	73.44	76.63	4.54
DconnNet[39]	77.23	86.39	83.48	82.33	88.02	1.87	78.23	86.97	85.99	85.83	88.34	4.02
Guide-Ynet[3]	79.72	87.98	84.60	84.84	87.63	1.78	78.23	87.19	86.55	86.88	87.09	4.04
C^3TMUNet[14]	*80.59*	*88.72*	*85.47*	*85.77*	*87.97*	1.80	*81.87*	*89.51*	*87.55*	*87.94*	*89.39*	*3.78*
MEAUNet[33]	80.09	88.43	85.27	86.19	86.86	*1.78*	79.48	87.61	83.01	85.38	85.20	3.89
MdCo-Unet[25]	79.81	87.95	84.11	84.65	87.96	1.81	79.12	87.50	84.42	85.18	87.47	3.94
Ours	81.31	89.19	86.63	86.58	88.66	1.78	82.79	90.03	88.53	88.92	90.32	3.74
$\text{Ratio}_{\text{highest}}$(↑)	0.72%	0.47%	1.16%	0.39%	0.69%	0.00	0.92%	0.52%	0.98%	0.98%	0.93%	0.04
$\text{Ratio}_{\text{lowest}}$(↑)	6.17%	4.66%	5.42%	4.23%	3.67%	0.16	17.09%	15.25%	30.04%	17.6%	18.14%	0.80

Table 2. The comparative experiment results (underline: best; italics: second best; bold: lowest). $\text{Ratio}_{\text{highest}}$ (↑)represents the ratio of DBW-Net experimental results to the results of the highest results among the comparison methods in Miou, Dice, VOE, RVD, and Recall. Ratiolowest (↑) represents the ratio of DBW-Net experimental results to the results of the lowest results among the comparison methods in Miou, Dice, VOE, RVD, and Recall.

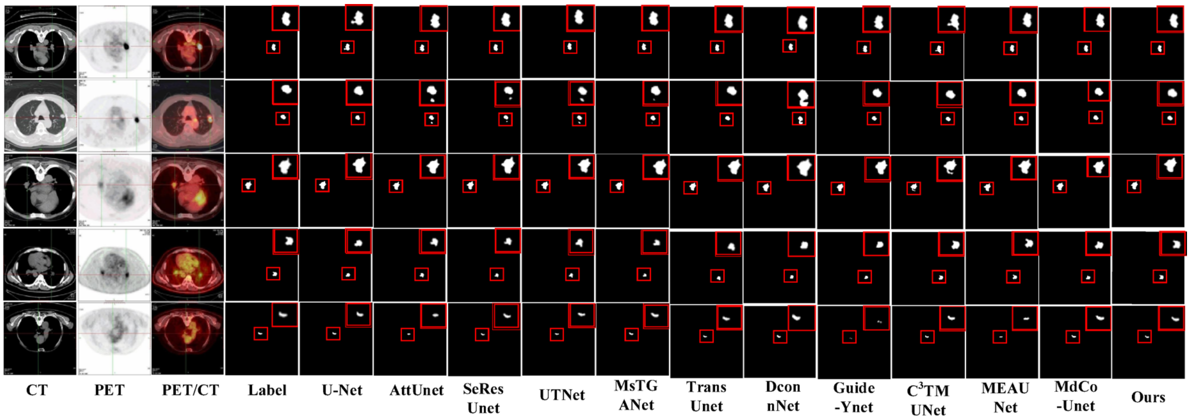

Fig. 5. Comparative experimental visual segmentation result map of lung tumor medical images.

