Contents lists available at ScienceDirect

Information Fusion

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

Full length article

XpertDx: Expert-level diabetic retinopathy lesion segmentation with cross-domain feature fusion

Fei Ma [a], Guangmei Jia [a,b,*], Fen Yan [c], Yuefeng Ma [a], Ronghua Cheng [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*
[c] *Ultrasound Medicine Department Qufu People's Hospital, Shandong, China*

ARTICLE INFO

Keywords:
Deep learning
OCTA
Segmentation
Retinal lesion
Contrastive learning

ABSTRACT

Diabetic retinopathy is a microvascular disease that poses a significant threat to visual health, and its automatic segmentation is crucial for early diagnosis and intervention. Optical Coherence Tomography Angiography (OCTA) is a non-invasive imaging technique capable of obtaining high-resolution structures of retinal and choroidal vasculature. However, due to the minimal blood flow changes associated with early microlesions, these subtle abnormalities are often overlooked in imaging. Furthermore, traditional segmentation methods primarily rely on information from a single perspective provided by a single network, making it challenging to effectively capture the complex characteristics of such lesions. To address these issues, we propose a novel lesion segmentation framework for the precise segmentation of retinal lesion regions in OCTA images. Specifically, we design a frequency-domain encoder with multi-level discrete wavelet transform to capture multi-scale texture features. An adaptive fusion perception module (AFPM) is then employed to facilitate deep interaction and alignment between spatial and frequency domain features. In addition, we developed a comparative monitoring module that embeds a contrastive learning mechanism at the patch level. Furthermore, we propose a consistency learning strategy with multi-path decoding and consistency correction to capture details in complex lesion regions. Experimental results on two OCTA DR datasets show that our method outperforms existing state-of-the-art methods. This consolidates that the analysis of retinal lesions may offer a new scheme for the study of various neurodegenerative diseases.

1. Introduction

Diabetic Retinopathy (DR) is a common and severe microvascular complication in diabetic patients, posing a significant global threat to visual health [1]. According to the International Diabetes Federation (IDF), the number of individuals with diabetes worldwide is expected to reach approximately 700 million by 2045 [2]. The primary pathological manifestations of DR include microvascular abnormalities, oxidative stress, and metabolic dysregulation, all of which progressively lead to irreversible retinal damage as the disease advances [3–6]. Therefore, early and accurate diagnosis and intervention are critical to slowing disease progression.

Optical Coherence Tomography Angiography (OCTA), as a non-invasive imaging technology, can image the microvasculature of the retina and the choroid, offering a valuable tool for the early diagnosis and progression monitoring of DR [7,8]. However, OCTA imaging faces several challenges, including stripe noise, instability in detecting low blood flow regions, and image quality degradation caused by eye motion and media opacity [9–11]. These issues complicate the differentiation between lesion regions and background areas, making it particularly difficult to delineate retinal lesion boundaries. Fig. 1 shows the typical noise of OCTA fundus image. One can observe that there are disconnected regions in the results of spatial-domain processing (Gaussian filtering operation) for Fig. 1(b), while many details of blood can be obtained by using Frequency transformation from Fig. 1(b).

Although spatial domain-based Convolutional Neural Networks (CNNs) have been widely applied in medical image analysis and excel in local feature extraction [12,13], their limited receptive fields hinder their ability to capture global context and handle long-range dependencies [14]. Additionally, wavelet transforms, known for their multi-scale capabilities in image denoising [15], reconstruction [16], and texture analysis [17], are less effective in capturing high-level

* Corresponding authors at: School of Computer Science, Qufu Normal University, Rizhao, 276826, China.
E-mail addresses: mafei@qfnu.edu.cn (F. Ma), gmjia@qfnu.edu.cn (G. Jia), guanlichm@163.com (F. Yan), rzmyf1976@163.com (Y. Ma), chengronghua98@163.com (R. Cheng), jingmeng@qfnu.edu.cn (J. Meng).

https://doi.org/10.1016/j.inffus.2025.103556
Received 25 March 2025; Received in revised form 5 June 2025; Accepted 17 July 2025
Available online 25 July 2025

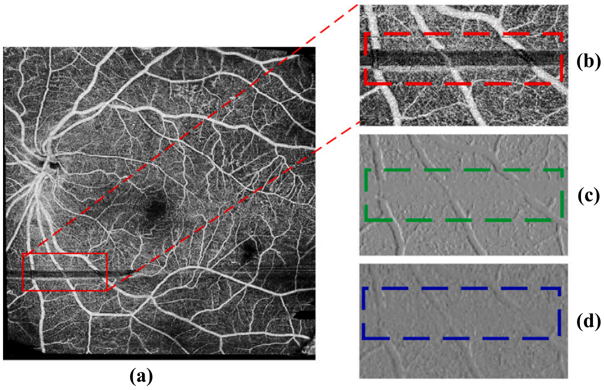

Fig. 1. The typical noises in OCTA fundus sample. (a) The original sample. (b) The details of noised region. (c) Result of Gaussian filtering for (b). (d) Result of Frequency transform for (b).

semantic features. Consequently, combining the advantages of multi-domain features while addressing the insufficient representation of global and local features in segmentation tasks is a critical challenge in OCTA lesion segmentation.

As above statements, we attempt to design a frequency-domain encoder constructed using multi-level discrete wavelet transforms to effectively extract multi-scale texture features. Simultaneously, the frequency-domain encoder optimizes input feature quality by removing stripe noise. To facilitate deep interaction and alignment between spatial and frequency-domain features, we will design an Adaptive Fusion Perception Module to achieve comprehensive feature integration. Inspired by the clinical practice of multiple expert consultations, we introduce a Consistency Learning Module, which strengthens the recognition and segmentation of lesion regions.

The main contributions of this paper are as follows:

- An adaptive fusion perception module is introduced to achieve the alignment and interaction of spatial and frequency domain features, providing a novel method for joint learning of multi-domain features.
- To share the low-level features and avoid excessive feature divergence between different paths, we design a Consistency Learning Module (CLM). Through multi-path decoding and consistency correction, the proposed module can capture details in complex lesion regions.
- To distinguish pixels between lesion regions and background areas, we introduce the Comparative Monitoring Module (CMM). By applying contrastive learning at the patch level, this module can differentiate pixels between lesion regions and background areas and offers a new approach to efficient feature extraction.
- The extensive experiments have been conducted on two datasets. Experimental results on the DRAC2022 and WF-OCTA datasets demonstrate that our method outperforms existing state-of-the-art approaches in lesion segmentation tasks.

The remainder of this paper is organized as follows: Section 2 introduces related works on convolutional neural networks, wavelet transforms, contrastive learning, and medical image segmentation. Section 3 provides a detailed explanation of the proposed method. The experimental setup and results are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related work

2.1. Medical image segmentation

Medical image segmentation is a crucial step in medical image processing [18]. Its goal is to extract regions of interest from medical images, providing accurate information for subsequent diagnosis, treatment, and surgery [19]. U-Net [20], as a typical deep learning architecture, has been widely applied in medical image segmentation due to its efficient segmentation performance. ResUnet [21] introduces residual connections to effectively address the gradient vanishing problem in deep networks. Subsequently, Jha et al. [22] enhanced the feature fusion capability by improving the encoder–decoder structure. To address the inherent challenges specific to the segmentation of Diabetic Retinopathy (DR) lesions, researchers have investigated various enhancements to network architectures and feature augmentation strategies [23–25]. Reza et al. [26] proposed RetiFluidNet, a Convolutional Neural Network (CNN) designed for multi-class retinal fluid segmentation in Optical Coherence Tomography (OCT) images. While RetiFluidNet enhances segmentation precision through an adaptive attention mechanism capable of accommodating variations in fluid morphology and scan characteristics, its complex attention modules incur significant computational time complexity. Recognizing the diverse manifestations of fundus lesions, several studies have also explored frequency-domain analysis approaches [27–29]. For instance, a lightweight network employing a frequency recalibration module [30] aims to improve the differentiation of texture and shape information characteristic of DR lesions. Furthermore, attention mechanisms inspired by the Transformer architecture are increasingly being utilized in medical image analysis [24,31]. Wang et al. [32] introduced MsTGANet, utilizing a multi-scale Transformer module (MsTNL) and a multi-semantic global channel and spatial joint attention module to improve the segmentation accuracy of drusen in OCT images. By integrating multi-scale features, this method better addresses drusen's morphological variations and noise interference, yet segmentation in complex backgrounds remains challenging.

