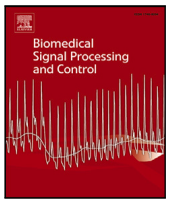
Biomedical Signal Processing and Control 109 (2025) 107945

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



Multi-domain fusion network: A novel approach to OCTA image segmentation in diabetic retinopathy

Guangmei Jia [a], Fei Ma [a,b,*], Sien Li [a], Zhaohui Zhang [a], Hongjuan Liu [a], Yanfei Guo [a], Jing Meng [a,b]

[a] School of Computer Science, Qufu Normal University, Rizhao, 276826, China
[b] Joint Technology Transfer Center of Rizhao and Qufu Normal University, Rizhao, 276826, China

ARTICLE INFO

Keywords:
Deep learning
Frequency domain
OCTA
Segmentation
Retinal lesion

ABSTRACT

Retinal lesions signify a cascade of pathological alterations disrupting normal retinal function, which makes their automatic segmentation pivotal for the timely detection of eye diseases. Optical coherence tomography angiography (OCTA) non-invasively visualizes the 3D vascular structure of the retina by analyzing blood flow signals at different depths. However, early microvascular lesions, such as microaneurysms, show minimal changes in blood flow, rendering them subtle and easily missed in OCTA images. Furthermore, the unique OCTA imaging process introduces inherent stripe noise. Existing deep learning-based segmentation algorithms mainly rely on spatial domain information from a single network, making it challenging to accurately capture such subtle changes. To address this problem, we propose a Multi-Domain Fusion Network (MFNet) that captures both spatial and frequency domain features from OCTA images for retinal lesion segmentation. It is the first time to design a novel frequency domain encoder by fusing multi-level Discrete Wavelet Transform (DWT), capturing multi-scale texture features while reducing noise. Moreover, we design a Domain Fusion Module (DFM) that employs a multi-level fusion strategy and gating mechanism to fully integrate spatial and frequency features, addressing the shortcomings of simple concatenation or addition in existing methods. Experimental results show that MFNet outperforms current methods on multiple datasets. For example, on the Diabetic Retinopathy Analysis Challenge 2022 (DRAC2022) dataset, MFNet achieved dice coefficients of 54.48% and 75.36% for neovascularization, intraretinal microvascular abnormalities, and non-perfusion areas, with intersection over union (IoU) values of 65.02%, 37.44%, and 60.47%, respectively. Our code is available at https://github.com/GM-JIa/MFNet.

1. Introduction

1.1. Motivations

The International Diabetes Federation (IDF) predicts that by 2045, the global population of diabetes mellitus (DM) patients will reach 700 million [1]. Retinopathy is a serious complication of diabetes, encompassing various pathological changes that affect the normal structure and function of the retina [2]. These pathological changes include microvascular abnormalities, increased cellular oxidative stress, and metabolic disorders [3–5]. As shown in Fig. 1, common retinal lesions associated with diabetic retinopathy include neovascularization (NV), non-perfusion areas (NPA) and intraretinal microvascular abnormalities (IRMA). The reversibility of retinopathy decreases as the disease progresses. Therefore, early diagnosis and timely intervention are crucial for maximizing the reversible potential of retinopathy [6].

Optical Coherence Tomography Angiography (OCTA) is an emerging, non-invasive eye imaging technology that can display high-resolution blood flow information and vascular structures of the retina and choroid by analyzing blood flow signals across multiple depth layers [7]. Compared to traditional fluorescein angiography, OCTA does not require the injection of contrast agents, significantly reducing patient discomfort and the risk of potential allergic reactions [8]. OCTA technology is based on the principles of Optical Coherence Tomography (OCT), identifying blood vessels by detecting changes in scattered light signals caused by the movement of red blood cells [9]. OCTA can generate three-dimensional images of the retinal and choroidal vascular networks, which is of great importance for the early diagnosis and monitoring of retinal diseases such as age-related macular degeneration, diabetic retinopathy, and other microvascular diseases [10].

* Corresponding author at: School of Computer Science, Qufu Normal University, Rizhao, 276826, China.
E-mail address: mafei@qfnu.edu.cn (F. Ma).

https://doi.org/10.1016/j.bspc.2025.107945
Received 23 June 2024; Received in revised form 28 February 2025; Accepted 14 April 2025
Available online 13 May 2025

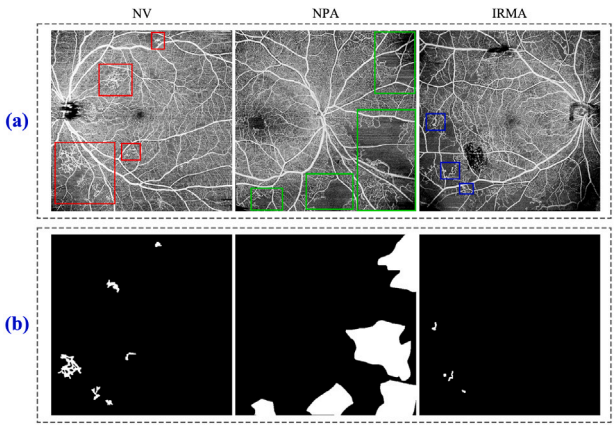


Fig. 1. Illustration of fundus images with retinal lesions. NV: neovascularization; NPA: non-perfusion areas; IRMA: intraretinal microvascular abnormalities. (a) Original images with red, blue and green bounding boxes indicate regions of lesion; (b) The ground truth for lesions.

The accurate segmentation of retinal lesions in ophthalmic diseases is of paramount importance for the diagnosis and subsequent treatment of these conditions. However, since the OCTA imaging principle relies on thresholds to differentiate between static tissue and blood flow, it may lead to instability in detection and alignment in areas with slow blood flow [11]. Early microlesions, such as microaneurysms, which exhibit minimal blood flow changes, may appear extremely subtle in OCTA images and could potentially be overlooked in practical applications [12]. Additionally, eye movements and refractive media opacities in patients may cause displacement variations during the reconstruction process [13], which presents challenges in accurately segmenting the boundary regions of retinal lesions, potentially affecting the assessment of disease progression.

Convolutional Neural Networks (CNNs) have shown the effectiveness in medical image analysis due to their ability to extract and learn image features [14,15]. Although CNNs are effective at extracting local features, they do not have the inherent ability to build global context and long-range dependencies in the images due to their limited receptive field [16]. Due to the specific imaging method of OCTA, inherent stripe noise is present in the images. Effectively addressing this stripe noise is crucial for lesion identification and segmentation. Wavelet transforms are robust mathematical methods used in image denoising [17,18], image reconstruction [19,20] and image segmentation [21]. The wavelet transform is designed to capture effectively texture features by decomposing the image at different scales. In contrast, capturing texture features with traditional CNN architectures is very difficult. While the wavelet transform is good at extracting texture features, it struggles with high-level abstract features of an image. In recent years, the Transformer model has achieved great success in the field of natural language processing and has been gradually introduced into the computer vision [22,23]. The Transformer model and its variants can efficiently construct long-range dependencies between sequences of tokens, which is particularly useful for medical image processing, especially for the complex microscopic lesion regions. However, transformers usually require a large amount of accurately labeled data to achieve good generalization [24]. But it is often difficult to obtain a sufficient number of high-quality medical images. In addition, the transformer models usually require more computational resources due to the presence of a fully connected layer, which limits their application.

1.2. Innovations

To address these issues, we propose a multi-domain fusion network for accurately distinguishing and segmenting retinal lesions in OCTA images. We attempt to leverage the complementary features from both the spatial and frequency domains to accurately segment retinal lesion areas in OCTA images. Unlike traditional methods that rely solely on convolutional operations to extract spatial domain features, we introduce a multi-level Discrete Wavelet Transform to construct a frequency domain encoder. The frequency domain encoder can capture texture features at different scales in OCTA images more effectively, enhancing the ability of the model to detect early microlesions. OCTA images, based on OCT technology, often contain stripe-like noise. The frequency domain encoder can remove noise by decomposing the image and applying thresholding at specific scales while preserving important edge information.

1.3. Contributions

In summary, the main contributions of our work are as follows:

- We propose the multi-domain fusion network for retinal lesion segmentation in OCTA images. To the best of our knowledge, this is the first time to design a novel model based on Discrete Wavelet Transform (DWT) for retinal lesion segmentation in OCTA images. Our study not only demonstrates that frequency domain features can be utilized for lesion segmentation, but also expands the possibilities of interaction between the spatial and frequency domains.
- We establish a targeted feature alignment strategy that effectively aligns and fully integrates spatial and frequency domain information at different scales. Our alignment strategy provides a new perspective for achieving effective interaction between different feature spaces.
- We design a Global Cognition Module (GCM) to capture long-range dependencies between lesions. Results show that the ConvTrans Unit (CTU), with its carefully designed three-branch structure, retains more lesion details during decoding, enhancing the model's overall understanding and perception of the image.
- The extensive experiments on the Diabetic Retinopathy Analysis Challenge 2022 (DRAC2022) dataset [25] and the WF-OCTA dataset [26] show that our approach performs better than other competing methods for the lesion segmentation in OCTA images, such as non-perfused areas, neovascularization and intraretinal microvascular abnormalities.

1.4. Manuscript organization

The remainder of this paper is organized as follows: In Section 2, we introduce related work on Convolutional Neural Networks, Vision Transformers, Wavelet Transforms, and medical image segmentation. Section 3 explains our proposed method in detail, and Section 4 presents the experimental setup and results. Section 5 presents a detailed discussion of model performance, interpretability, and limitations. Finally, we conclude our work in Section 6.