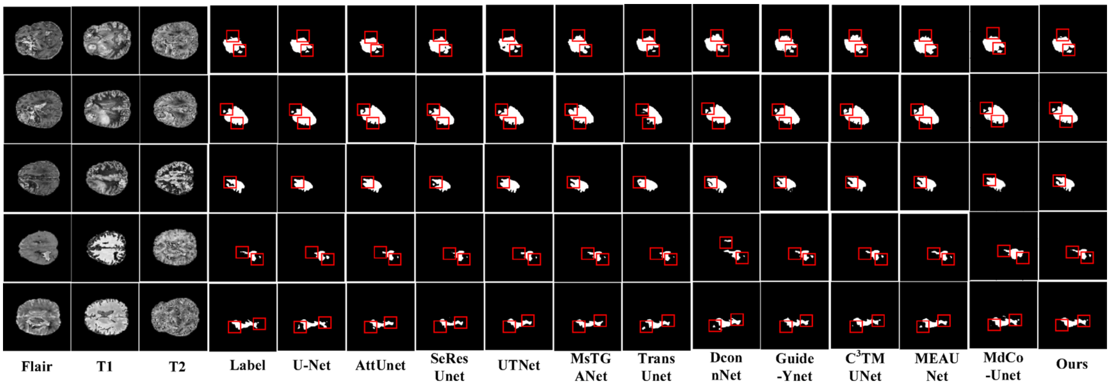

Fig. 6. Comparative experimental visual segmentation result map of BtraTS brain tumor.

in incomplete lesion segmentation (as shown in columns 12 and 13 in the fourth row).MEAUNet has an over-segmentation phenomenon (as shown in column 13 of row 5).There is an undersegmentation phenomenon in MdCo-Unet (as shown in the first row and column 15 of the fifth row). Table 2 shows the experimental results of the comparison between the proposed network and the other 11 networks. Compared with the suboptimal results of Miou, Dice, VOE, RVD and Recall, the proposed method improves the results by 0.92%, 0.52%, 0.98%, 0.98% and 0.93%, respectively.

The above results show that the existing methods are not sufficient for the extraction and fusion of cross-modal space-frequency features. However, the proposed method has the largest coverage of the coordinate axes of each evaluation index in radar Figs. 7 and 8, indicating that the model has obvious advantages in overall performance. This performance improvement is mainly due to the cross-modal two-domain bidirectional feature interaction mechanism proposed in this paper. This bidirectional interaction mechanism of "the compensation of frequency domain features to spatial features + the feedback of spatial features to frequency domain fusion" enables the model to have stronger feature expression power in areas with complex lesion structures and significant modal differences, thus achieving better performance in multimodal segmentation tasks.

Ablation experiment

The ablation experiments in this paper are divided into two parts. The first part conducts ablation experiments on the method proposed in this paper, verifying the effectiveness and necessity of introducing the proposed method into the medical image segmentation task on two datasets. The second part conducted ablation experiments within the CMFE (F-> S) module and the CMFC (S-> F) module on the multi-modal lung tumor dataset. The selection of the number k of frequency components in the CMFE (F-> S) module of the Cross-Modal Feature Extractor is verified, as well as the necessity of using low-frequency information for PET and high-frequency information for CT in the modal feature mapping strategy of the CMFC (S-> F) module.

Ⅰ: As shown in Table 3, this section will conduct ablation experiments on the method proposed in this paper on two datasets. A total of four sets of experiments have been designed to verify the necessity of adding the method proposed in this paper to the network. The first experiment is based on the U-Net structure of three encoders and one decoder. Experiment 2 is based on the encoder of Experiment 1 and adds the CMFE (F-> S) module. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall indices of Experiment 2 increased by 2.6%, 2.05%, 4.02%, 2.08% and 3.83% respectively, and the HD

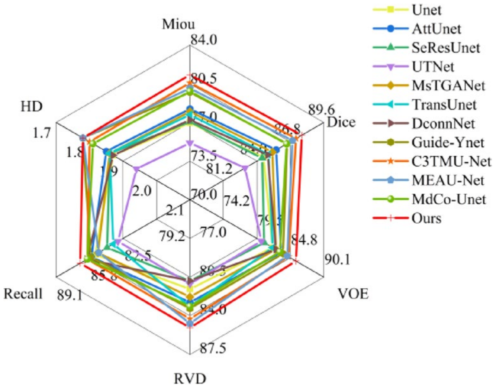

Fig. 7. Lung tumor comparison experimental radar map.

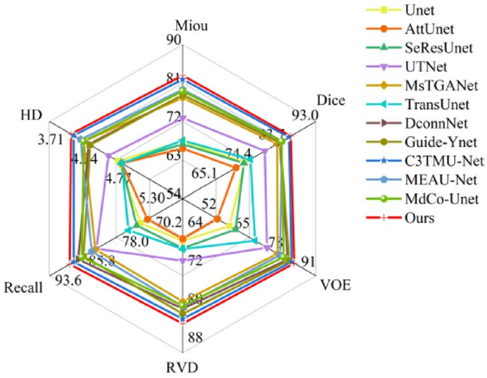

Fig. 8. BraTS brain tumor comparison.