2.2. Contrastive learning

Contrastive Learning (CL), as an important paradigm of self-supervised learning, has found widespread application in deep learning [33–35]. In particular, in image processing tasks, contrastive learning has become an effective approach to enhancing model performance, owing to its powerful feature extraction capability [36–38]. Most existing methods, such as SimCLR [39] and MoCo [40], typically perform contrastive learning at the instance level, optimizing the model's ability to represent different instances. However, this approach may lead to suboptimal representations in pixel-level prediction tasks, such as image segmentation [41]. With the evolving task requirements, some methods [42–44] have attempted to extend contrastive learning to the pixel level, aiming to further enhance model performance in segmentation tasks. However, due to the large number of pixels in the entire dataset, such methods may significantly increase the model's resource consumption. RegionContrast, proposed by Hu et al. [45], employs region-level contrastive learning to better capture deep semantic relationships in the image while maintaining computational efficiency. Additionally, another class of research [46–49] combines contrastive learning with other modules, such as attention mechanisms and multi-scale learning, in an attempt to further improve the segmentation accuracy of models in complex backgrounds. These studies offer effective pathways for addressing the shortcomings of traditional models in detail modeling.

2.3. Wavelet transform

Wavelet transform is a powerful tool for time-frequency analysis that decomposes an input signal into a set of fundamental waveforms [50]. The discrete wavelet transform (DWT) has been widely applied in signal analysis [51,52], image processing [53], and feature extraction in pattern recognition [54], among others. In recent years, wavelet transform has been increasingly incorporated into medical image processing. Singh et al. [55] designed MISegNet, a lightweight network for multi-modality medical image segmentation. It leverages DWT to efficiently extract frequency-domain features from the medical

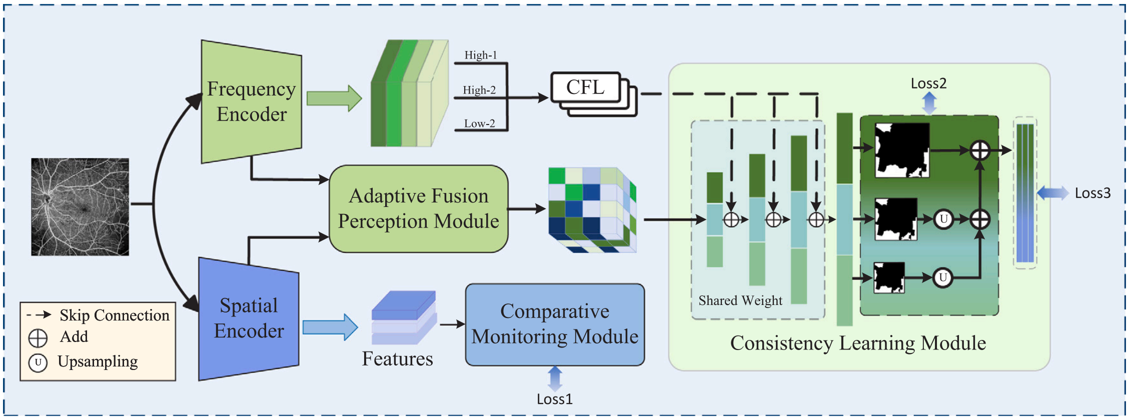

Fig. 2. Overview of XpertDx framework. The input image undergoes feature extraction via the Frequency Encoder and Spatial Encoder, which capture image features at different levels. Subsequently, the features are fused in the Adaptive Fusion Perception Module to generate a comprehensive feature representation with learned weights. In the feature alignment step, the input image is divided into multiple patches, and high-dimensional feature information for each patch is analyzed to optimize the fusion process. This is followed by a group perception correction during the decoding phase from multiple perspectives to produce the final segmentation mask.

images. Ramya et al. [56] introduced an efficient skin lesion segmentation method based on DWT. However, this approach still faces limitations when dealing with highly complex or low-contrast skin images. Pang et al. [57] proposed BLENet, a lightweight network for segmenting the left ventricle in echocardiography. Their model utilizes an adaptive wavelet fusion module in the decoder to improve feature integration specifically for cardiac ultrasound images. Similarly, to better suppress noise while preserving crucial features in ultrasound images, Khor et al. [58] utilized wavelet-based modules in a GAN for speckle reduction, exploiting the advantages of processing medical images in the wavelet domain. Furthermore, leveraging the inherent multi-scale frequency analysis capability of wavelet transforms, researchers [59,60] have also investigated their application in medical image fusion. Notwithstanding these valuable applications, the optimal integration of multi-scale frequency features derived from wavelet transforms with spatial context, particularly for precisely delineating subtle or complex features in challenging medical imaging modalities, remains less explored.

3. Methodology

As illustrated in Fig. 2, the proposed XpertDx framework is designed to achieve high-precision segmentation of lesion regions in OCTA images of patients with diabetic retinopathy. Centered around the core principle of spatial-frequency domain synergy, the architecture employs a dual-path encoder composed of a frequency encoder and a spatial encoder to extract both structural and fine-grained texture features in parallel. The Adaptive Fusion Perception Module (AFPM) is introduced to perform precise alignment and saliency recalibration of these heterogeneous features, thereby enhancing the model's capacity for detailed texture modeling. During the decoding phase, the framework incorporates a Consistency Learning Module (CLM), which simulates multi-perspective correction through group perception of multiple decoding paths, effectively strengthening the recognition and delineation of lesion areas. Concurrently, a Comparative Monitoring Module (CMM) is integrated to impose patch-level contrastive learning, optimizing the discriminative power of features through semantic differentiation, and significantly enhancing the model's sensitivity to lesion-background boundaries.

3.1. Encoder

3.1.1. Spatial encoder

The Spatial Encoder is designed to extract multi-scale spatial features from the input image. Given an input image $X \in \mathbb{R}^{H\times W\times C}$, where H (height), W (width), and C (number of channels) represent the spatial dimensions. As shown in Fig. 3, the Spatial Encoder consists of multiple convolutional layers that progressively extract features at different scales. The encoder is composed of multiple operational units, where each unit reduces the spatial resolution of the feature maps through convolutional operations while increasing the number of feature channels to capture higher-level semantic information. These convolutional layers systematically decrease the spatial resolution of the feature maps while enhancing the depth of the feature representation to capture features at various scales. Finally, the output features from the Spatial Encoder are passed to the Adaptive Fusion Perception Module for further integration and processing.

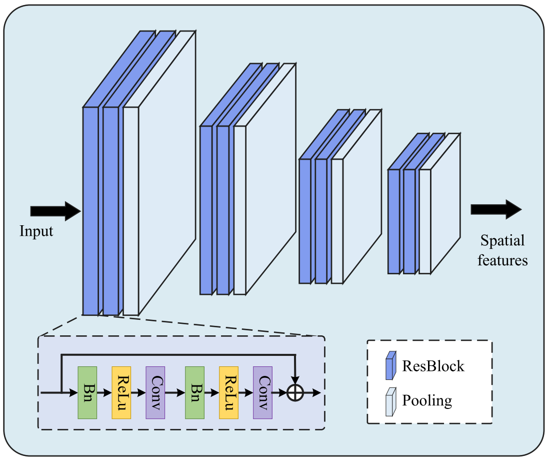

Fig. 3. The spatial encoder.

3.1.2. Frequency encoder

The primary function of the Frequency Encoder is to extract multi-scale frequency-domain features from the input image using multi-level Discrete Wavelet Transform (DWT), thereby enhancing the capability to capture texture details and fine-grained information, as illustrated in Fig. 4. This encoder consists of two levels of DWT units, each responsible for decomposing features across different frequency ranges and utilizing the decomposed features for subsequent image segmentation tasks.

Initially, the input image $X \in \mathbb{R}^{H\times W\times C}$ is fed into the first DWT unit to perform a two-dimensional wavelet transform. This transformation

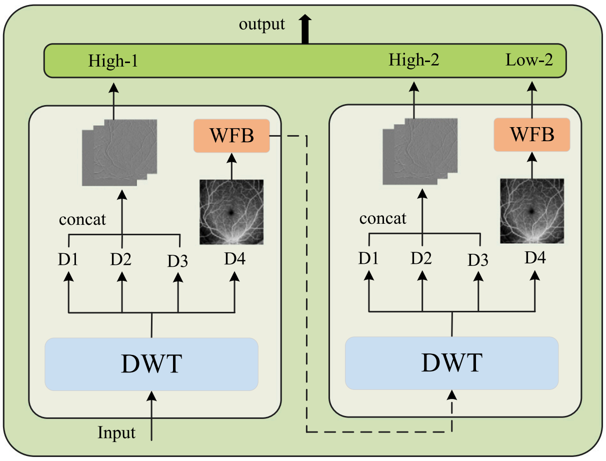

Fig. 4. The frequency encoder.

decomposes the input features into four sub-bands: one low-frequency sub-band (LL) and three high-frequency sub-bands (LH, HL, and HH), corresponding to horizontal, vertical, and diagonal directions. The high-frequency sub-bands, which capture edge and texture information, are directly output from the first DWT unit. The low-frequency sub-band undergoes a convolutional operation to extract deeper-level features and is passed to the second DWT unit as input.

The second DWT unit repeats the decomposition process, further splitting the input features into four sub-bands and extracting additional frequency-domain information. The high-frequency sub-bands at this level are directly used to provide high-frequency detail information, while the low-frequency sub-band is processed through a convolutional layer and serves as the final output of the Frequency Encoder.