2. Related work

2.1. CNN for image segmentation

In recent years, the convolutional neural networks [14,27–31] have been widely used in various image segmentation tasks by automatically extracting image features and classifying them at the pixel level. The fully convolutional networks (FCNs) can be used to segment images without fully connected layers [32]. FCNs successfully use traditional CNNs for pixel-level prediction and achieve image segmentation in an end-to-end manner. However, the convolutional layers of FCNs are usually stacked sequentially, resulting in a coarse output of FCNs. The U-Net [33] model further improves the accuracy of medical image segmentation through its unique U-shaped encoder–decoder structure. ResUnet effectively alleviates the vanishing gradient problem in deep network by introducing residual connectivity [34]. On the other hand, Jha et al. [35] optimized the feature fusion process by designing a more complex coder–decoder path. Although these CNN-based methods have achieved well performance in image segmentation tasks, they usually have difficulty in capturing long-range dependencies in images.

2.2. Vision Transformers

At present, the Transformer model has been gradually used for natural language processing (NLP) [36] and computer vision [37]. Dosovitskiy et al. [38] proposed the Vision Transformer (ViT), which has achieved well performance on image segmentation tasks. Since the self-attention in Transformer is computationally intensive with squared complexity, many approaches have been devoted to reducing its time. Liu et al. [39] presented a hierarchical Transformer whose representation is computed with shifted windows. The sequel [40] uses even larger window sizes. Touvron et al. [37] introduced a new distillation procedure with a distillation token. This procedure reduces the amount of training data required by the Transformer architecture, which can improve the performance under the zero-shot or few-shot scene. However, recent research [24] suggests that the self-attention does not exhibit any structural bias towards inputs. As a result, Transformers (w/o pre-training) are prone to have the over-fitting problem on small or moderate-sized datasets. Unfortunately, it is challenging to collect a large-scale labeled dataset accurately in the medical field, which limits the application of Transformers for medical image segmentation tasks.

2.3. Segmentation networks for medical images

Due to the specificity and complexity of medical images, it is often necessary to develop some specialized segmentation networks to improve the performance of medical image segmentation tasks. HiFormer can extract two multi-scale feature representations using the Swin Transformer module and a CNN-based encoder [41]. Reza et al. [42] designed a new convolutional neural network called RetiFluidNet for multi-class retinal fluid segmentation. Although RetiFluidNet benefits from hierarchical representation learning of texture, context and edge features, it has a high time complexity. Some recent researches [43–46] have focused on exploiting multi-scale features for accurate segmentation of medical images. For example, Bougourzi et al. [47] introduced PAG-TransYnet, which fuses CNN and Transformer mechanism. PAG-TransYnet uses attention mechanism in a Pyramid encoder architecture at different scales. Wang et al. [48] proposed a new network (MsTGANet), which is based on transformer module and for segmenting drusen in OCT images. MsTGANet aims to capture multi-scale non-local features with long-range dependencies from different layers of the encoder. Although these studies can extract the features in the spatial domain, they do not fully mine the rich frequency domain properties of images.

2.4. Wavelet transform

The wavelet transform is a method of decomposing an input signal of interest into a set of elementary waveforms, called wavelets. This provides a means of analyzing waveforms that are bounded in both frequency and duration [49]. The two primary categories of wavelet transforms include discrete wavelet transforms (DWT) and continuous wavelet transforms (CWT). Discrete wavelet transform was initially used in the field of signal analysis and has been gradually introduced in the field of image processing [50,51]. Imtiaz et al. [52] designed a new method to incorporate a boundary-aware unit with an attention mechanism based on wavelet domain in each stage of the encoder–decoder output. Ramya et al. [53] designed an effective method based on DWT for segmenting skin lesions. However, there exist limitations with highly complex or low contrast skin images.

3. Methodology

Fig. 2 shows the illustration of our proposed framework. Here, the spatial encoder branch and the frequency encoder branch interact with each other through multiple levels of domain fusion modules to exchange information at different scales. The conditional feature layer is used to introduce additional supervisory information beyond the encoder and decoder. The global cognition module improves segmentation boundaries by capturing long-range dependencies in the image. The detailed steps of this process are described in Algorithm 1. In the next section, we will describe the various components of MFNet in detail.

Algorithm 1 MFNet for Lesion Segmentation

Input: OCTA image $X \in \mathbb{R}^{H\times W\times C}$, model parameters Θ
Output: Segmented lesion map $\hat{Y}$
Initialize parameters Θ
for each epoch **do**
  $F_s \leftarrow \text{SpatialEncoder}(X, \Theta_s)$
  $\{F_{h1}, F_{h2}, F_l\} \leftarrow \text{FrequencyEncoder}(X, \Theta_f)$
  for each layer i in encoder **do**
    $F_s \leftarrow \text{Fuse}(F_s[i], F_{h1}[i], \Theta_{sfu1})$
    $F_s \leftarrow \text{Fuse}(F_s, F_{h2}[i], \Theta_{sfu2})$
    $F_s \leftarrow \text{Fuse}(F_s, F_l[i], \Theta_{cfu})$
  end for
  $F_{\text{cond\_enc}} \leftarrow \text{SharedEncoder}(X, \Theta_s)$ ▷ Shared with Spatial Encoder
  $F_{\text{cond}} \leftarrow \text{ConditionalDecoder}(F_{\text{cond\_enc}}, \Theta_{cfl})$
  for each layer j in decoder **do**
    $F_{\text{up}} \leftarrow \text{Upsample}(F_{\text{fused}}[j], \Theta_{\text{dec}}[j])$
    $F_{\text{up}} \leftarrow F_{\text{up}} + F_s[j]$ ▷ Skip connection with spatial features
  end for
  $F_{\text{final\_dec}} \leftarrow F_{\text{up}}$ ▷ Final decoded output
  $F_{\text{final}} \leftarrow F_{\text{final\_dec}} \times F_{\text{cond}}$ ▷ Multiply with conditional features
  $\hat{Y} \leftarrow \text{GlobalCognitionModule}(F_{\text{final}}, \Theta_{gcm})$
end for
return $\hat{Y}$

3.1. Encoder

3.1.1. Spatial encoder

In our designed spatial encoder, we use CNN as a feature extractor to obtain pyramid feature maps with different spatial resolutions. The input image $X \in R^{H\times W\times C}$ is with spatial dimensions H, W, and C channels. It is first fed into the spatial encoder. The spatial encoder consists of six layers, each of which is connected to the layer of the corresponding decoder through skip connections to compensate for missing information at the lower levels of the decoder and to recover local spatial information. The outputs of the second and final layers of the spatial encoder are connected through the Domain Fusion Module (DFM) to the frequency encoder.

3.1.2. Frequency encoder

In our MFNet, the main role of the frequency encoder is to extract high-quality frequency domain features from the OCTA image to capture the texture information in the image, which is crucial for understanding the details of the retinal lesion region. The frequency encoder is illustrated in Fig. 3. The multi-level discrete wavelet transform (DWT) was selected for its robust multi-resolution analysis capability, which decomposes images into low-frequency subbands that capture structural information and high-frequency subbands that capture edges and finer details. This hierarchical decomposition enables the encoder to effectively capture image features across different scales, allowing the model to process both intricate microvascular structures and broader retinal regions in OCTA images. Furthermore, as OCTA image

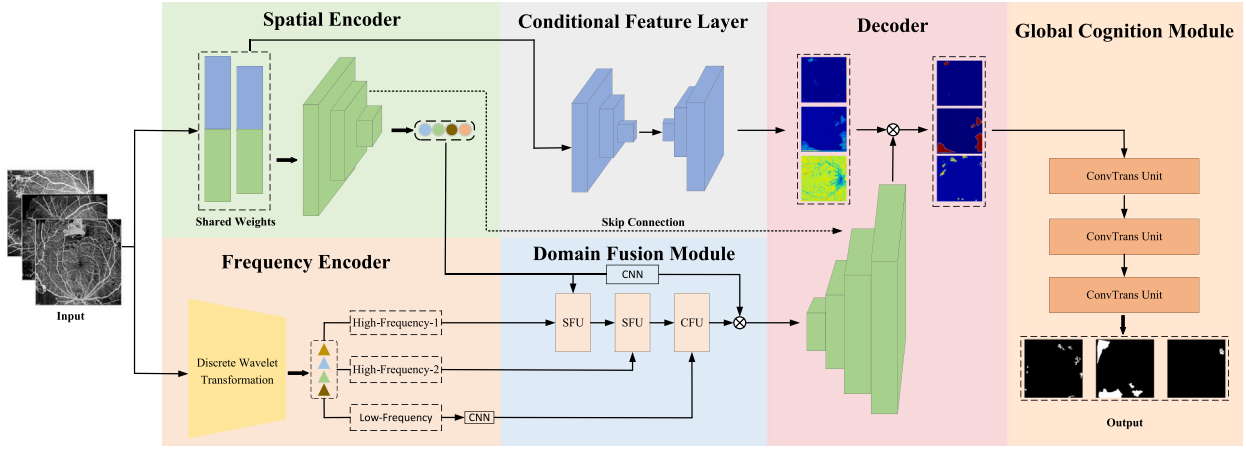


Fig. 2. Overview of our Multi-Domain Fusion Network (MFNet). MFNet employs a fusion of CNN and DWT to extract both spatial and frequency features from OCTA images. Furthermore, it integrates information from disparate domains using DFM. MFNet employs a conditional feature layer to provide supplementary supervisory information for the lesion segmentation process, subsequently enhancing the segmentation boundaries through the global cognition module.

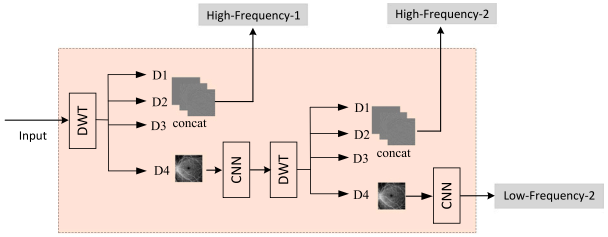


Fig. 3. The Frequency Encoder. The frequency encoder decomposes the input image into four frequency subbands through a two-stage DWT, where the first stage captures the basic frequency information and the second stage captures the more complex texture details, thus constructing a feature set that contains multi-scale texture and edge features.

noise often appears in high-frequency regions, DWT allows for effective noise separation, preserving essential features while suppressing noise. Its sparse representation also enhances computational efficiency, making it a flexible and efficient encoding choice.