Architecture	Lung tumor DataSet						Brats2019 DataSet					
	Miou↑	Dice↑	VOE↑	RVD↑	Recall↑	HD↓	Miou↑	Dice↑	VOE↑	RVD↑	Recall↑	HD↓
Baseline	78.29	86.87	81.97	83.95	84.50	1.88	79.73	87.72	83.85	86.28	84.56	3.81
+ CMFE(F-> S)	*80.89*	88.92	85.99	86.03	*88.33*	1.79	**81.14**	*89.09*	*87.48*	*88.19*	*87.56*	*3.81*
+ CMFC(S-> F)	80.66	88.76	*86.14*	*86.11*	88.31	1.77	80.69	88.44	84.97	86.44	87.41	3.82
Ours	81.31	89.19	86.63	86.58	88.66	1.78	82.79	90.03	88.53	88.92	90.32	3.74

Table 3. Results of the module ablation experiment.

index decreased by 0.09. For the multi-modal brain tumor dataset, the Miou, Dice, VOE, RVD and Recall indices of Trial 2 increased by 1.41%, 1.37%, 3.63%, 1.91% and 3% respectively. Experiment 3 is based on the encoder of Experiment 1 and adds the CMFC (S-> F) module. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall indices of Experiment 3 increased by 2.37%, 1.89%, 4.17%, 2.16% and 3.81% respectively. For the multi-modal brain tumor dataset, the Miou, Dice, VOE, RVD and Recall indices of Trial 3 increased by 0.96%, 0.72%, 1.12%, 0.16% and 2.85% respectively. Experiment 4 is the network proposed in this paper. For the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall index of Experiment 3 increased by 3.02%, 2.32%, 4.66%, 2.63% and 4.16% respectively. For the multi-modal brain tumor dataset, the Miou, Dice, VOE, RVD and Recall indices of Trial 3 increased by 3.06%, 2.31%, 4.68%, 2.64% and 5.76% respectively. Furthermore, in radar Figs. 9 and 10, the proposed model has the largest comprehensive coverage on the Miou, Dice, VOE, RVD, Recall, and HD evaluation index coordinate axes, which proves the effectiveness and necessity of adding the proposed method to the network.

II: As shown in Table 4, this subsection will conduct ablation experiments within the CMFE (F-> S) and CMFC (S-> F) modules proposed in this paper on the multi-modal lung tumor clinical dataset. A total of five sets of experiments have been designed for each module to verify the selection of the number k of frequency components in the CMFE (F-> S) module respectively. And the rationality of using low-frequency information for PET and high-frequency information for CT in the modal feature mapping strategy of the CMFC (S-> F) module.

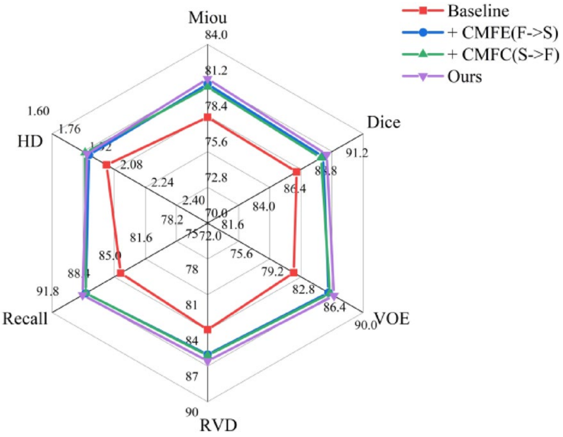

Fig. 9. Lung tumor ablation experimental radar map.

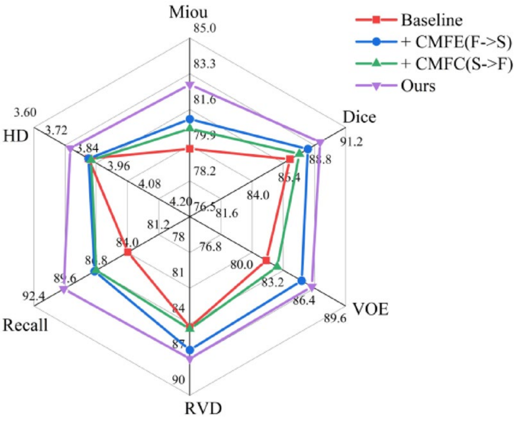

Fig. 10. BraTS brain tumor ablation experimental radar map.