The extracted features are divided into two main components for distinct purposes. On one hand, the second-level low-frequency features are combined with the Spatial Encoder's output features and passed to the Adaptive Fusion Perception Module to achieve fusion of frequency-domain and spatial-domain features. On the other hand, the high-frequency features from both the first and second levels, along with the second-level low-frequency features, are processed by the Conditional Feature Layer (CFL) and directly fed into the Consistency Learning Module. These multi-scale high-frequency and low-frequency features provide complementary information during the consistency learning phase, effectively enhancing the model's ability to capture texture and edge details.

The feature extraction process of the Frequency Encoder can be mathematically described using the Discrete Wavelet Transform. Specifically, for each pixel (x, y) in the input features, the 2D DWT performs the following transformations:

$$LL(x,y) = \sum_{m,n} h(m) \cdot h(n) \cdot I(2x+m, 2y+n), \tag{1}$$

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

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

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

where $h(\cdot)$ and $g(\cdot)$ represent the low-pass and high-pass filters, respectively. The low-frequency sub-band LL retains the overall information of the image, while the high-frequency sub-bands LH, HL, and HH capture edge features in the horizontal, vertical, and diagonal directions.

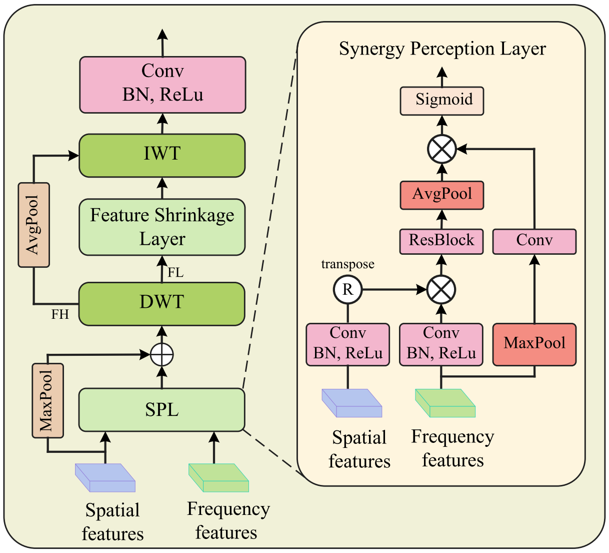

Fig. 5. The adaptive fusion perception module.

Through this multi-scale decomposition mechanism, the Frequency Encoder effectively separates the frequency-domain information of the input image into sub-bands at different resolutions, providing comprehensive feature support for subsequent modules.

3.2. Adaptive fusion perception module

In our proposed framework, the Adaptive Fusion Perception Module (AFPM) is designed to effectively integrate features from the Spatial Encoder and Frequency Encoder, bridging the representational gap between the spatial and frequency domains to achieve more comprehensive feature representations. Fig. 5 shows the structure of AFPM.

The AFPM takes the spatial features from the Spatial Encoder and the multi-scale frequency-domain features from the Frequency Encoder as inputs. First, the spatial and frequency-domain features are fed into the Synergy Perception Layer (SPL), which learns the saliency relationships between the two types of features through residual blocks and a cross-channel weighting mechanism. Specifically, the spatial and frequency-domain features are processed in the SPL through a series of convolutional operations, batch normalization, and activation functions (ReLU). This process calculates the cross-channel weight distribution and recalibrates the features accordingly, explicitly enhancing the significance of high-saliency feature channels.

Simultaneously, the preliminarily fused features undergo further decomposition via a two-dimensional Discrete Wavelet Transform (DWT), enabling the capture of fine-grained texture information. The decomposed high-frequency components (FH) and low-frequency components (FL) are processed separately: the high-frequency components pass through the Feature Shrinkage Layer, while the low-frequency components are processed using Inverse Wavelet Transform (IWT) to synthesize a feature map enriched with multi-scale information. This stage ensures effective integration of spatial and frequency-domain information while reducing computational overhead caused by redundant information.

Finally, the fused features are further processed through a convolutional module to produce the final fused feature representation. The output features encompass both spatial and frequency-domain information, retaining local textures and edge details while capturing global structural information. The AFPM makes it an efficient feature fusion module, providing critical support for subsequent consistency learning and segmentation prediction.

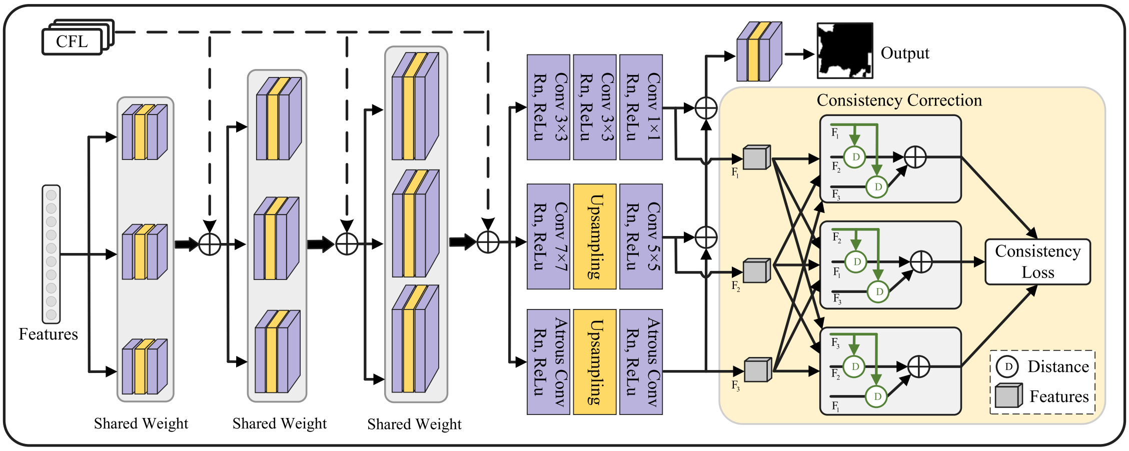

Fig. 6. Consistency learning module.

3.3. *Consistency learning module*

The Consistency Learning Module (CLM) draws inspiration from the clinical practice of multi-expert consultations and proposes a Group Perception (GP) collaborative learning strategy. This approach leverages multi-path decoding to simulate the diagnostic process of different experts during the decoding phase, correcting the model's outputs from multiple perspectives. This strategy enhances the recognition and segmentation of lesion regions. Fig. 6 shows the structure of CLM.

Specifically, CLM processes the input features through multiple decoding paths, simulating diverse decision-making processes. The first three stages of each path share the same weights, a design that reduces the number of parameters and computational cost while enforcing consistency across paths via shared weight constraints. This ensures shared representation of low-level features, avoiding excessive feature divergence between paths. The shared stages consist of a series of convolutional operations and upsampling modules that progressively restore the spatial resolution of the features. The shared weight mechanism can be mathematically expressed as:

$$F_i^k = \mathrm{Conv}\left(F_{i-1}^k; \Theta_{\mathrm{shared}}\right), \tag{5}$$

where F_i^k denotes the feature output of path k at stage i, and Θ_{shared} represents the shared weight parameters.

Simultaneously, the low-frequency features (F_{low1}) and high-frequency features ($F_{\mathrm{high1}}, F_{\mathrm{high2}}$) produced by the Frequency Encoder are incorporated into the decoding paths via skip connections, processed by the Conditional Feature Layer (CFL). This integration enhances the multi-scale feature representation during the decoding phase.

For the low-frequency and high-frequency features generated by the Frequency Encoder, the CFL processes these features as follows:

$$F_{\mathrm{high1}'} = \mathrm{ReLU}\left(\mathrm{BN}(\mathrm{Conv}(F_{\mathrm{high1}}; \Theta_{\mathrm{cfl3}}))\right), \tag{6}$$

$$F_{\mathrm{low2}'} = \mathrm{ReLU}\left(\mathrm{BN}(\mathrm{Conv}(F_{\mathrm{low2}}; \Theta_{\mathrm{cfl2}}))\right), \tag{7}$$

$$F_{\mathrm{high2}'} = \mathrm{ReLU}\left(\mathrm{BN}(\mathrm{Conv}(F_{\mathrm{high2}}; \Theta_{\mathrm{cfl4}}))\right), \tag{8}$$

where $F_{\mathrm{high1}'}, F_{\mathrm{low2}'}, F_{\mathrm{high2}'}$ are the processed features output by CFL, respectively.

The features from CFL are fed into the corresponding stages of the decoding paths via skip connections and fused with the progressively generated features in the decoding paths. After the shared stages, the three decoding paths proceed to independent branches for feature processing. Each path generates distinct feature representations through dedicated convolutional layers, aiming to introduce diverse feature perspectives and simulate the varied diagnostic viewpoints of multiple experts during clinical consultations. The outputs from the three paths are then upsampled to the same resolution to enable consistency correction.