To effectively extract texture features from OCTA images, we have developed two levels of the DWT Unit in the frequency encoder, each focused on extracting features at different scales. The DWT Unit works by decomposing the input image into multiple frequency subbands, a process that can be represented by the following equation:

$$X_{l+1} = DWT\left(X_l\right), \tag{1}$$

where X_l is the input image of the lth layer, while X_{l+1} is a set of subband images obtained after the DWT. In our structure, the DWT module contains two levels of transform. The first level of the DWT transform captures the underlying frequency information, while the second level allows us to capture more complex image textures that may not be fully represented at the first level. With this multi-scale decomposition, we can create a frequency domain feature set that contains both coarse and detailed texture information.

In each DWT Unit, we first perform a 2D discrete wavelet transform on the input feature map I. The four frequency sub-bands obtained capture the texture and edge information in different directions of the image. For each pixel (x,y), the transform can be expressed as:

$$LL(x,y) = \sum_{m,n} h(m)\cdot h(n)\cdot I(2x+m, 2y+n), \tag{2}$$

$$LH(x,y) = \sum_{m,n} h(m)\cdot g(n)\cdot I(2x+m, 2y+n), \tag{3}$$

$$HL(x,y) = \sum_{m,n} g(m)\cdot h(n)\cdot I(2x+m, 2y+n), \tag{4}$$

$$HH(x,y) = \sum_{m,n} g(m)\cdot g(n)\cdot I(2x+m, 2y+n), \tag{5}$$

where $h(m)$ and $h(n)$ represent the low-pass filters and $g(m)$ and $g(n)$ represent the high-pass filters. $LL(x,y)$ denotes the low-frequency part, $LH(x,y)$ is the horizontal high-frequency part, $HL(x,y)$ represents the vertical high-frequency part and $HH(x,y)$ denotes the high-frequency part.

3.2. Domain fusion module

Fig. 4 shows the illustration of Domain Fusion Module (DFM), which is designed to integrate spatial and frequency domain features for retinal lesion segmentation. The DFM consists of three stages, each of which can facilitate the spatial and frequency domain features at different scales.

3.2.1. Spatial fusion unit

The spatial fusion unit is illustrated in Fig. 5. The spatial fusion unit is located in the first and second layers of the DFM and is the key module used to process and fuse feature maps from different input domains. $X_f \in R^{H_f \times W_f \times C}$ is the frequency feature, where H_f, W_f and C denote spatial dimensions and channel number respectively. X_f is first fed into a convolution layer with a convolution kernel size 3×3, which is used to extract features in the frequency domain. The convolution is followed by a batch normalization layer and a ReLu activation function to stabilize the learning process and speed up training. Next, the output of the convolutional layer is passed through a channel-dimension max-pooling layer and an Next, the output of the convolutional layer is passed through a channel-dimension max-pooling layer and an average pooling layer, respectively. Then, two intermediate feature maps of the form $X_f \in R^{H_f \times W_f \times 1}$ are obtained. This process reduces the spatial dimensions of the feature maps and improves the noise immunity of the model.

The frequency domain feature $F_w \in R^{H_f \times W_f \times 1}$ are then obtained by sequentially passing through a convolutional layer with a convolutional kernel size 3×3, a batch normalization layer and a sigmoid activation function. Given a spatial input $X_s \in R^{H_s \times W_s \times C}$ with spatial dimensions H_s and W_s and channels C, X_s is first passed through a max-pooling layer to reduce the dimensions of the feature map while preserving important spatial information. These feature maps are then fed into a 3×3 convolutional layer and a ReLU activation function. The spatial domain features $X_s \in R^{H_f \times W_f \times C}$ are then Hadamard transform with the frequency domain weight $F_w \in R^{H_f \times W_f \times 1}$ to perform feature recalibration. This recalibration method allows the model to enhance important features and suppress irrelevant ones, thus better retaining

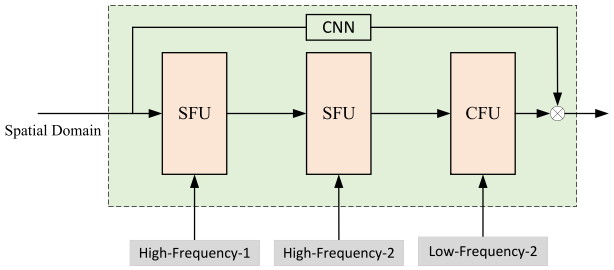


Fig. 4. The Domain Fusion Module (DFM). The DFM is a complex multi-stage process involving two types of underlying components, the Spatial Fusion Unit and the Channel Fusion Unit.

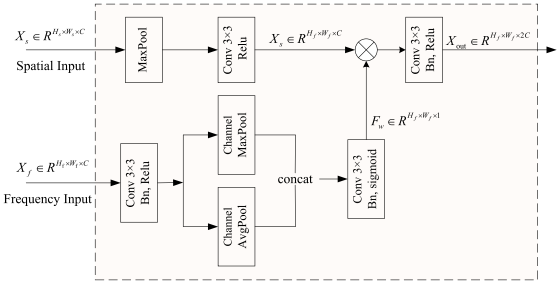


Fig. 5. Spatial Fusion Unit (SFU). The SFU first processes the frequency domain inputs along the channel dimensions and then recalibrates the spatial domain features using the generated frequency domain weighting coefficients to finally output the fused features.

the information useful for the final task. Finally, the calibrated features are fed into the final convolutional layer, changing the number of output channels to obtain the final output feature $X_{out} \in R^{H_f \times W_f \times 2C}$.

In the first stage of the domain fusion module, we use the SFU to integrate the second-stage output of the spatial encoder with the high-frequency-1 obtained from the first-stage DWT. The SFU is designed to take into account the different nature of the features, i.e. the spatial features in the output of the CNN versus the frequency domain features in the output of the DWT. This fusion process can be defined as:

$$SFU_{out}^{(1)} = f\left(S_{CNN_2}, H_{DWT_1}; \theta_{SFU}\right), \tag{6}$$

where S_{CNN_2} is the output of the second stage of the spatial encoder, H_{DWT_1} is the high frequency component of the first stage of the DWT, θ_{SFU} is the parameter of the SFU, $f(\cdot)$ is the fusion function of the SFU.

The second stage includes a SFU which is responsible for integrating the output of the first stage SFU with the High-Frequency-2 obtained from the second stage DWT. This step further strengthens the high frequency features and increases the sensitivity of the model to the details of the OCTA image. Similarly, this fusion process can be represented as:

$$SFU_{out}^{(2)} = f\left(SFU_{out}^{(1)}, H_{DWT_2}; \theta_{SFU}^{(2)}\right), \tag{7}$$

where $SFU_{out}^{(1)}$ is the output of the first SFU, H_{DWT_2} is the second high frequency component of the second stage DWT, $\theta_{SFU}^{(2)}$ is the parameter of the second SFU.

3.2.2. Channel fusion unit

The channel fusion unit is located in the last layer of the DFM and the main purpose of the CFU is to efficiently integrate both spatial and frequency features while preserving them. Fig. 6 shows the architecture of the proposed channel fusion unit. Given a frequency input $X_f \in R^{H_f \times W_f \times C}$ with spatial dimensions H_f and W_f, and channels C, the CFU first feeds X_f into the global average pooling layer and the maximum pooling layer, and obtains two intermediate feature maps in the form of $X_f \in R^{1\times1\times C}$. They are immediately fed into two convolutional layers and the ReLu activation layer to refine and enhance

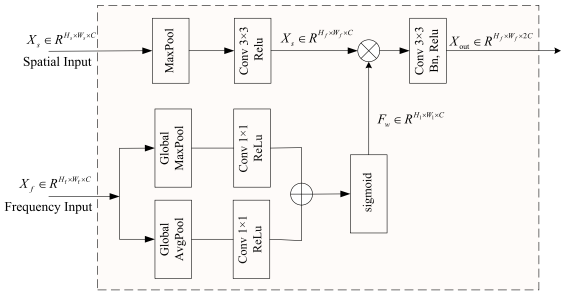


Fig. 6. Channel Fusion Unit (CFU).

the pooled frequency domain features, respectively. The obtained two features are then summed and passed through a sigmoid activation function to obtain the frequency domain feature $F_w \in R^{1\times1\times C}$. Similar to the CFU, the spatial domain inputs are sequentially passed through a pooling layer and a convolutional layer and the frequency domain weights are Hadamard multiplied to perform feature recalibration. Finally, the calibrated merged features are passed through the final convolutional layer and the number of output channels is changed to obtain the final output features $X_{out} \in R^{H_f \times W_f \times 2C}$.

The third stage of the DFM uses a CFU which focuses on fusing the outputs of the low frequency and the previous SFU to produce a composite representation of the features. The operation of the CFU can be represented as follows:

$$CFU_{out} = g\left(SFU_{out}^{(2)}, L_{DWT_2}; \theta_{CFU}\right), \tag{8}$$

where $SFU_{out}^{(2)}$ represents the output of the second SFU. L_{DWT_2} is the low frequency of the second level DWT. θ_{CFU} denotes the parameter of the CFU. $g(\cdot)$ is the fusion function of the CFU.