	CMFE ablation							CMFC ablation					
Architecture	Miou	Dice	VOE	RVD	Recall	HD	Architecture	Miou	Dice	VOE	RVD	Recall	HD
Baseline	78.29	86.87	81.97	83.95	84.50	1.88	Baseline	78.29	86.87	81.97	83.95	84.50	1.88
+ CMFE(k = 1)	79.56	87.68	83.56	84.86	87.52	1.83	+ CMFC(PET_H,CT_H)	80.30	88.44	84.72	85.85	86.72	1.79
+ CMFE(k = 2)	80.22	88.39	85.09	85.46	87.41	1.82	+ CMFC(PET_L,CT_L)	*80.45*	*88.62*	*85.45*	*86.25*	*87.54*	*1.79*
+ CMFE(k = 8)	*80.43*	*88.66*	*85.69*	*85.84*	*88.13*	*1.80*	+ CMFC(PET_H,CT_L)	79.75	88.01	83.79	85.21	85.95	1.81
+ CMFE(k = 4)	80.89	88.92	85.99	86.03	88.33	1.79	+ CMFC(PET_L,CT_H)	80.66	88.76	86.14	86.11	88.31	1.77

Table 4. The internal ablation experimental results of the CMFE module and the CMFC module.

For the ablation experiment of CMFE (F-> S), Experiment 1 is based on the U-Net structure of three encoders and one decoder. Experiment 2 is based on Experiment 1 and adopted the CMFE module with the frequency component k = 1. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall index of Experiment 2 increased by 1.27%, 0.81%, 1.59%, 0.91% and 3.02% respectively. The HD index decreased by 0.05. Experiment 3 is based on Experiment 1 and adopted the CMFE module with the frequency component k = 2. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall index of Experiment 3 increased by 1.93%, 1.52%, 3.12%, 1.51% and 2.91% respectively. The HD index decreased by 0.06. Experiment 4 is based on Experiment 1 and adopted the CMFE module with frequency component k = 8. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall index of Experiment 4 increased by 2.14%, 1.79%, 3.72%, 1.89% and 3.63% respectively. The HD index decreased by 0.08. Experiment 5 is based on Experiment 1 and adopted the CMFE module with the frequency component k = 4. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall index of Experiment 5 increased by 2.6%, 2.05%, 4.02%, 2.08% and 3.83% respectively. The HD index decreased by 0.09. Furthermore, in radar Fig. 11, the CMFE

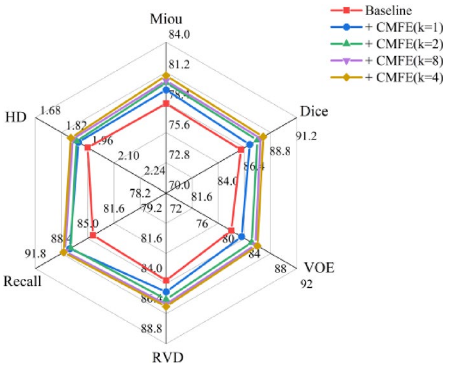

Fig. 11. CMFE ablation experimental radar map.

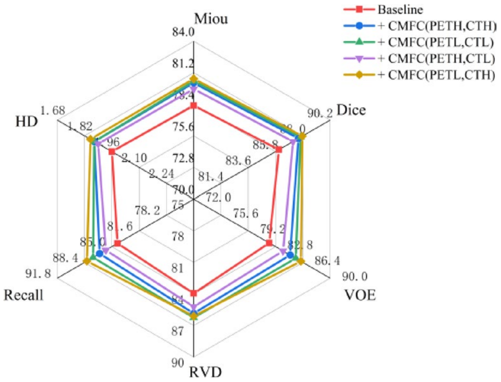

Fig. 12. CMFC ablation experimental radar map.

module with the frequency component k = 4 has the largest comprehensive coverage on the Miou, Dice, VOE, RVD, Recall, and HD evaluation index coordinate axes, which proves the effectiveness and necessity of the CMFE module when the frequency component k = 4 is adopted.