Consistency correction is the core step of CLM. By calculating the similarity between path outputs, the model is guided to reduce discrepancies among the outputs of different paths. Specifically, let the outputs of the three paths be O_1, O_2, O_3. The consistency loss $\mathcal{L}_{\mathrm{consistency}}$ is defined as:

$$\mathcal{L}_{\mathrm{consistency}} = \frac{1}{3} \sum_{i,j \in \{1,2,3\}, i \neq j} \|O_i - O_j\|_2^2, \tag{9}$$

where $\|\cdot\|_2^2$ denotes the Euclidean distance. By minimizing the consistency loss, the model effectively reduces the deviation among path outputs, yielding more stable segmentation results from the multi-path predictions.

The consistency-corrected features are further processed by residual convolutional blocks to extract global semantic features and generate the final segmentation result. The innovation of CLM lies in its combination of multi-path learning and Consistency Correction (CC), which simulates the multi-perspective analysis process in real clinical diagnosis. Additionally, the use of shared weights and residual mechanisms significantly enhances the model's segmentation capability and robustness in predicting lesion regions.

3.4. *Comparative monitoring module*

The core idea of the Comparative Monitoring Module (CMM) is to apply contrastive learning at the patch level to distinguish pixels between lesion regions and background areas. By dividing the input image into fixed-sized patches, CMM encodes each patch into a high-dimensional feature representation. Through a carefully designed contrastive loss function, CMM optimizes the similarity of features within the same category in feature space while maximizing the separation between features of different categories, thereby refining feature representation. The CMM is illustrated in Fig. 7 and consists of the following key steps.

First, the input image $X \in \mathbb{R}^{B \times C \times H \times W}$ is divided into patches with size of $P \times P$ pixels, resulting in a total of $N = \frac{H}{P} \times \frac{W}{P}$ patches. Here, B represents the batch size, C is the number of channels, and H and W are the height and width of the image, respectively. A label tensor $Y \in \mathbb{R}^{B \times N}$ is used to denote the class of each patch, where $y_{i,j} \in \{0,1\}$. Label 1 indicates that the patch contains a lesion region, while label 0 represents a background region.

Next, each patch is encoded into a high-dimensional feature vector $F_i \in \mathbb{R}^D$, where D is the feature dimension. The similarity between

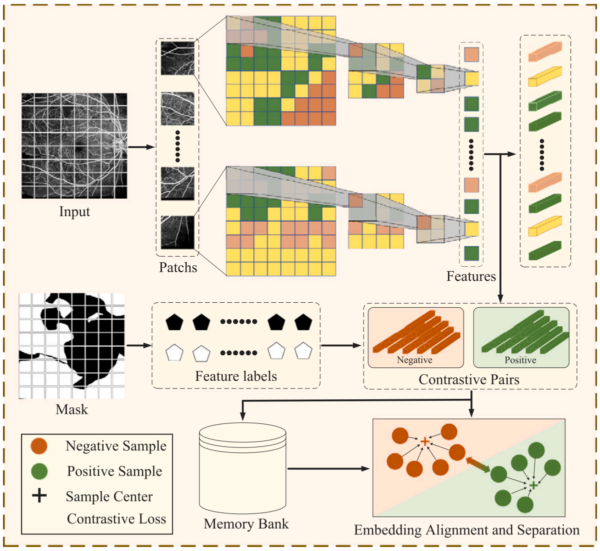

Fig. 7. The Comparative Monitoring Module.

feature vectors in the feature space is calculated using cosine similarity, defined as:

$$\mathrm{sim}(F_i, F_j) = \frac{F_i \cdot F_j}{\|F_i\|_2 \|F_j\|_2}, \tag{10}$$

where $F_i \cdot F_j$ denotes the dot product of two vectors, and $\|\cdot\|_2$ represents the L2 norm. The pairwise cosine similarities of all patches form a similarity matrix $S \in \mathbb{R}^{B\times N\times N}$.

To optimize feature representation, CMM introduces a label-based contrastive loss function. A label similarity matrix $L \in \mathbb{R}^{B\times N\times N}$ is defined for each element as:

$$L_{i,j} = \begin{cases} 1, & \text{if } y_i = y_j, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

This matrix distinguishes patches of the same class (positive samples) from those of different classes (negative samples). Using the similarity matrix and the label similarity matrix, the positive sample loss $\mathcal{L}_{\mathrm{pos}}$ and negative sample loss $\mathcal{L}_{\mathrm{neg}}$ are defined as follows:

$$\mathcal{L}_{\mathrm{pos}} = \frac{1}{N_{\mathrm{pos}}} \sum_{i,j} \mathbb{1}_{\{L_{i,j}=1\}} \cdot (1 - \mathrm{sim}(F_i, F_j))^2, \tag{12}$$

$$\mathcal{L}_{\mathrm{neg}} = \frac{1}{N_{\mathrm{neg}}} \sum_{i,j} \mathbb{1}_{\{L_{i,j}=0\}} \cdot (\mathrm{sim}(F_i, F_j))^2, \tag{13}$$

where N_{pos} and N_{neg} denote the number of positive and negative samples, respectively, and $\mathbb{1}_{\{\cdot\}}$ is the indicator function. The total contrastive loss is then defined as:

$$\mathcal{L}_{\mathrm{contrastive}} = \mathcal{L}_{\mathrm{pos}} + \mathcal{L}_{\mathrm{neg}}. \tag{14}$$

To further enhance the effectiveness of contrastive learning, particularly when batch sizes are constrained during training, the CMM module incorporates a cross-batch Memory Bank. This Memory Bank stores the patch features from previous training batches, with dimensions $K \times D$ (storing K features, each of dimension D). When computing the contrastive loss, each patch feature in the current batch is compared not only with other features within the same batch but also with a large number of historical features stored in the Memory Bank. The Memory Bank is dynamically updated using a First-In, First-Out (FIFO) queue mechanism to maintain a feature repository of fixed size K. Through these mechanisms, CMM significantly enhances the model's ability to distinguish lesion regions from background areas, contributing to more accurate and robust segmentation performance.

The optimization procedure of our approach is summarized in Algorithm 1.

Algorithm 1 XpertDx Framework for Consistency-Aware Lesion Segmentation with Comparative Monitoring

```
Input: Input image $X \in \mathbb{R}^{B\times C\times H\times W}$, ground truth mask $Y$, model parameters $\Theta$
Output: Predicted segmentation mask $\hat{Y}$
Initialize $\Theta = \{\Theta_{\mathrm{freq}}, \Theta_{\mathrm{spatial}}, \Theta_{\mathrm{afpm}}, \Theta_{\mathrm{decoder}}, \Theta_{\mathrm{loss}}\}$
Step 1: Feature Extraction
$F_{\mathrm{freq}} \leftarrow \mathrm{FrequencyEncoder}(X; \Theta_{\mathrm{freq}})$
$F_{\mathrm{spatial}} \leftarrow \mathrm{SpatialEncoder}(X; \Theta_{\mathrm{spatial}})$
Step 2: Feature Fusion
$F_{\mathrm{fused}} \leftarrow \mathrm{AFPM}(F_{\mathrm{freq}}, F_{\mathrm{spatial}}; \Theta_{\mathrm{afpm}})$
$F_{\mathrm{fused}} = \sigma\left(W_f \cdot [F_{\mathrm{freq}}, F_{\mathrm{spatial}}]\right)$        ▷ Learned fusion weight
Step 3: Feature Alignment
Divide $X$ into $N$ patches
for each patch pair do
    Compute similarity between patch features
end for
Step 4: Decoding and Prediction
$\hat{Y} \leftarrow \mathrm{Decoder}(F_{\mathrm{fused}}; \Theta_{\mathrm{decoder}})$
Step 5: Loss Computation and Optimization
Compute total loss:
    $\mathcal{L}_{\mathrm{total}} = \alpha \cdot \mathcal{L}_{\mathrm{dice}} + \beta \cdot \mathcal{L}_{\mathrm{contrastive}} + \gamma \cdot \mathcal{L}_{\mathrm{consistency}}$
Update $\Theta$ using gradient descent:
    $\Theta \leftarrow \Theta - \eta \cdot \nabla_{\Theta} \mathcal{L}_{\mathrm{total}}$
return $\hat{Y}$
```

4. Experiments

4.1. Datasets

We evaluated the performance of our model on two public OCTA DR datasets:

(1) DRAC2022 Challenge Dataset: This dataset provides standardized ultra-widefield (swept-source) optical coherence tomography angiography (UW-OCTA) data for automated image quality assessment, lesion segmentation, and diabetic retinopathy grading. The dataset was acquired using the VG200D SS-OCTA system, which operates at a working wavelength of 1050 nm and a scan speed of 200,000 A-scans per second. In this study, we utilized 106 lesion images associated with non-perfusion areas from this dataset.

(2) WFDR Dataset: Provided by Dai et al. this widefield optical coherence tomography angiography retinal dataset (WFDR) includes data from 288 diabetic patients. The lesion images were meticulously annotated by a team of experienced ophthalmologists from Shanghai Sixth People's Hospital, affiliated with Shanghai Jiao Tong University. For our study, we selected 80 high-quality images containing non-perfusion areas for further analysis and investigation.

To ensure the scientific rigor of model training and evaluation, all datasets were strictly divided into training, validation, and test sets with ratio 8:1:1.