3.3. Decoder

In MFNet, we design a decoder to reconstruct the feature maps learned by the frequency encoder, spatial encoder and DFM. The decoder with hierarchical multi-scale feature reconstruction will make progressive feature recovery of the feature maps from coarse to fine. Assuming that the feature of the lth layer in the spatial encoder is F_{enc}^{l} and features of the lth layer in the decoder is F_{dec}^{l}, the feature maps of the previous layer of the decoder, F_{dec}^{l-1}, are first up-sampled by inverse convolution, which increases the spatial dimensions of the feature maps and reduces the number of channels. Then, the up-sampling feature map is concatenated with the feature map F_{dec}^{l} of the corresponding layer of the spatial encoder using skip connections to achieve direct fusion of the deeper features of the encoder and the shallower features of the spatial encoder. Finally, the concatenated feature map is refined by two convolutional layers and the feature representation is refined by a batch normalization function and a nonlinear activation function as follows:

$$F_{dec}^{l} = Conv(Concat(TransConv(F_{dec}^{l-1}),\ F_{dec}^{l})), \tag{9}$$

where $TransConv(\cdot)$ is the inverse convolution operation, $Concat(\cdot)$ concatenates features along the channel dimension and $Conv(\cdot)$ is the convolution operation. This progressive feature recovery strategy is the key to achieving high-precision medical image segmentation.

3.4. Global cognition module

The Global Cognition Module (GCM), a key component of MFNet. It processes the entire image to improve the ability of the model to identify and segment complex lesion regions. In our architecture, the GCM consists of three ConvTrans units that capture global features. Unlike traditional Transformer models, which are used for sequence

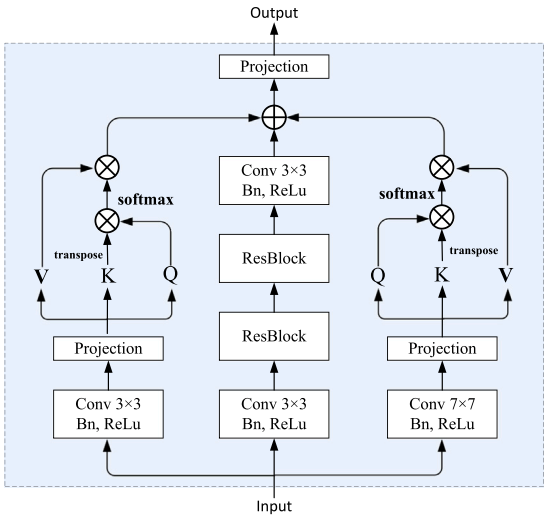


Fig. 7. ConvTrans Unit. In order to more effectively accommodate the characteristics of image data, each CTU incorporates two enhanced multi-head self-attention branches and one convolution branch.

data, the multi-head self-attention mechanism in our ConvTrans Unit (CTU) is specifically redesigned for 2D medical images.

Fig. 7 depicts the architectural configuration of the proposed CTU. In the CTU, the input image passes through three convolutional layers for initial feature extraction, followed by batch normalization and a ReLU activation function. Each CTU contains two multi-head self-attention branches and one convolution branch. The self-attention branches project the feature map into three matrices: query (Q), key (K), and value (V). These projections are generated through convolutional layers, which preserve the spatial structure of the image while learning relationships between different regions. The attention score is computed by taking the dot product of Q and K, followed by softmax normalization to generate the attention weights. These attention weights are applied to the values (V) to emphasize important regions of the input image. The two self-attention branches use different convolutional layers and projection spaces, allowing the CTU to capture diverse image features. The convolution branch adjusts the input feature map to match the dimension of the attention branches, using a residual block to combine the attention-weighted features. Ultimately, the CTU integrates global and local features, improving the accuracy of lesion segmentation.

3.5. Conditional feature layer

The Conditional Feature Layer is used to induce the model to adaptively learn how to perform feature selection and fusion by introducing additional information beyond the encoder and decoder. The CFL can generate specific conditional features based on different lesions in the OCTA image. The CFL consists of several convolutional layers, batch normalization layers and ReLu activation function layers. The last convolutional layer with a sigmoid function is used to fit the input feature map to a single channel probabilistic map. The spatial encoder input is also used as the CFL input. The output of the CFL is weighted and fused with the output of the decoder to provide the decoder with additional contextual information.

3.6. Loss function

In our model, we use a hybrid loss function that combines the Binary Cross-Entropy (BCE) loss and the Dice loss. The BCE loss is commonly used in binary classification tasks to compute pixel-wise errors between predicted outputs and ground truth masks. In contrast, the Dice loss is particularly effective for dealing with class imbalance in segmentation tasks, as it measures the overlap between the predicted and actual masks. To balance these two loss components, a learnable weight parameter, α, is introduced. The initial value of α is set to 0.5 and is dynamically adjusted during training. The total loss function is defined as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{BCE}} + (1-\alpha) \cdot \mathcal{L}_{\text{Dice}}, \tag{10}$$

where the BCE loss is formulated as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \tag{11}$$

and the Dice loss is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \epsilon}, \tag{12}$$

where p_i denotes the predicted probability, y_i is the ground truth, N is the total number of samples, and ϵ is a smoothing factor to ensure numerical stability.

4. Experiments and results

4.1. Datasets

We evaluated the performance of all methods on three public datasets. (1) The DRAC2022 Challenge provides a standardized ultra-wide (swept source) optical coherence tomography angiography (UW-OCTA) dataset with images and grade labels for automated image quality assessment, lesion segmentation and DR grading. There exist three types of lesions in the DRAC2022 lesion segmentation dataset, which are intraretinal microvascular abnormalities, non-perfusion areas and neointimal formation. There are 109 raw images in the dataset, including the ground truth label of each image. Each raw image may contain multiple types of lesion. Specifically, the dataset contained 86 intraretinal microvascular abnormalities, 106 non-perfusion areas and 35 neovascularization. (2) Dai et al. collected WF-OCTA images of the retina of 288 individuals with diabetes. Two professional ophthalmologists from the Sixth People's Hospital, Shanghai Jiao Tong University, Shanghai, China, annotated the WF-OCTA for clinical diagnosis. We used 80 of these, which contained areas of non-perfusion. (3) The Indian Diabetic Retinopathy Image Dataset (IDRiD) [54] provides a standardized color fundus image dataset for diabetic retinopathy research. This dataset is included to evaluate the generalizability of our methods, with a variety of lesion types annotated for segmentation and grading tasks.

4.2. Implementation details

(1) Experimental Settings: The proposed network framework was implemented in Pytorch and trained on a workstation equipped with an NVIDIA GeForce GTX A4000 GPU with 16 GB of RAM. During training, the AdamW optimizer was used to update the model parameters. AdamW was chosen for its ability to decouple weight decay from the gradient update process, allowing for more effective regularization. This leads to better generalization, especially in high-dimensional tasks such as image segmentation, where overfitting and complex parameter spaces are common. The initial learning rate was set to 0.0001, which balances stability and convergence, ensuring that the model progresses steadily in the early stages of training. A weight decay of 0.01 was applied to control model complexity and prevent overfitting. The batch size was set to 4, as a trade-off between memory limitations and the need for frequent updates during training.

(2) Data Preprocessing: For each lesion, flipping, rotation, cropping, scaling, brightness-contrast adjustment and affine transformation were performed to add diversity to the data. Data augmentation is a useful method for training a model to avoid overfitting problems. This involves normalizing all images to a range of [0, 1] instead of [0, 255] before inputting them into the model.

Table 1
Segmentation results of NV, NPA and IRMA on the DRAC2022 dataset.

Lesion	Methods	DICE	IoU	SEN	PRE	SPE
NV	Unet [33]	0.7286	0.5730	0.7624	0.6976	0.9975
	ResUnet [34]	0.7417	0.5894	0.7849	0.7029	0.9975
	ResUnet++ [35]	0.7479	0.5973	0.7701	0.7269	0.9978
	RetiFluidNet [42]	0.7583	0.6106	0.7517	0.7650	0.9981
	MsTGANet [48]	0.7655	0.6201	0.8029	0.7315	0.9978
	PAG-TransYnet [47]	0.7546	0.6058	0.7414	**0.7681**	0.9945
	MFNet	**0.7880**	**0.6502**	**0.8113**	0.7660	**0.9983**
NPA	Unet [33]	0.7044	0.5437	0.7441	0.6688	0.9366
	ResUnet [34]	0.7222	0.5652	0.7364	0.7087	0.9479
	ResUnet++ [35]	0.7195	0.5619	0.7354	0.7043	0.9469
	RetiFluidNet [42]	0.7316	0.5768	**0.7991**	0.6746	0.9337
	MsTGANet [48]	0.7455	0.5943	0.7708	0.7218	**0.9535**
	PAG-TransYnet [47]	0.7396	0.5868	0.7511	0.7285	0.9519
	MFNet	**0.7536**	**0.6047**	0.7728	**0.7355**	0.9522
IRMA	Unet [33]	0.4898	0.3243	0.3833	0.6782	0.9990
	ResUnet [34]	0.4950	0.3289	0.4033	0.6408	0.9992
	ResUnet++ [35]	0.5110	0.3431	0.4184	0.6561	0.9992
	RetiFluidNet [42]	0.5226	0.3537	0.4174	0.6986	0.9994
	MsTGANet [48]	0.5228	0.3539	0.4104	0.7201	0.9995
	PAG-TransYnet [47]	0.5106	0.3428	0.3862	0.7533	0.9993
	MFNet	**0.5448**	**0.3744**	**0.4239**	**0.7623**	**0.9996**

Table 2
Segmentation Results of NPA on the WF-OCTA Dataset.

Lesion	Methods	DICE	IoU	SEN	PRE	SPE
NPA	Unet [33]	0.7277	0.5720	0.6882	0.7720	0.9535
	ResUnet [34]	0.7124	0.5533	0.6849	0.7422	0.9455
	ResUnet++ [35]	0.6923	0.5295	0.6390	0.7553	0.9526
	MsTGANet [48]	0.7540	0.6051	0.7358	0.7731	0.9506
	RetiFluidNet [42]	0.7537	0.6048	0.7404	0.7675	0.9487
	PAG-TransYnet [47]	0.7246	0.5681	0.6776	0.7786	**0.9559**
	MFNet	**0.7869**	**0.6486**	**0.7790**	**0.7949**	0.9540

4.3. Evaluation metrics

We use five metrics to evaluate the performance of all methods as follows:

- Dice coefficient (DICE) = 2×TP/(FP+2×TP+FN));
- Intersection over Union (IoU) = TP/(TP+FP+FN);
- Sensitivity (SEN) = TP/(TP+FN);
- Specificity (SPE) = TN/(TN+FP);
- Precision (PRE) = TP/(TP+FP);

where TP (True Positives) is the number of correctly predicted lesion pixels and TN (True Negatives) is the number of correctly predicted non-lesion (normal) pixels. FP (False Positives) refers to the non-lesion pixels that were incorrectly predicted as lesion pixels and then FN (False Negatives) refers to the lesion pixels that were incorrectly predicted as non-lesion pixels.