For the ablation experiment of CMFC (S-> F), Experiment 1 is based on the U-Net structure of three encoders and one decoder. Experiment 2 is based on Experiment 1 and adopted the CMFC module using high-frequency PET and high-frequency CT. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall index of Experiment 2 increased by 2.01%, 1.57%, 2.75%, 1.9% and 2.22% respectively. The HD index decreased by 0.09. Experiment 3 is based on Experiment 1 and adopted the CMFC module with low-frequency PET and low-frequency CT. Compared with Experiment 1, for the multi-modal lung tumor dataset, the Miou, Dice, VOE, RVD and Recall indices of Experiment 3 increased by 2.16%, 1.75%, 3.48%, 2.3% and 3.04% respectively. The HD index decreased by 0.09. Experiment 4 is based on Experiment 1 and adopted the CMFC module with high-frequency PET and low-frequency CT. Compared with Experiment 1, it is aimed at the multi-modal lung tumor dataset. The Miou, Dice, VOE, RVD and Recall. indices of Experiment 4 increased by 1.46%, 1.14%, 1.82%, 1.26% and 1.45% respectively, while the HD index decreased by 0.07. Experiment 5 is based on Experiment 1 and adopts the CMFC module that uses low-frequency PET and high-frequency CT. Compared with Experiment 1, it is aimed at the multi-modal lung tumor dataset. The Miou, Dice, VOE, RVD and Recall indices of Experiment 5 increased by 2.37%, 1.89%, 4.17%, 2.16% and 3.81% respectively, and the HD index decreased by 0.11. Furthermore, in radar Fig. 12, the CMFC module that uses PET with high frequency and CT with high frequency has the largest comprehensive coverage on the Miou, Dice, VOE, RVD, Recall, and HD evaluation index coordinate axes, which proves the effectiveness and necessity of the CMFC module when adopting this strategy.

Conclusion

From the perspective of frequency domain and spatial domain, this paper proposes a cross-modal dual-domain bi-directional feature interaction semantic segmentation network DBW-Net, which is "from frequency to spatial " and "from spatial to frequency". This paper effectively solves the problem that multi-modal segmentation methods usually focus on enhancing the extraction features ability from the spatial domain perspective, ignoring the dependence of multi-modal spatial domain features on frequency domain information and the effective fusion of fine-grained frequency domain features. Cross-Modal Feature Extractor "from frequency domain to spatial domain" (CMFE(F-> S)) and cross-modal feature coupler "from spatial to frequency" (CMFC(S-> F)) are introduced into the encoder and bottleneck layer. It improves the model's ability to extract, fuse and express fine-grained features in multi-modal images. About lung tumor segmentation, the experimental results show that the Miou, Dice, Voe, Rvd and Recall are increased by 3.02%, 2.32%, 4.66%, 2.63% and 4.16%, respectively. About brain tumor segmentation, the Miou, Dice, Voe, Rvd, Recall are increased by 3.06%, 2.31%, 4.68%, 2.64%, 5.76%,respectively. It shows that the model has high precision and relatively low redundancy about complex shape lesion segmentation, which significantly improves the segmentation accuracy and robustness of the lesion area, and provides technical support for accurate identification and early lesions diagnosis. In addition, this study not only proves the importance of frequency domain information in medical image segmentation tasks, but also provides a new idea for multi-modal image processing. By introducing 2D Discrete Cosine Transform (2D DCT) and 2D wavelet transform (2D DWT) to capture frequency domain features, the bi-directional interaction of cross-modal spatial-frequency features is realized, which effectively bridges the semantic differences of multi-modal images, and has good generalization ability and application potential.

Future work

This paper reveals the importance of frequency domain information in multi-modal medical image segmentation through the study of cross-modal dual-domain bidirectional feature interaction. However, for a wider range of application scenarios and practical requirements, future work can be carried out in depth from the following aspects: Firstly, this paper attempts to extend the space-frequency interaction method to other multi-modal medical image analysis tasks. For example, in cardiac image segmentation, the frequency domain can effectively extract the boundary information between heart tissue and surrounding structures, and in breast tumor segmentation, the high-frequency components of the frequency domain can significantly improve the accuracy of tumor boundaries. Medical image data in these fields has complex modal characteristics and diversified segmentation requirements. Secondly, incorporating more advanced learning mechanisms will further improve the model performance. For example, self-supervised learning uses unlabeled medical data for feature pre-training. By learning frequency domain features, the model can extract meaningful information from a wider range of data, thereby improving its performance on unlabeled data. In addition, cross-modal pre-training can help the model learn shared frequency domain and spatial features on large-scale cross-modal data, and effectively improve the generalization ability of the model in different tasks. Finally, the efficiency and lightweight design of the method are the key research directions in the future. In terms of frequency domain processing, it is possible to explore how to further reduce the computational and memory overhead by more efficient frequency domain conversion algorithms or by exploiting deep compression techniques such as quantization, pruning, etc. In conclusion, future research will continue to focus on the spatial-frequency feature interaction, and further promote the practical development of multi-modal medical image segmentation technology by expanding application fields, enhancing learning mechanisms and improving computational efficiency, and providing technical support for precision medicine.