4.2. Implementation details

(1) Experimental Setup: The proposed model was implemented using PyTorch and trained on a workstation equipped with an NVIDIA GeForce GTX A4000 GPU (16 GB VRAM). The model parameters were updated using the AdamW optimizer, selected for its ability to decouple weight decay from the gradient update process, thereby enabling more efficient regularization. This characteristic makes AdamW particularly suitable for high-dimensional tasks such as image segmentation, which often face challenges like overfitting and complex parameter spaces.

Table 1
Performance comparison of different models on the DRAC2022 dataset. Best results are in bold.

Methods	DICE	IoU	BA	G-Mean	ACC	SPE
U-Net [20]	0.7044 ± 0.0270	0.5437 ± 0.0330	0.8404 ± 0.0170	0.8348 ± 0.0190	0.9957 ± 0.0041	0.9366 ± 0.0080
ResUnet [21]	0.7222 ± 0.0240	0.5652 ± 0.0290	0.8421 ± 0.0150	0.8354 ± 0.0175	0.9971 ± 0.0025	0.9479 ± 0.0065
ResUnet++ [22]	0.7195 ± 0.0255	0.5619 ± 0.0310	0.8412 ± 0.0160	0.8343 ± 0.0180	0.9973 ± 0.0019	0.9469 ± 0.0068
RetiFluidNet [26]	0.7316 ± 0.0195	0.5768 ± 0.0245	0.8664 ± 0.0053	0.8636 ± 0.0100	0.9980 ± 0.0012	0.9337 ± 0.0055
MsTGANet [32]	0.7455 ± 0.0130	0.5943 ± 0.0142	0.8622 ± 0.0095	0.8573 ± 0.0115	0.9982 ± 0.0015	0.9535 ± 0.0018
PAG-TransYnet [61]	0.7396 ± 0.0190	0.5868 ± 0.0225	0.8515 ± 0.0130	0.8455 ± 0.0155	0.9975 ± 0.0018	0.9519 ± 0.0050
FAI [62]	0.7379 ± 0.0175	0.5847 ± 0.0210	0.8486 ± 0.0135	0.8447 ± 0.0160	0.9976 ± 0.0017	0.9299 ± 0.0075
MADGNet [63]	0.7454 ± 0.0140	0.5941 ± 0.0155	0.8504 ± 0.0125	0.8461 ± 0.0145	0.9979 ± 0.0014	0.9353 ± 0.0065
XpertDx	**0.7620 ± 0.0111**	**0.6145 ± 0.0138**	**0.8735 ± 0.0052**	**0.8697 ± 0.0085**	**0.9989 ± 0.0010**	**0.9536 ± 0.0026**

The initial learning rate was set to 0.0001 and was progressively adjusted using a cosine annealing schedule to ensure smooth and effective convergence during training. The batch size was set to 4, balancing memory constraints and training efficiency. Furthermore, we applied data augmentation techniques to enhance the model's generalization ability and employed an early stopping strategy to monitor the validation loss, preventing overfitting.

(2) Data Augmentation: Multiple data augmentation techniques were applied to each lesion sample, including flipping, rotation, cropping, scaling, brightness and contrast adjustments, and affine transformations. These augmentations effectively improved the model's robustness and mitigated the risk of overfitting. Before feeding the images into the model, all images were standardized by scaling pixel values from $[0, 255]$ to $[0, 1]$, ensuring more stable training performance.

4.3. Evaluation metrics

We utilized six evaluation metrics to assess the performance of all methods, including Dice coefficient (DICE), Intersection over Union (IoU), Balanced Accuracy (BA), G-Mean, Accuracy (ACC), and Specificity (SPE). These metrics are defined as follows:

$$\text{DICE} = \frac{2 \times \text{TP}}{\text{FP} + 2 \times \text{TP} + \text{FN}}, \tag{15}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{16}$$

$$\text{BA} = \frac{1}{2} \cdot \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right), \tag{17}$$

$$\text{G-Mean} = \sqrt{\frac{\text{TP} \cdot \text{TN}}{(\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP})}}, \tag{18}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{19}$$

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{20}$$

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 non-lesion pixels incorrectly predicted as lesion pixels, and FN (False Negatives) refers to lesion pixels incorrectly predicted as non-lesion pixels.

4.4. Experimental results

4.4.1. Quantitative analysis

In this section, we evaluate the performance of the proposed model through quantitative analysis on two different datasets. We employ DICE, IoU, BA, G-Mean, ACC, and SPE as evaluation metrics and compare the results with existing methods.

The results in Table 1 demonstrate the superiority of our XpertDx compared to existing models on the DRAC2022 dataset. XpertDx achieves the highest DICE score of 0.7620 and an IoU of 0.6145, outperforming other strong baseline models including MsTGANet, RetiFluidNet, and MADGNet. These results indicate that our method effectively captures lesion boundaries and improves segmentation accuracy. The Balanced Accuracy (BA) and G-Mean scores further validate the robustness of our approach. With a BA of 0.8735 and a G-Mean of 0.8697, our model surpasses all competing methods, showing enhanced capability to balance the detection of lesion and non-lesion regions. Moreover, the Specificity (SPE) score of 0.9535 reflects the model's ability to accurately identify non-lesion areas, minimizing false positives. The consistent improvements across multiple metrics highlight the effectiveness of integrating frequency-domain and spatial-domain features in our framework. The proposed multi-path decoding strategy and consistency learning mechanisms enable the model to perform well even on challenging non-perfusion region segmentation tasks.

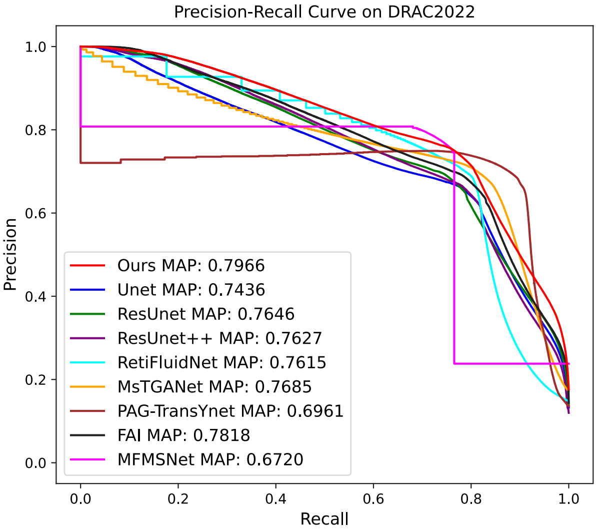

Fig. 8. Precision-Recall curves of different methods on DRAC2022 dataset.

Table 2 provides the performance comparison of various models on the second dataset, showing consistent improvement achieved by the proposed method. With a DICE score of 0.7715 and an IoU of 0.6590, our method achieves leading results, surpassing other considered methods including RetiFluidNet and MADGNet, indicating its ability to effectively segment non-perfusion areas. The BA score of 0.8713 and G-Mean of 0.8670 confirm the model's balanced performance in detecting lesion and non-lesion regions. Compared to other methods such as MsTGANet, PAG-TransYnet, and MADGNet, the proposed method demonstrates superior performance in capturing fine-grained details while maintaining a high specificity (SPE) of 0.9580.

For a more comprehensive assessment of segmentation accuracy, Fig. 8 presents the Precision-Recall (PR) curves of the evaluated models. Our model achieved the highest Mean Average Precision (MAP) score, underscoring its superior performance.

To statistically assess the performance improvement of the XpertDx model, paired t-tests were conducted comparing XpertDx with each baseline method based on their Dice coefficient and IoU scores on both the DRAC2022 (Table 3) and WFDR (Table 4) datasets. The results indicate that on both datasets, the performance improvements of XpertDx over all compared methods, in terms of both Dice and IoU

Table 2
Performance comparison of different models on the WFDR dataset. Best results are in bold.

Methods	DICE	IoU	BA	G-Mean	ACC	SPE
U-Net [20]	0.7277 ± 0.0215	0.5720 ± 0.0270	0.8209 ± 0.0140	0.8101 ± 0.0160	0.9914 ± 0.0028	0.9535 ± 0.0045
ResUnet [21]	0.7124 ± 0.0240	0.5533 ± 0.0295	0.8152 ± 0.0165	0.8043 ± 0.0185	0.9920 ± 0.0026	0.9455 ± 0.0055
ResUnet++ [22]	0.6923 ± 0.0270	0.5295 ± 0.0325	0.7958 ± 0.0200	0.7802 ± 0.0220	0.9926 ± 0.0023	0.9526 ± 0.0050
RetiFluidNet [26]	0.7537 ± 0.0155	0.6048 ± 0.0185	0.8446 ± 0.0095	0.8381 ± 0.0090	0.9940 ± 0.0015	0.9487 ± 0.0038
MsTGANet [32]	0.7540 ± 0.0140	0.6051 ± 0.0170	0.8432 ± 0.0092	0.8369 ± 0.0100	**0.9956 ± 0.0012**	0.9506 ± 0.0035
PAG-TransYnet [61]	0.7246 ± 0.0220	0.5681 ± 0.0280	0.8168 ± 0.0155	0.8048 ± 0.0170	0.9935 ± 0.0020	0.9559 ± 0.0037
FAI [62]	0.7468 ± 0.0170	0.5959 ± 0.0200	0.8457 ± 0.0105	0.8402 ± 0.0110	0.9928 ± 0.0019	0.9423 ± 0.0042
MADGNet [63]	0.7624 ± 0.0135	0.6161 ± 0.0160	0.8568 ± 0.0080	0.8523 ± 0.0085	0.9929 ± 0.0018	0.9442 ± 0.0039
XpertDx	**0.7755 ± 0.0110**	**0.6590 ± 0.0142**	**0.8713 ± 0.0060**	**0.8670 ± 0.0070**	0.9930 ± 0.0017	**0.9580 ± 0.0022**

Table 3
P-values from paired t-tests comparing XpertDx with other methods on the DRAC2022 dataset based on Dice and IoU scores.