4.4. Quantitative comparisons

In this section, we evaluate the performance of the MFNet model on the tasks of IRMA, NPAs and NV segmentation. To demonstrate the superiority of the proposed method on different tasks, we present the quantitative results of our method on several segmentation tasks in Tables 1 and 2.

4.4.1. Results on the DRAC2022 dataset

Table 1 shows the segmentation results on theDRAC2022 dataset with the competitive methods. MFNet achieved the best performance on three lesion types in the DRAC2022 dataset, especially excelling in NV and NPA segmentation tasks. In the NV segmentation, MFNet attained a DICE score of 78.8%, outperforming MsTGANet (76.55%) by 2.25%. Furthermore, MFNet achieved higher IoU (65.02%) and sensitivity (81.13%) than PAG-TransYnet, with respective improvements of 4.44% and 7.99%. These results underscore the strength of MFNet in precise boundary localization and refinement, attributed primarily to its DFM, which enables comprehensive integration of spatial and frequency features, effectively capturing the intricate characteristics of NV regions.

In the NPA segmentation task, MFNet demonstrated a leading DICE score of 75.36%, surpassing all other methods. In comparison, MsTGANet and RetiFluidNet achieved DICE scores of 74.55% and 73.16%, respectively. MFNet also ranked highest in both IoU (60.47%) and precision (73.55%), highlighting its capacity to handle low-contrast, irregular NPA shapes. The ability of the model to extract multi-scale texture information through the DWT was critical in reducing missed detections and enhancing edge detection accuracy.

For the IRMA segmentation, the most challenging task, MFNet again delivered the best performance with a DICE score of 54.48% and IoU of 37.44%. This represents a 2.2% and 2.22% improvement in DICE over MsTGANet and RetiFluidNet, respectively. Additionally, MFNet achieved superior precision (76.23%) and specificity (99.96%), indicating high reliability in identifying intraretinal microvascular abnormalities. The GCM also contributed to the performance of MFNet by enhancing the ability of the model to capture global dependencies and fine-scale details, thereby enabling more accurate detection of complex IRMA boundaries.

4.4.2. Results on the WF-OCTA dataset

Table 2 shows the segmentation results on the WF-OCTA Dataset with the competitive methods. One can see that the proposed MFNet demonstrated superior performance compared to state-of-the-art segmentation models and medical segmentation models on the most performance metrics. Specifically, it outperformed the state-of-the-art MsTGANet by 3.29% in DICE and 4.39% in IoU. Furthermore, it surpassed the RetiFluidNet by 3.85% in SEN. The advantage is that MFNet can learn multiple domains features in OCTA images, allowing it to

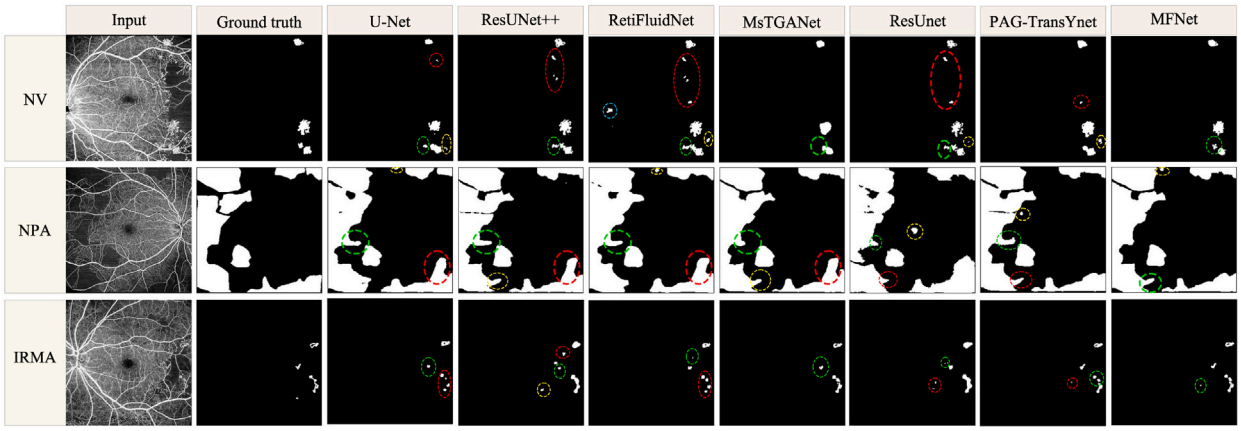


Fig. 8. The segmentation of different methods on the DRAC2022 dataset. Each row: the results of NV segmentation, NPA segmentation and IRMA segmentation, respectively.

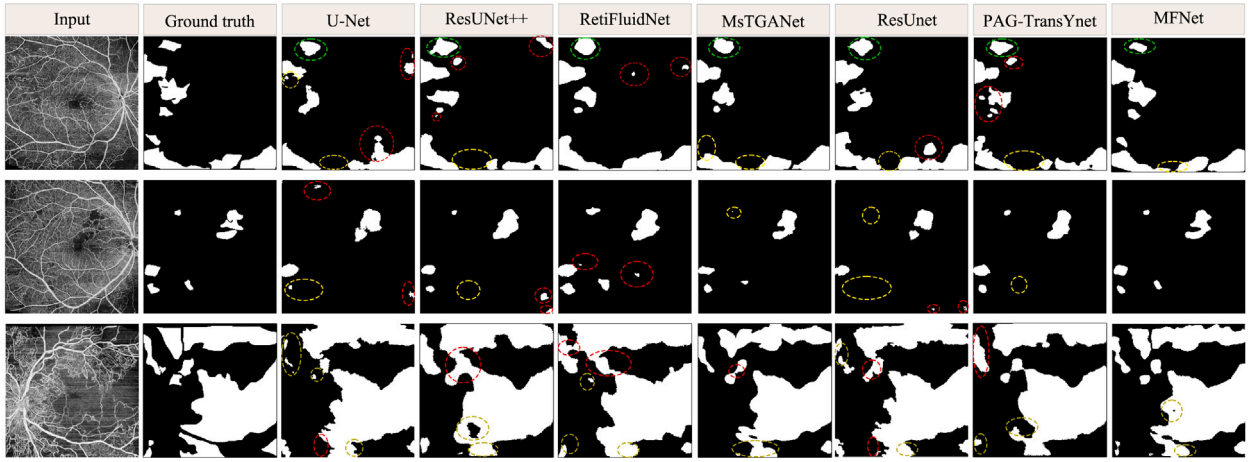


Fig. 9. The segmentation of different methods on the WF-OCTA dataset.

simultaneously extract and integrate information from both spatial and frequency domains. Compared with the single-domain networks, this multi-domain learning framework is the key to improving segmentation performance.

4.5. Lesion segmentation visualization

This section visualizes the segmentation results of different models, with the objective of evaluating the effectiveness of the MFNet model for retinal image segmentation. The efficacy of the MFNet model in addressing complex retinal lesion segmentation can be discerned from the series of visualization results presented in Figs. 8 and 9.

The results demonstrate that our proposed method can improve the accuracy of the retinal lesion segmentation. As shown in Fig. 8, our approach achieves very competitive segmentation performances of the IRMA lesion, which shows the effectiveness of our proposed method in producing accurate pixel-wise lesion segmentation. Our approach can obtain more accurate lesion edges and complex structures. However, the other methods such as U-Net and ResUnet++ have many disconnected areas and erroneous segmentation in complex cases. Although RetiFluidNet and MsTGANet are specially designed for OCTA images, their overall segmentation performance is lower than MFNet.

Fig. 9 illustrates the segmentation results of non-perfusion areas (NPA) on the WF-OCTA dataset, comparing the proposed MFNet against several state-of-the-art models, including U-Net, ResUNet++, RetiFluidNet, MsTGANet, ResUnet, and PAG-TransYnet. The first and second columns show the input OCTA image and the corresponding ground truth, respectively. In the lesion segmentation visualization, green circles indicate regions where there are shape differences between the segmentation boundaries and the ground truth, while the circles in red and yellow color indicate over-segmentation and under-segmentation areas, respectively.

Among all the methods, MFNet demonstrates superior segmentation accuracy. The advantage is to design the multi-domain fusion, which integrates both spatial and frequency domain information. Specifically, MFNet utilizes the Discrete Wavelet Transform (DWT) to capture fine texture features and reduce noise, enabling more accurate boundary alignment, as shown in the green-circled regions. In comparison, methods such as U-Net and ResUNet++ exhibit over-segmentation in certain regions, highlighted by the red circles, where unnecessary areas are incorrectly identified as non-perfused. This over-segmentation issue is less pronounced in MFNet, thanks to its Domain Fusion Module (DFM), which allows for a more refined integration of features from both domains. However, it is important to note that MFNet still exhibits some under-segmentation in some regions, such as the area marked by the yellow circles. However, MFNet achieves a better overall balance between sensitivity and specificity compared to other models and can reduce the over-segmentation and more accurate segmentation of NPA boundaries.

4.6. Ablation study

In this section, we conduct an ablation study on the DRAC2022 dataset to evaluate the impact of key modules in MFNet on segmentation performance.

Table 3
Effects of DFM modules.