Data availability

The Brats2019 public brain tumor dataset can be obtained by registering at the following link: https://www.med.upenn.edu/cbica/brats2019/data.html.

Received: 13 September 2025; Accepted: 3 November 2025
Published online: 04 December 2025

References

1. Nam, J. H., Syazwany, N. S., Kim, S. J. & Lee, S. C. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11480–11491. (2024).
2. Wang, X., Li, Z., Huang, Y. & Jiao, Y. Multimodal medical image segmentation using multi-scale context-aware network. *Neurocomputing* **486**, 135–146 (2022).
3. Zhou, T. et al. Guided-YNet: saliency feature-Guided interactive feature enhancement lung tumor segmentation network. *Computers Mater. Continua*. **80**, 4813–4832 (2024).
4. Fu, X., Bi, L., Kumar, A., Fulham, M. & Kim, J. Multimodal Spatial attention module for targeting multimodal PET-CT lung tumor segmentation. *IEEE J. Biomedical Health Inf.* **25**, 3507–3516 (2021).
5. Kumar, A., Fulham, M., Feng, D. & Kim, J. Co-learning feature fusion maps from PET-CT images of lung cancer. *IEEE Trans. Med. Imaging*. **39**, 204–217 (2019).
6. Fan, X., Liu, L. & Zhang, H. Multimodal information interaction for medical image segmentation. https://arXiv.org/abs/2404.16371. (2024).
7. Hou, Q., Peng, Y., Wang, Z., Wang, J. & Jiang, J. MFD-Net: Modality fusion diffractive network for segmentation of multimodal brain tumor image. *IEEE J. Biomed.Health Inf.* 5958–5969. (2023).
8. Gao, S., Yang, X., Jiang, L., Fu, Z. & Du, J. Global feature-based multimodal semantic segmentation. *Pattern Recogn.* **151**, 110340 (2024).
9. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference.* part III **18**, 234–241. (Springer, 2015).
10. Zhang, Z. & Wu, H. A novel deep learning model for medical image segmentation with convolutional neural network and transformer. *Interdisciplinary Sciences: Comput. Life Sci.* **15** (4), 663–677 (2023).
11. Yuan, Y. Medical image segmentation with UNet-based multi-scale context fusion. *Sci. Rep.* **14** (1), 15687 (2024).
12. Zhu, W. et al. P.Qiu,. Selfreg-unet: Self-regularized unet for medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 601–611. (Springer, 2024).
13. Zhang, Z. & Miao, Y. Deep learning and radiomics-based approach to meningioma grading: exploring the potential value of peritumoral edema regions[J]. *Phys. Med. Biol.* **69** (10), 105002 (2024).
14. Zhou, T. et al. Cross-modal cross-scale cross-dimension PET/CT image segmentation model based on transformer. *J. Electron. Inform. Technol.* **45**, 3529–3537 (2023).
15. Lin, J. et al. CKD-TransBTS: clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *IEEE Trans. Med. Imaging*. **42**, 2451–2461 (2023).