Metric	DRAC2022							
	U-Net	ResUnet	ResUnet++	RetiFluidNet	MsTGANet	PAG-TransYnet	FAI	MADGNet
Dice	0.0012	0.0021	0.0025	0.0088	0.0135	0.0287	0.0095	0.0142
IoU	0.0015	0.0028	0.0031	0.0105	0.0162	0.0311	0.0118	0.0175

Table 4
P-values from paired t-tests comparing XpertDx with other methods on the WFDR dataset based on Dice and IoU scores.

Metric	WFDR							
	U-Net	ResUnet	ResUnet++	RetiFluidNet	MsTGANet	PAG-TransYnet	FAI	MADGNet
Dice	0.0028	0.0015	0.0009	0.0102	0.0351	0.0046	0.0075	0.0415
IoU	0.0035	0.0019	0.0011	0.0127	0.0388	0.0059	0.0018	0.0082

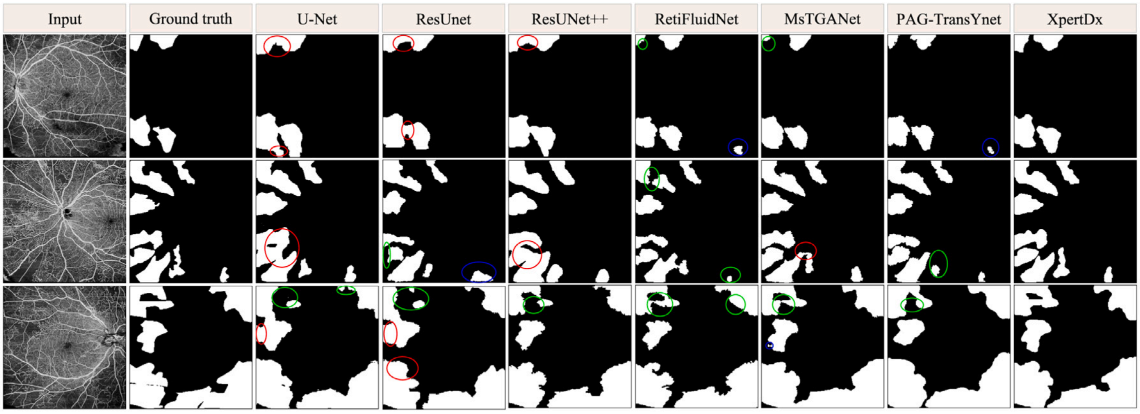

Fig. 9. Comparison of lesion segmentation results by different methods on the DRAC2022 dataset.

metrics, are statistically significant (all p-values < 0.05). This robustly substantiates the superiority of our proposed XpertDx model for the task of diabetic retinopathy lesion segmentation.

These results highlight the generalizability of our method across different datasets. The integration of frequency-domain features via multi-level wavelet transforms, along with the adaptive fusion module, contributes to the model's ability to handle diverse and complex lesion structures. This demonstrates that our framework can consistently achieve state-of-the-art performance on challenging medical image segmentation tasks.

4.4.2. Qualitative analysis

To qualitatively evaluate the effectiveness of different models on OCTA images, we compared our proposed XpertDx with several state-of-the-art models, including U-Net, ResUNet, ResUNet++, RetiFluidNet, MsTGANet, and PAG-TransYnet. The segmentation results of different models are visualized in Fig. 9. The first column in Fig. 9 represents the input retinal images, while the second column displays the corresponding ground truth segmentation. Columns three to nine show the segmentation outputs of various methods, where red circles indicate over-segmentation, green circles denote under-segmentation, and blue circles highlight mis-segmentation.

As shown in Fig. 9, it is evident that classical CNN-based architectures such as U-Net, ResUNet, and ResUNet++ exhibit a higher tendency for over-segmentation. This phenomenon may be attributed to the simple skip connections, which introduce unnecessary erroneous information into the model. Benefiting from specialized optimizations for medical imaging, RetiFluidNet, MsTGANet, and PAG-TransYnet demonstrate notable improvements over traditional CNN architectures. However, it is worth noting that these models still exhibit certain limitations when segmenting complex lesion regions, struggling to accurately capture the shape and location of subtle pathological features.

In contrast, our proposed XpertDx demonstrates a significant advantage in accurately segmenting non-perfusion areas (NPA). As shown in Fig. 9, XpertDx effectively mitigates over-segmentation and achieves a more precise delineation of the location and boundaries of NPAs, outperforming existing models in handling intricate pathological structures.

Table 5
Effects of Consistency Learning Module (CLM). Best results are in bold. GP: Group Perception, CC: Consistency Correction.

Dataset	GP	CC	DICE	IoU	BA	G-Mean	ACC	SPE
DRAC2022	×	×	0.7420 ± 0.0180	0.5890 ± 0.0220	0.8620 ± 0.0110	0.8570 ± 0.0130	0.9978 ± 0.0025	0.9480 ± 0.0040
	✓	×	0.7520 ± 0.0150	0.6010 ± 0.0180	0.8680 ± 0.0075	0.8630 ± 0.0105	0.9983 ± 0.0018	0.9500 ± 0.0032
	✓	✓	**0.7620 ± 0.0111**	**0.6145 ± 0.0138**	**0.8735 ± 0.0052**	**0.8697 ± 0.0085**	**0.9989 ± 0.0010**	**0.9536 ± 0.0026**
WFDR	×	×	0.7450 ± 0.0190	0.5930 ± 0.0235	0.8550 ± 0.0125	0.8500 ± 0.0145	0.9840 ± 0.0030	0.9520 ± 0.0042
	✓	×	0.7510 ± 0.0165	0.6020 ± 0.0190	0.8520 ± 0.0130	0.8470 ± 0.0150	0.9948 ± 0.0020	0.9510 ± 0.0030
	✓	✓	**0.7755 ± 0.0110**	**0.6590 ± 0.0142**	**0.8713 ± 0.0060**	**0.8670 ± 0.0070**	**0.9930 ± 0.0017**	**0.9580 ± 0.0022**

Table 6
Effects of Comparative Monitoring Module (CMM). Best results are in bold.

Dataset	CMM	DICE	IoU	BA	G-Mean	ACC	SPE
DRAC2022	×	0.7480 ± 0.0165	0.5960 ± 0.0200	0.8710 ± 0.0075	0.8660 ± 0.0100	0.9984 ± 0.0016	0.9510 ± 0.0035
	✓	**0.7620 ± 0.0111**	**0.6145 ± 0.0138**	**0.8735 ± 0.0052**	**0.8697 ± 0.0085**	**0.9989 ± 0.0010**	**0.9536 ± 0.0026**
WFDR	×	0.7510 ± 0.0170	0.6020 ± 0.0210	0.8520 ± 0.0090	0.8470 ± 0.0105	**0.9948 ± 0.0024**	0.9510 ± 0.0038
	✓	**0.7755 ± 0.0110**	**0.6590 ± 0.0142**	**0.8713 ± 0.0060**	**0.8670 ± 0.0070**	0.9930 ± 0.0017	**0.9580 ± 0.0022**

4.5. Ablation study

In this section, we perform an ablation study of key modules in the XpertDx framework to evaluate the different impacts on segmentation performance.

4.5.1. Effects of consistency learning module

To investigate the impact of the Consistency Learning Module on the image segmentation performance, we first replaced the Group Perception-based multi-path decoding strategy with the traditional single-path i.e., decoding approach, and removing the consistency correction step from the Comparative Monitoring Module (CMM). This ablation operation resulted in the degradation of the model's decoding process to a simple decoder, where features were progressively restored to their original size using only simple upsampling and skip connections along a single path. Experimental results indicated that, compared to the original framework, the removal of the Group Perception strategy and consistency correction significantly impaired the model's segmentation performance. For instance, as shown in Table 5, on the DRAC2022 dataset, the Dice score dropped by 2%, IoU decreased by 2.55%, and BA reduced by 1.15%. This demonstrates that, at this point, the model's learning perspective becomes limited, failing to learn multi-view representations of a single target, which in turn reduces its ability to recognize and segment complex lesion areas.