Lesion	Methods	DICE	IoU	SEN	PRE	SPE
NV	MAX	0.7568	0.6088	0.7780	0.7367	0.9979
	MIN	0.7460	0.5948	0.7788	0.7158	0.9977
	AVG	0.7461	0.5950	0.7624	0.7304	0.9979
	SUM	0.7680	0.6234	0.7762	0.7599	**0.9984**
	DFM	**0.7880**	**0.6502**	**0.8113**	**0.7660**	0.9983
NPA	MAX	0.7468	0.5959	0.7603	0.7338	0.9521
	MIN	0.7391	0.5862	0.7521	0.7265	0.9513
	AVG	0.7211	0.5638	0.7701	0.6780	0.9371
	SUM	0.7503	0.6004	**0.7791**	0.7236	0.9488
	DFM	**0.7536**	**0.6047**	0.7728	**0.7355**	**0.9522**
IRMA	MAX	0.5091	0.3414	0.3942	0.7183	0.9994
	MIN	0.5294	0.3601	0.4327	0.6818	0.9993
	AVG	0.5141	0.3460	0.4039	0.7071	0.9994
	SUM	0.5230	0.3541	0.4119	0.7163	0.9994
	DFM	**0.5448**	**0.3744**	**0.4239**	**0.7623**	**0.9996**

Table 4
Effects of DWT modules.

Lesion	DWT	DICE	IoU	SEN	PRE	SPE
NV	×	0.7680	0.6234	0.7762	0.7599	0.9979
	✓	**0.7880**	**0.6502**	**0.8113**	**0.7660**	**0.9983**
NPA	×	0.7415	0.5892	0.7620	0.7221	0.9496
	✓	**0.7536**	**0.6047**	**0.7728**	**0.7355**	**0.9522**
IRMA	×	0.5220	0.3532	0.4148	0.7041	0.9994
	✓	**0.5448**	**0.3744**	**0.4239**	**0.7623**	**0.9996**

Table 5
Performance on GCM modules.

Lesion	GCM	DICE	IoU	SEN	PRE	SPE
NV	×	0.7605	0.6135	0.7698	0.7514	0.9981
	✓	**0.7880**	**0.6502**	**0.8113**	**0.7660**	**0.9983**
NPA	×	0.7464	0.5954	0.7667	0.7271	0.9505
	✓	**0.7536**	**0.6047**	**0.7728**	**0.7355**	**0.9522**
IRMA	×	0.5358	0.3660	0.4281	0.7162	0.9994
	✓	**0.5448**	**0.3744**	**0.4239**	**0.7623**	**0.9996**

4.6.1. Effects of DFM

To investigate the effectiveness of DFM in each task, we replace DFM with MAX, MIN, AVG and SUM operations. Table 3 shows the lesion segmentation results on the DRAC2022 dataset. One can observe that the segmentation performance of our model have decreased on three types of lesions while removing the DFM module. Specifically, the DICE and IoU metrics decreased significantly. As the above analysis, one can see that DFM plays a central role in our proposed framework.

4.6.2. Effects of DWT

DWT can capture the texture features of images, which is a crucial component of MFNet. To verify its effectiveness, we replace the original DWT with the corresponding convolution operation. In Table 4, one can see that the results are significantly degraded without DWT. For example, the DICE and IoU metrics of the model on the segmentation task of IRMA lesions decreased by 2.29% and 2.12%, respectively. These results demonstrate the effectiveness of DWT in the proposed framework.

4.6.3. Effects of GCM

The GCM is one of the important components in MFNet, which uses CTU to capture long-range dependencies. Table 5 shows the effects of GCM. After removing the GCM, the model degrades to the low effective network. By adding the GCM, the performance is promoted to 78.80%, 75.36% and 54.48% for segmenting NV, NPA and IRMA, respectively. It demonstrates that the GCM can improve the performance of the proposed framework.

Table 6
Comparison of model parameters and GFLOPs.

Model	Parameters (M)	GFLOPs
Unet [33]	7.8	17.54
ResUnet [34]	8.3	19.29
ResUnet++ [35]	9.2	19.98
MsTGANet [48]	30.57	67.86
RetiFluidNet [42]	34.26	71.22
PAG-TransYnet [47]	23.61	54.19
MFNet	27.76	59.88

5. Discussion

5.1. Heatmap visualization

We performed a detailed analysis of the heatmaps shown in Fig. 10 to evaluate the performance of the model in handling different types of retinal diseases.

It is clear that the spatial and frequency encoders each have their own advantages, allowing the model to capture both large-scale structures and finer details. The spatial encoder is particularly effective at detecting macroscopic changes. For example, in the case of neovascularization (NV), it accurately localizes abnormal regions around the retinal vessels and optic disc. The frequency encoder, on the other hand, excels at detecting subtle texture variations, showing strong sensitivity to high-frequency information in all lesions, especially at the edges of vascular bifurcations and abnormal areas.

The Domain Fusion Module (DFM) further enhances the ability of the model to detect critical regions by fusing spatial and frequency domain features. For example, in NV detection, the DFM enables the model to focus on neovascular areas while suppressing non-vascular areas. In non-perfusion area (NPA) detection, DFM helps to sharpen lesion boundaries. The fusion capability of DFM enhances the accuracy of the model in recognizing complex lesion contours while reducing attention to irrelevant regions.

The heatmaps also show that the decoder provides an initial localization of the lesion areas. Although some false positives and missed detections remain at this stage, the decoder provides a solid foundation for more refined processing. The Conditional Feature Layer (CFL), in conjunction with the decoder, enhances the attention of the model to lesion regions. For example, in NV, the highlighted neovascular areas become more concentrated, and in NPA, the boundaries of the non-perfused areas are more clearly defined.

Finally, the Global Cognition Module (GCM) plays a critical role by incorporating the broader context of the entire image, enabling the model to more accurately assess potential lesion areas. In the case of intraretinal microvascular abnormalities (IRMA), the GCM successfully filters out some false positive regions while retaining the true locations of IRMA lesions. This global perspective significantly improves the overall accuracy and reliability of the model.

5.2. Model complexity comparison

Table 6 provides a comprehensive comparison of the model complexity in terms of the number of parameters and GFLOPs (Giga Floating Point Operations per second) for seven segmentation models: Unet, ResUnet, ResUnet++, MsTGANet, RetiFluidNet, PAG-TransYnet, and our proposed MFNet. The number of parameters is represented in millions (M), and the GFLOPs measure the computational cost.

From Table 6, one can see that traditional models like Unet and ResUnet have relatively low parameter counts (7.8M and 8.3M, respectively) and correspondingly low GFLOPs (17.54 and 19.29), indicating lower computational complexity. These models are well-suited for scenarios with limited computational resources but may be less effective in handling more complex segmentation tasks due to their simpler

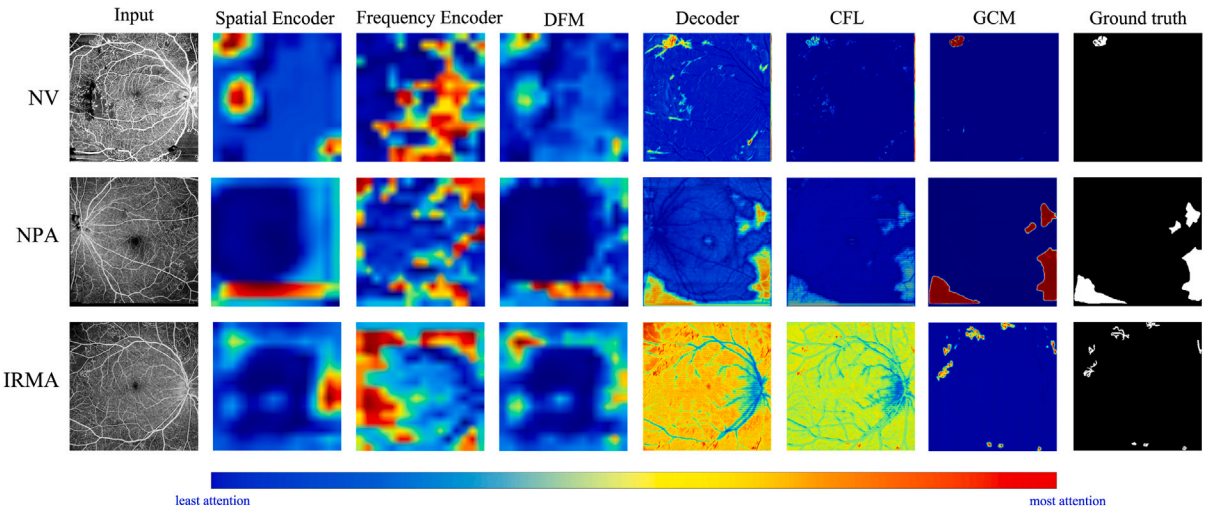


Fig. 10. Heatmap visualization on the DRAC2022 dataset. The 1st column shows the input images, the rest columns show the heatmaps of MFNet and ground truth, respectively.

Table 7
Segmentation Results of EX on the IDRiD.

Methods	DICE	IoU	SEN	PRE	SPE
Unet [33]	0.5892	0.4177	0.4902	0.7384	0.9967
ResUnet [34]	0.6174	0.4465	0.5661	0.6789	0.9969
ResUnet++ [35]	0.5914	0.4199	0.5013	0.7210	0.9964
MsTGANet [48]	0.6533	0.4851	0.5523	**0.7994**	0.9972
RetiFluidNet [42]	0.6359	0.4661	0.6337	0.6381	0.9970
PAG-TransYnet [47]	0.6310	0.4610	0.5215	0.7989	**0.9973**
MFNet	**0.6723**	**0.5063**	**0.5966**	0.7699	0.9969

Table 8
Segmentation Results of HE on the IDRiD.