16. Xu, C., Li, Q., Jiang, X., Yu, D. & Zhou, Y. Dual-space graph-based interaction network for RGB-thermal semantic segmentation in electric power scene. *IEEE Trans. Circuits Syst. Video Technol.* **33**, 1577–1592 (2022).
17. Diao, Y., Fang, H., Yu, H., Li, F. & Xu, Y. Multimodal invariant feature prompt network for brain tumor segmentation with missing modalities. *Neurocomputing* **616**, 128847 (2025).
18. Wang, Z. et al. Multimodal parallel attention network for medical image segmentation. *Image Vis. Comput.* **147**, 105069 (2024).
19. Wang, Q. et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 11534–11542. (2020).
20. Misra, D., Nalamada, T., Arasanipalai, A. U. & Hou, Q. Rotate to attend: Convolutional triplet attention module. *Proceedings of the IEEE/CVF winter conference on applications of computer vision* 3139–3148. (2021).
21. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13713–13722 (2021).
22. Zhang, H. et al. Resnest: Split-attention networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2736–2746. (2022).
23. Ouyang, D. et al. Efficient multi-scale attention module with cross-spatial learning. *ICASSP 2023–2023 IEEE Int. Conf. Acoust. Speech Signal. Process. (ICASSP)*. **IEEE**, 1–5 (2023).
24. Yu, Y., Zhang, Y., Cheng, Z., Song, Z. & Tang, C. Multidimensional collaborative attention in deep convolutional neural networks for image recognition. *Eng. Appl. Artif. Intell.* **126**, 107079 (2023).
25. Zhou, T. et al. Model-data co-driven U-net segmentation network for multimodal lung tumor images. *Appl. Soft Comput.* **180** https://doi.org/10.1016/j.asoc.2025.113410 (2025).
26. Huang, Y. et al. Medical frequency domain learning: Consider inter-class and intra-class frequency for medical image segmentation and classification. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 897–904. (IEEE, 2021).
27. Azad, R., Bozorgpour, A., Asadi-Aghbolaghi, M., Merhof, D. & Escalera, S. Deep frequency re-calibration u-net for medical image segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision* 3274–3283. (2021).
28. Tang, X., Peng, J., Zhong, B., Li, J. & Yan, Z. Introducing frequency representation into convolution neural networks for medical image segmentation via twin-Kernel fourier convolution, computer methods and programs in biomedicine. **205** 106110. (2021).
29. Mu, N. et al. Exploring a frequency-domain attention-guided cascade U-Net: towards spatially tunable segmentation of vasculature. *Comput. Biol. Med.* **167**, 107648 (2023).
30. Chen, J., Yang, G., Liu, A., Chen, X. & Liu, J. Spatial-Frequency enhancement network for robust nuclei segmentation in histopathology images. *Comput. Biol. Med.* **171**, 108131 (2024).
31. Qin, Z., Zhang, P., Wu, F. & Li, X. Fcanet: Frequency channel attention networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 783–792. (2021).
32. Farnoosh, R. & Aghagoli, F. Pseudo-deep unsupervised model-based clustering for brain tumor detection in magnetic resonance images. *Appl. Soft Comput.* **174**, 1568–4946. https://doi.org/10.1016/j.asoc.2025.112940 (2025).
33. Zhou, T. et al. Cross-modal multi-encoding hybrid attention U-Net for lung tumor image segmentation. *Acta Photonica Sinica*. **51**, 0410006 (2022).
34. Oktay, O. et al. Attention u-net: learning where to look for the pancreas. https://arXiv.org/abs/1804.03999. (2018).
35. Wang, M. et al. MsTGANet: automatic Drusen segmentation from retinal OCT images. *IEEE Trans. Med. Imaging*. **41** (2), 394–406 (2021).
36. Cao, Z. et al. Cascaded SE-ResUnet for segmentation of thoracic organs at risk. *Neurocomputing* **453**, 357–368 (2021).
37. Chen, J. & Lu, Y. Q.Yu, et al. Transunet: Transformers make strong encoders for medical image segmentation. https://arXiv.org/abs/2102.04306. (2021).
38. Y.Gao, M. & Zhou, D. M. UTNet: a hybrid transformer architecture for medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI. : 24th International Conference*. Part III **24**. 61–71. (Springer International Publishing, 2021).
39. Yang, Z. S.Farsiu. Directional connectivity-based segmentation of medical images[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11525–11535. (2023).

Author contributions

T.Zhou: Funding acquisition, Resources, Supervision, Writing –review & editing. Q. Liu: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. K.Song: Data collection, Writing – review & editing. W.Chai: Writing – review & editing. K.Chen: Organize references, Validation. H.Lu: Validation.

Funding

This work is supported by the National Natural Science Foundation of China (No.62561002, 62576009), Natural Science Foundation of Ningxia Province(No.2025AAC020006).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.