Next, we conduct the experiments with the Group Perception strategy while without consistency correction. The size of features from different decoding paths were resized to same and directly fused for output. The experimental results showed an improvement in segmentation performance with the Group Perception strategy. For example, as shown in Table 5, on the DRAC2022 dataset, the Dice score increased by 1%, IoU rose by 1.2%, and BA improved by 0.6%. However, the model's segmentation performance was still lower than the original framework. There exist a similar trend on the WFDR dataset. This could be attributed to the absence of consistency correction, which amplified the feature discrepancies between different decoding paths, thereby diminishing the model's robustness.

4.5.2. Effects of comparative monitoring module

We conducted an ablation study on the Comparative Monitoring Module (CMM) to comprehensively assess its contribution to the model's performance. The core ideas of CMM lie in optimizing features at the patch level through a contrastive learning mechanism, thereby enhancing the model's ability to distinguish between lesion and background areas. In the ablation experiment, we performed performance evaluations under two conditions: one with the inclusion of CMM and one without. As shown in Table 6, the experimental results demonstrated that the introduction of CMM significantly improved the model's performance across multiple evaluation metrics. On the DRAC2022 dataset, when CMM was not used, the model achieved a Dice coefficient of 0.7480, IoU of 0.5960, and G-Mean of 0.8660. However, with CMM incorporated, the Dice coefficient increased to 0.7620, IoU rose to 0.6145, and G-Mean improved to 0.8697. Additionally, there exist improvements on other metrics such as BA, ACC, and SPE. Similarly, on the WFDR dataset, the inclusion of CMM resulted in significant performance enhancement. These results indicate that CMM, by optimizing the similarity distribution within the feature space, enhances the model's ability to distinguish between lesion and background regions, thereby improving segmentation accuracy in complex image segmentation tasks.

4.5.3. Effects of adaptive fusion perception module

As shown in Table 7, to validate the effectiveness of the Adaptive Fusion Perception Module (AFPM), we conducted comparative experiments on the DRAC2022 and WFDR datasets, contrasting it with MAX, MIN, AVG, and SUM operations. The experimental results demonstrate that AFPM significantly outperforms traditional feature fusion methods in key metrics such as DICE and IoU, and removing this module substantially degrades the quality of feature fusion. In comparison to simple extreme value operations (MAX/MIN) or linear fusion methods (AVG/SUM), AFPM bridges the semantic gap between different representation domains effectively through the cross-channel weighting mechanism in the Synergy Perception Layer. The fused features not only preserve rich local texture information but also maintain global structural consistency through inverse wavelet transformation, thereby providing more robust feature representations for subsequent segmentation tasks.

4.6. Model complexity comparison

Table 8 shows a complexity comparison between the XperDx model and several other typical segmentation models, including U-Net, ResUnet, ResUnetPlusPlus, MsTGANet, RetiFluidNet, and PAG-TransYnet. The primary comparison metrics are the number of model parameters (in millions) and the computational cost (GFLOPs).

The U-Net and ResUnet models have relatively fewer parameters, 7.84M and 8.35M, respectively, with corresponding computational costs of 17.54 and 19.29 GFLOPs. The fewer parameters indicate that these models have lower computational complexity and are suitable for use in resource-constrained environments. However, the reduced computational complexity also suggests that their performance in more complex tasks may not be as strong as that of more sophisticated models. Among the more complex models, MsTGANet and RetiFluidNet exhibit a significant increase in both parameters and computational cost, with 30.57M parameters and 67.86 GFLOPs, and 34.26M parameters and 71.22 GFLOPs, respectively. While these models may

Table 7
Effects of Adaptive Fusion Perception Module (AFPM). Best results are in bold.

Dataset	Methods	DICE	IoU	BA	G-Mean	ACC	SPE
DRAC2022	MAX	0.7379 ± 0.0170	0.5750 ± 0.0215	0.8632 ± 0.0085	0.8580 ± 0.0110	0.9970 ± 0.0022	0.9520 ± 0.0038
	MIN	0.7210 ± 0.0200	0.5550 ± 0.0240	0.8485 ± 0.0105	0.8430 ± 0.0125	0.9962 ± 0.0028	0.9415 ± 0.0045
	AVG	0.7295 ± 0.0180	0.5650 ± 0.0225	0.8558 ± 0.0090	0.8505 ± 0.0115	0.9966 ± 0.0024	0.9468 ± 0.0040
	SUM	0.7345 ± 0.0175	0.5700 ± 0.0220	0.8595 ± 0.0088	0.8543 ± 0.0112	0.9968 ± 0.0023	0.9492 ± 0.0039
	AFPM	**0.7620 ± 0.0111**	**0.6145 ± 0.0138**	**0.8735 ± 0.0052**	**0.8697 ± 0.0085**	**0.9989 ± 0.0010**	**0.9536 ± 0.0026**
WFDR	MAX	0.7553 ± 0.0160	0.6020 ± 0.0195	0.8425 ± 0.0095	0.8380 ± 0.0105	0.9928 ± 0.0025	0.9560 ± 0.0030
	MIN	0.7410 ± 0.0190	0.5850 ± 0.0220	0.8305 ± 0.0115	0.8260 ± 0.0125	0.9919 ± 0.0031	0.9545 ± 0.0036
	AVG	0.7482 ± 0.0175	0.5935 ± 0.0205	0.8365 ± 0.0105	0.8320 ± 0.0115	0.9924 ± 0.0028	0.9553 ± 0.0032
	SUM	0.7518 ± 0.0168	0.5980 ± 0.0200	0.8395 ± 0.0100	0.8350 ± 0.0110	0.9926 ± 0.0027	0.9568 ± 0.0028
	AFPM	**0.7755 ± 0.0110**	**0.6590 ± 0.0142**	**0.8713 ± 0.0060**	**0.8670 ± 0.0070**	**0.9930 ± 0.0017**	**0.9580 ± 0.0022**

Table 8
Comparison of model parameters and GFLOPs.

Model	Parameters (M)	GFLOPs
U-Net [20]	7.84	17.54
ResUnet [21]	8.35	19.29
ResUnetPlusPlus [22]	9.23	19.98
MsTGANet [32]	30.57	67.86
RetiFluidNet [26]	34.26	71.22
PAG-TransYnet [61]	23.61	54.19
XperDx	25.72	52.97

offer better performance in complex tasks, they require more computational resources and memory, which could limit their applicability in resource-limited scenarios. PAG-TransYnet, with 23.61M parameters and 54.19 GFLOPs, strikes a moderate level of complexity, demonstrating a reasonable demand for computational resources when compared to MsTGANet and RetiFluidNet.

Our proposed XperDx model has 25.72M parameters and 52.97 GFLOPs, positioning it between these models. Compared to PAG-TransYnet, XperDx has slightly more parameters but slightly lower computational cost. XperDx achieves an optimal balance between complexity and performance, offering high performance while avoiding excessive computational overhead. As a result, XperDx not only maintains high efficiency when tackling complex segmentation tasks, but also effectively utilizes computational resources, making it adaptable to a wide range of application scenarios.

5. Conclusion

This study presents a novel image segmentation framework for retinal lesion segmentation in optical coherence tomography angiography (OCTA) images. By combining features from both the spatial and frequency domains, the model effectively addresses several key challenges faced by traditional methods in segmentation tasks. The frequency domain encoder employs multi-level discrete wavelet transform to capture multi-scale texture features and attenuate strip noise in the image. The Adaptive Fusion Perception Module facilitates sufficient interaction between spatial and frequency domain features through a feature alignment mechanism. The Comparative Monitoring Module enhances the model's ability to distinguish between lesion and background areas via patch-level contrastive learning. The Consistency Learning Module further improves the robustness and accuracy of the segmentation results through multi-path decoding and consistency correction. Experimental results demonstrate that the proposed method outperforms existing mainstream approaches on several datasets, providing an efficient and reliable solution for automated segmentation tasks.

Although the framework proposed in this study achieves breakthroughs in several aspects, several issues remain for further investigation. First, the performance of the model may vary for different types of lesions, and future work could improve its generalization ability by incorporating additional annotated data from a wider range of lesion categories. In addition, the effectiveness of contrastive learning depends on the amount and diversity of the samples; future studies could combine self-supervised learning or generative models to further improve the efficiency of feature learning. Overall, future research will focus on further optimizing the framework design while exploring its potential applications in other medical image analysis tasks.

CRediT authorship contribution statement

Fei Ma: Writing – original draft, Project administration, Conceptualization, Data curation, Writing – review & editing, Supervision, Funding acquisition. **Guangmei Jia:** Software, Visualization, Conceptualization, Writing – original draft, Writing – review & editing, Resources, Methodology, Investigation. **Fen Yan:** Formal analysis, Data curation. **Yuefeng Ma:** Formal analysis, Investigation. **Ronghua Cheng:** Project administration, Conceptualization. **Jing Meng:** Supervision, Project administration, Conceptualization, Funding acquisition.