Methods	DICE	IoU	SEN	PRE	SPE
Unet [33]	0.4143	0.2613	0.3871	0.4457	0.9957
ResUnet [34]	0.4351	0.2780	0.4062	0.4684	0.9959
ResUnet++ [35]	0.4404	0.2824	0.4356	0.4453	0.9952
MsTGANet [48]	0.5136	0.3455	0.4783	0.5544	0.9966
RetiFluidNet [42]	0.4964	0.3301	**0.4868**	0.5063	0.9958
PAG-TransYnet [47]	0.5061	0.3388	0.4533	0.5730	0.9970
MFNet	**0.5293**	**0.3599**	0.4772	**0.5941**	**0.9971**

architecture. On the other hand, models such as MsTGANet and RetiFluidNet demonstrate significantly higher parameter counts (30.57M and 34.26M, respectively) and computational costs (67.86 GFLOPs and 71.22 GFLOPs). The increase in complexity is expected to improve performance, but the cost of requires more computing power and memory, which will limit their deployment in resource constrained environments. Our MFNet achieves a balance between these two extremes, with a parameter count of 27.76M and 59.88 GFLOPs. Although the computational requirements of MFNet are slightly higher than those of models such as PAG TransYnet (23.61M parameters, 54.19GFLOP), MFNet provides a favorable trade-off between model complexity and potential performance. This balance ensures that MFNet can effectively handle segmentation tasks in complex scenes while maintaining an acceptable level of computational efficiency.

5.3. Generalization ability

To evaluate the generalization ability of MFNet in different types of samples, we conducted additional experiments on the Indian Diabetic Retinopathy Image Dataset (IDRiD) [54], which is a standard benchmark for diabetic retinopathy containing various lesion types. In this study, we focused on two specific lesion types: hard exudates (EX) and haemorrhages (HE). The experimental setup, including hyperparameters, data augmentation strategies and performance metrics, was kept consistent with those used for the DRAC2022 dataset.

(1) Hard Exudates Detection: As shown in Table 7, MFNet demonstrated superior performance in the detection of hard exudates on the IDRiD dataset, achieving an increase Dice coefficient and IoU of 2.3% and 1.8% compared to MsTGANet and RetiFluidNet, respectively. This improvement indicates that MFNet has a strong generalization capability in this task. Notably, although the frequency domain feature extraction module of MFNet was originally designed for OCTA images, it still effectively captures the primary features of hard exudates in color fundus images while maintaining high segmentation accuracy. However, there is still room for optimization in certain detail processing, suggesting that further adjustments may be needed to better adapt to the characteristics of color fundus images.

(2) Haemorrhage Detection: In the haemorrhage detection task (Table 8), MFNet also showed a strong detection capability. The Dice coefficient and IoU reached 52.93% and 35.99% respectively, an improvement of 2.3% and 2.1% over PAG-TransYnet. This result demonstrates the promising generalization performance of the model.

5.4. Study limitations and alternatives

(1) Computational Complexity: While MFNet improves performance through multi-level fusion and gating mechanisms, it also increases computational complexity. Table 6 shows a comparison of the parameters and GFLOPs of MFNet with other models. Although MFNet achieves a good balance between performance and efficiency, its computational cost is still relatively high, which may limit its use in resource-constrained or real-time applications. In future work, we aim to reduce the computational complexity through techniques such as model pruning and quantization, making MFNet more suitable for resource-constrained environments.

(2) Handling Noisy Images: MFNet demonstrates strong multiscale feature extraction, but it may still face stability challenges when processing highly noisy OCTA images, primarily due to high-frequency noise that interferes with subtle lesion features. To improve the robustness to noise, we plan to incorporate more advanced image preprocessing and noise reduction techniques in future research to improve the performance of the model under noisy conditions.

(3) Generalizability and Limited Data: In the experiment, we used data augmentation to increase the quantity of samples. The effectiveness of MFNet depends to some extent on the amount of training data. In situations where data is limited, models may struggle to learn effectively, resulting in poor performance for certain types of lesions. To address this issue, we attempt to improve generalization ability by generating more diverse synthetic datasets and introducing domain

adaptation techniques. In addition, we have explored unsupervised and weakly supervised learning methods to improve performance under limited or incomplete data labels.

5.5. Clinical implications

In this study, by integrating spatial and frequency domain features, the proposed MFNet significantly improves the segmentation accuracy of lesion regions and demonstrates strong clinical potential. First, MFNet shows excellent performance in detecting microvascular abnormalities associated with early diabetic retinopathy, providing clinicians with an earlier opportunity for intervention and reducing the risk of further visual impairment. Second, the incorporation of frequency domain feature analysis effectively reduces the noise and artifacts commonly observed in OCTA imaging, greatly improving both the accuracy and stability of segmentation results and providing a more reliable basis for clinical diagnosis. Furthermore, although this study focuses primarily on diabetic retinopathy, experimental results indicate that MFNet is equally applicable to other ophthalmic diseases involving hard exudates and haemorrhages. The method can now be extended to color fundus images, demonstrating robust adaptability. The high accuracy of MFNet in retinal lesion segmentation suggests its future integration into automated retinal screening systems. Combined with electronic health records and AI-assisted diagnostic systems, it could enable automated progression analysis, thereby reducing the diagnostic workload of clinicians and improving the accessibility of screening services.

6. Conclusion

This study demonstrates that incorporating both spatial and frequency domain features into a neural network notably improves the accuracy of retinal lesion segmentation in OCTA images. By using multi-level discrete wavelet transforms and domain fusion module, our approach effectively captures fine details and global structures of retinal lesions and is particularly effective at detecting complex lesions, such as neovascularization. The experimental results confirm that MFNet outperforms existing methods on key performance metrics such as IoU and Dice, establishing its potential as a powerful tool for medical image segmentation. Additionally, the inclusion of frequency domain information mitigates the impact of noise, thereby enhancing the accuracy of lesion boundary detection. However, the performance of the model is still influenced by the quality, size and imbalance of the dataset.

Future research will focus on several key areas. First, to improve the robustness of the model to noisy images, we aim to explore more advanced image pre-processing and noise reduction techniques to further mitigate the impact of noise on segmentation results. Second, given the limitations in size and diversity of current medical imaging datasets, we will develop synthetic data generation methods capable of generating different lesion types and their corresponding labels, thereby improving the generalizability of the model, especially for small and unbalanced datasets. Additionally, we will focus on optimizing the computational efficiency of MFNet, using techniques such as model pruning and quantization to reduce computational complexity, enhancing the suitability of the model for low-resource environments and real-time applications. Finally, we will also explore the application of MFNet to other medical imaging modalities, such as OCT and color fundus imaging, to further validate its versatility and effectiveness.

CRediT authorship contribution statement

Guangmei Jia: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Conceptualization. **Fei Ma:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Sien Li:** Formal analysis, Data curation. **Zhaohui Zhang:** Investigation, Formal analysis. **Hongjuan Liu:** Project administration, Conceptualization. **Yanfei Guo:** Conceptualization. **Jing Meng:** Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported by Natural Science Foundation of Shandong Province, China (No:ZR2020MF105), Guangdong Provincial Key Laboratory of Biomedical Optical Imaging Technology (No:2020B12 1201010), the National Natural Science Foundation of China (62175156, 61675134), Science and technology innovation project of Shanghai Science and Technology Commission (19441905800, 22S31903000). Qufu Normal University Foundation for High Level Research (116-607001).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2025.107945.

Data availability

Data will be made available on request.

References

[1] Z.L. Teo, Y.-C. Tham, M. Yu, M.L. Chee, T.H. Rim, N. Cheung, M.M. Bikbov, Y.X. Wang, Y. Tang, Y. Lu, et al., Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis, Ophthalmology 128 (11) (2021) 1580–1591, http://dx.doi.org/10.1016/j.ophtha.2021.04.027.

[2] Z. Fu, Y. Gong, C. Löfqvist, A. Hellström, L.E. Smith, Review: adiponectin in retinopathy, Biochim. Biophys. Acta (BBA)- Mol. Basis Dis. 1862 (8) (2016) 1392–1400, http://dx.doi.org/10.1016/j.bbadis.2016.05.002.

[3] R. Simo, C. Hernández, Neurodegeneration in the diabetic eye: new insights and therapeutic perspectives, Trends Endocrinol. Metab. 25 (1) (2014) 23–33, http://dx.doi.org/10.1016/j.tem.2013.09.005.

[4] S.Z. Safi, R. Qvist, S. Kumar, K. Batumalaie, I.S.B. Ismail, Molecular mechanisms of diabetic retinopathy, general preventive strategies, and novel therapeutic targets, BioMed Res. Int. 2014 (1) (2014) 801269, http://dx.doi.org/10.1155/2014/801269.

[5] H. Khalil, Diabetes microvascular complications—A clinical update, Diabetes Metab. Syndr.: Clin. Res. Rev. 11 (2017) S133–S139, http://dx.doi.org/10.1016/j.dsx.2016.12.022.

[6] T.Y. Wong, C.M.G. Cheung, M. Larsen, S. Sharma, R. Simó, Diabetic retinopathy, Nat. Rev. Dis. Prim. 2 (1) (2016) 16012, http://dx.doi.org/10.1038/nrdp.2016.12.

[7] R.F. Spaide, J.G. Fujimoto, N.K. Waheed, S.R. Sadda, G. Staurenghi, Optical coherence tomography angiography, Prog. Retin. Eye Res. 64 (2018) 1–55, http://dx.doi.org/10.1016/j.preteyeres.2017.11.003.

[8] A.C. Tan, G.S. Tan, A.K. Denniston, P.A. Keane, M. Ang, D. Milea, U. Chakravarthy, C.M.G. Cheung, An overview of the clinical applications of optical coherence tomography angiography, Eye 32 (2) (2018) 262–286, http://dx.doi.org/10.1038/eye.2017.181.

[9] Y. Jia, O. Tan, J. Tokayer, B. Potsaid, Y. Wang, J.J. Liu, M.F. Kraus, H. Subhash, J.G. Fujimoto, J. Hornegger, et al., Split-spectrum amplitude-decorrelation angiography with optical coherence tomography, Opt. Express 20 (4) (2012) 4710–4725, http://dx.doi.org/10.1364/OE.20.004710.

[10] B. Tombolini, E. Crincoli, R. Sacconi, M. Battista, F. Fantaguzzi, A. Servillo, F. Bandello, G. Querques, Optical coherence tomography angiography: A 2023 focused update on age-related macular degeneration, Ophthalmol. Ther. 13 (2) (2024) 449–467, http://dx.doi.org/10.1007/s40123-023-00870-2.

[11] A.H.K. Nissen, A.S. Vergmann, Clinical utilisation of wide-field optical coherence tomography and angiography: A narrative review, Ophthalmol. Ther. 13 (4) (2024) 903–915, http://dx.doi.org/10.1007/s40123-024-00905-2.

[12] J. Chua, B. Tan, D. Wong, G. Garhöfer, X.W. Liew, A. Popa-Cherecheanu, C.W.L. Chin, D. Milea, C.L.-H. Chen, L. Schmetterer, Optical coherence tomography angiography of the retina and choroid in systemic diseases, Prog. Retin. Eye Res. (2024) 101292, http://dx.doi.org/10.1016/j.preteyeres.2024.101292.
[13] N.K. Waheed, R.B. Rosen, Y. Jia, M.R. Munk, D. Huang, A. Fawzi, V. Chong, Q.D. Nguyen, Y. Sepah, E. Pearce, Optical coherence tomography angiography in diabetic retinopathy, Prog. Retin. Eye Res. 97 (2023) 101206, http://dx.doi.org/10.1016/j.preteyeres.2023.101206.
[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90, http://dx.doi.org/10.1145/3065386.
[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 5987–5995, http://dx.doi.org/10.1109/CVPR.2017.634.
[16] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, Q. Ye, Conformer: Local features coupling global representations for recognition and detection, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 9454–9468, http://dx.doi.org/10.1109/TPAMI.2023.3243048.
[17] J.-J. Huang, P.L. Dragotti, WINNet: Wavelet-inspired invertible network for image denoising, IEEE Trans. Image Process. 31 (2022) 4377–4392, http://dx.doi.org/10.1109/TIP.2022.3184845.
[18] Y. Cui, W. Ren, X. Cao, A. Knoll, Image restoration via frequency selection, IEEE Trans. Pattern Anal. Mach. Intell. 46 (2) (2024) 1093–1108, http://dx.doi.org/10.1109/TPAMI.2023.3330416.
[19] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, C. Miao, WaveFill: A wavelet-based generation network for image inpainting, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 14094–14103, http://dx.doi.org/10.1109/ICCV48922.2021.01385.
[20] B. Li, B. Zheng, H. Li, Y. Li, Detail-enhanced image inpainting based on discrete wavelet transforms, Signal Process. 189 (2021) 108278, http://dx.doi.org/10.1016/j.sigpro.2021.108278.
[21] V.K. Singh, E.Y. Kalafi, S. Wang, A. Benjamin, M. Asideu, V. Kumar, A.E. Samir, Prior wavelet knowledge for multi-modal medical image segmentation using a lightweight neural network with attention guided features, Expert Syst. Appl. 209 (2022) 118166, http://dx.doi.org/10.1016/j.eswa.2022.118166.
[22] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, 2020, http://dx.doi.org/10.48550/arXiv.2006.03677, arXiv preprint arXiv:2006.03677.
[23] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F.E.H. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on ImageNet, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 538–547, http://dx.doi.org/10.1109/ICCV48922.2021.00060.
[24] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open 3 (2022) 111–132, http://dx.doi.org/10.1016/j.aiopen.2022.10.001.
[25] B. Qian, H. Chen, X. Wang, Z. Guan, T. Li, Y. Jin, Y. Wu, Y. Wen, H. Che, G. Kwon, et al., DRAC 2022: A public benchmark for diabetic retinopathy analysis on ultra-wide optical coherence tomography angiography images, Patterns (2024) http://dx.doi.org/10.1016/j.patter.2024.100929.
[26] B. Dong, X. Wang, X. Qiang, F. Du, L. Gao, Q. Wu, G. Cao, C. Dai, A multi-branch convolutional neural network for screening and staging of diabetic retinopathy based on wide-field optical coherence tomography angiography, IRBM 43 (6) (2022) 614–620, http://dx.doi.org/10.1016/j.irbm.2022.04.004.
[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1–9, http://dx.doi.org/10.1109/CVPR.2015.7298594.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.
[29] Y. Yin, Z. Han, M. Jian, G.-G. Wang, L. Chen, R. Wang, AMSUnet: A neural network using atrous multi-scale convolution for medical image segmentation, Comput. Biol. Med. 162 (2023) 107120, http://dx.doi.org/10.1016/j.compbiomed.2023.107120.
[30] T.D. Ngo, B.-S. Hua, K. Nguyen, ISBNet: a 3D point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 13550–13559, http://dx.doi.org/10.1109/CVPR52729.2023.01302.
[31] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, et al., Self-supervised pretraining of visual features in the wild, 2021, http://dx.doi.org/10.48550/arXiv.2103.01988, arXiv preprint arXiv:2103.01988.
[32] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 3431–3440, http://dx.doi.org/10.1109/CVPR.2015.7298965.
[33] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: 2015 18th Medical Image Computing and Computer-Assisted Intervention, MICCAI, Springer, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.
[34] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th International Conference on Information Technology in Medicine and Education, ITME, 2018, pp. 327–331, http://dx.doi.org/10.1109/ITME.2018.00080.
[35] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T.D. Lange, P. Halvorsen, H. D. Johansen, ResUNet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia, ISM, 2019, pp. 225–2255, http://dx.doi.org/10.1109/ISM46123.2019.00049.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, 2020, http://dx.doi.org/10.48550/arXiv.2012.12877, arXiv preprint arXiv:2012.12877.
[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv preprint arXiv:2010.11929.
[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 9992–10002, http://dx.doi.org/10.1109/ICCV48922.2021.00986.
[40] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer V2: Scaling up capacity and resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 11999–12009, http://dx.doi.org/10.1109/CVPR52688.2022.01170.
[41] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E.K. Aghdam, J. Cohen-Adad, D. Merhof, Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2023, pp. 6191–6201, http://dx.doi.org/10.1109/WACV56688.2023.00614.
[42] R. Rasti, A. Biglari, M. Rezapourian, Z. Yang, S. Farsiu, RetiFluidNet: A self-adaptive and multi-attention deep convolutional network for retinal OCT fluid segmentation, IEEE Trans. Med. Imaging 42 (5) (2023) 1413–1423, http://dx.doi.org/10.1109/TMI.2022.3228285.
[43] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2020) 1856–1867, http://dx.doi.org/10.1109/TMI.2019.2959609.
[44] X. Fang, P. Yan, Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, IEEE Trans. Med. Imaging 39 (11) (2020) 3619–3629, http://dx.doi.org/10.1109/TMI.2020.3001036.
[45] J. Liu, M. Li, Q. Gao, S. Gong, Z. Tang, Y. Xie, A. Mohammadzadeh, Toward automated right ventricle segmentation via edge feature-induced self-attention multiscale feature aggregation full convolution network, IEEE Trans. Instrum. Meas. 72 (2023) 1–12, http://dx.doi.org/10.1109/TIM.2022.3206810.
[46] M.I. Razzak, M. Imran, G. Xu, Efficient brain tumor segmentation with multiscale two-pathway-group conventional neural networks, IEEE J. Biomed. Heal. Inform. 23 (5) (2019) 1911–1919, http://dx.doi.org/10.1109/JBHI.2018.2874033.
[47] F. Bougourzi, F. Dornaika, A. Taleb-Ahmed, V.T. Hoang, Rethinking attention gated with hybrid dual pyramid transformer-CNN for generalized segmentation in medical imaging, 2024, http://dx.doi.org/10.48550/arXiv.2404.18199, arXiv preprint arXiv:2404.18199.
[48] M. Wang, W. Zhu, F. Shi, J. Su, H. Chen, K. Yu, Y. Zhou, Y. Peng, Z. Chen, X. Chen, MsTGANet: Automatic drusen segmentation from retinal OCT images, IEEE Trans. Med. Imaging 41 (2) (2022) 394–406, http://dx.doi.org/10.1109/TMI.2021.3112716.
[49] X. Li, Y. Zheng, M. Zang, W. Jiao, Wavelet transform and edge loss-based three-stage segmentation model for retinal vessel, Biomed. Signal Process. Control. 86 (2023) 105355, http://dx.doi.org/10.1016/j.bspc.2023.105355.
[50] C. Tian, M. Zheng, W. Zuo, B. Zhang, Y. Zhang, D. Zhang, Multi-stage image denoising with the wavelet transform, Pattern Recognit. 134 (2023) 109050, http://dx.doi.org/10.1016/j.patcog.2022.109050.
[51] C. Liu, M. Pang, Automatic lung segmentation based on image decomposition and wavelet transform, Biomed. Signal Process. Control. 61 (2020) 102032, http://dx.doi.org/10.1016/j.bspc.2020.102032.
[52] T. Imtiaz, S.A. Fattah, S.-Y. Kung, BAWGNet: Boundary aware wavelet guided network for the nuclei segmentation in histopathology images, Comput. Biol. Med. 165 (2023) 107378, http://dx.doi.org/10.1016/j.compbiomed.2023.107378.
[53] J. Ramya, H. Vijaylakshmi, H.M. Saifuddin, Segmentation of skin lesion images using discrete wavelet transform, Biomed. Signal Process. Control. 69 (2021) 102839, http://dx.doi.org/10.1016/j.bspc.2021.102839.
[54] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, et al., Idrid: Diabetic retinopathy–segmentation and grading challenge, Med. Image Anal. 59 (2020) 101561, http://dx.doi.org/10.1016/j.media.2019.101561.