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, China (No: 2020B121201010), the National Natural Science Foundation of China (62175156, 61675134), Science and technology innovation project of Shanghai Science and Technology Commission, China (19441905800, 22S31903000). Qufu Normal University Foundation for High Level Research, China (116-607001). Academic-Enterprise Joint Research and Development Project, China (KJ2025HX019) (Corresponding author: Fei Ma.)

Data availability

Data will be made available on request.

References

[1] 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.

[2] 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.

[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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] S.T.M. Ataky, A.L. Koerich, Multiresolution texture analysis of histopathologic images using ecological diversity measures, Expert Syst. Appl. 224 (2023) 119972, http://dx.doi.org/10.1016/j.eswa.2023.119972.

[18] M.M. Rahman, M. Munir, R. Marculescu, Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 11769–11779, http://dx.doi.org/10.1109/cvpr52733.2024.01118.

[19] M.E. Rayed, S.S. Islam, S.I. Niha, J.R. Jim, M.M. Kabir, M. Mridha, Deep learning for medical image segmentation: State-of-the-art advancements and challenges, Inform. Med. Unlocked (2024) 101504, http://dx.doi.org/10.1016/j.imu.2024.101504.

[20] 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.

[21] 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.

[22] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T.D. Lange, P. Halvorsen, H.a. 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.

[23] Y. Fu, G. Zhang, X. Lu, H. Wu, D. Zhang, RMCA U-net: hard exudates segmentation for retinal fundus images, Expert Syst. Appl. 234 (2023) 120987, http://dx.doi.org/10.1016/j.eswa.2023.120987.

[24] S. Huang, J. Li, Y. Xiao, N. Shen, T. Xu, RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation, IEEE Trans. Med. Imaging 41 (6) (2022) 1596–1607, http://dx.doi.org/10.1109/tmi.2022.3143833.

[25] H. Wang, Y. Zhou, J. Zhang, J. Lei, D. Sun, F. Xu, X. Xu, Anomaly segmentation in retinal images with poisson-blending data augmentation, Med. Image Anal. 81 (2022) 102534, http://dx.doi.org/10.1016/j.media.2022.102534.

[26] 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.

[27] H. Li, H. Li, H. Shu, J. Chen, Y. Hu, J. Liu, Self-supervision boosted retinal vessel segmentation for cross-domain data, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, IEEE, 2023, pp. 1–5, http://dx.doi.org/10.1109/ISBI53787.2023.10230561.

[28] J. Xu, W. Yang, C. Wan, J. Shen, Weakly supervised detection of central serous chorioretinopathy based on local binary patterns and discrete wavelet transform, Comput. Biol. Med. 127 (2020) 104056, http://dx.doi.org/10.1016/j.compbiomed.2020.104056.

[29] H. Li, H. Liu, H. Fu, Y. Xu, H. Shu, K. Niu, Y. Hu, J. Liu, A generic fundus image enhancement network boosted by frequency self-supervised representation learning, Med. Image Anal. 90 (2023) 102945, http://dx.doi.org/10.1016/j.media.2023.102945.

[30] Y. Fu, M. Liu, G. Zhang, J. Peng, Lightweight frequency recalibration network for diabetic retinopathy multi-lesion segmentation, Appl. Sci. 14 (16) (2024) 6941, http://dx.doi.org/10.3390/app14166941.

[31] Y. Fu, J. Liu, J. Shi, TSCA-Net: Transformer based spatial-channel attention segmentation network for medical images, Comput. Biol. Med. 170 (2024) 107938, http://dx.doi.org/10.1016/j.compbiomed.2024.107938.

[32] 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.

[33] Q. Hao, R. Ren, K. Wang, S. Niu, J. Zhang, M. Wang, EC-Net: General image tampering localization network based on edge distribution guidance and contrastive learning, Knowl.-Based Syst. 293 (2024) 111656, http://dx.doi.org/10.1016/j.knosys.2024.111656.

[34] F. Liu, X. Qian, L. Jiao, X. Zhang, L. Li, Y. Cui, Contrastive learning-based dual dynamic GCN for SAR image scene classification, IEEE Trans. Neural Netw. Learn. Syst. 35 (1) (2022) 390–404, http://dx.doi.org/10.1109/TNNLS.2022.3174873.

[35] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, H. Hu, Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2021, pp. 16679–16688, http://dx.doi.org/10.1109/CVPR46437.2021.01641.

[36] R. Zhu, B. Zhao, J. Liu, Z. Sun, C.W. Chen, Improving contrastive learning by visualizing feature transformation, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 10286–10295, http://dx.doi.org/10.1109/ICCV48922.2021.01014.

[37] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, L. Van Gool, Unsupervised semantic segmentation by contrasting object mask proposals, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2021, pp. 10032–10042, http://dx.doi.org/10.1109/ICCV48922.2021.00990.

[38] H. Wu, B. Zhang, C. Chen, J. Qin, Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning, IEEE Trans. Med. Imaging (2023) http://dx.doi.org/10.1109/TMI.2023.3314430.

[39] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, 2020, http://dx.doi.org/10.48550/arXiv.2002.05709, arXiv preprint arXiv:2002.05709.

[40] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 9726–9735, http://dx.doi.org/10.1109/CVPR42600.2020.00975.

[41] Q. Quan, Q. Yao, H. Zhu, S.K. Zhou, IGU-Aug: Information-guided unsupervised augmentation and pixel-wise contrastive learning for medical image analysis, IEEE Trans. Med. Imaging 44 (1) (2024) 154–164, http://dx.doi.org/10.1109/TMI.2024.3436713.

[42] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, A.C. Murillo, Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8219–8228, http://dx.doi.org/10.1109/ICCV48922.2021.00811.

[43] X. Wang, R. Zhang, C. Shen, T. Kong, L. Li, Dense contrastive learning for self-supervised visual pre-training, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 3023–3032, http://dx.doi.org/10.1109/CVPR46437.2021.00304.

[44] J. Wei, Y. Wang, X. Gao, R. He, Z. Sun, Multi-faceted knowledge-driven graph neural network for iris segmentation, IEEE Trans. Inf. Forensics Secur. 19 (2024) 6015–6027, http://dx.doi.org/10.1109/TIFS.2024.3407508.

[45] H. Hu, J. Cui, L. Wang, Region-aware contrastive learning for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16291–16301, http://dx.doi.org/10.1109/ICCV48922.2021.01598.

[46] Y. Tan, K.-F. Yang, S.-X. Zhao, Y.-J. Li, Retinal vessel segmentation with skeletal prior and contrastive loss, IEEE Trans. Med. Imaging 41 (9) (2022) 2238–2251, http://dx.doi.org/10.1109/TMI.2022.3161681.

[47] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, Y. Wang, Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning, Med. Image Anal. 79 (2022) 102447, http://dx.doi.org/10.1016/j.media.2022.102447.

[48] B. Zeng, L. Chen, Y. Zheng, X. Chen, Adaptive multi-dimensional weighted network with category-aware contrastive learning for fine-grained hand bone segmentation, IEEE J. Biomed. Heal. Inform. (2024) http://dx.doi.org/10.1109/JBHI.2024.3391387.

[49] X. Liu, Y. Ding, Y. Zhang, J. Tang, Multi-scale local-global transformer with contrastive learning for biomarkers segmentation in retinal OCT images, Biocybern. Biomed. Eng. 44 (1) (2024) 231–246, http://dx.doi.org/10.1016/j.bbe.2024.02.001.

[50] 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.

[51] W. Xie, T. Feng, M. Zhang, J. Li, D. Ta, L. Cheng, Q. Cheng, Wavelet transform-based photoacoustic time-frequency spectral analysis for bone assessment, Photoacoustics 22 (2021) 100259, http://dx.doi.org/10.1016/j.pacs.2021.100259.

[52] A.H. Abdulwahhab, A.H. Abdulaal, A.H.T. Al-Ghrairi, A.A. Mohammed, M. Valizadeh, Detection of epileptic seizure using EEG signals analysis based on deep learning techniques, Chaos Solitons Fractals 181 (2024) 114700, http://dx.doi.org/10.1016/j.chaos.2024.114700.

[53] 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.

[54] M. Fallahian, E. Ahmadi, F. Khoshnoudian, A structural damage detection algorithm based on discrete wavelet transform and ensemble pattern recognition models, J. Civ. Struct. Heal. Monit. 12 (2) (2022) 323–338, http://dx.doi.org/10.1007/s13349-021-00546-0.

[55] 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.

[56] 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.

[57] X. Pang, F. Yao, Y. Zhang, Y. Sun, E.P.L. Lao, C. Lin, P.C.-I. Pang, W. Wang, W. Li, Z. Gao, et al., BLENet: a bio-inspired lightweight and efficient network for left ventricle segmentation in echocardiography, IEEE Trans. Circuits Syst. Video Technol. (2025) http://dx.doi.org/10.1109/tcsvt.2025.3558496.

[58] H.G. Khor, G. Ning, X. Zhang, H. Liao, Ultrasound speckle reduction using wavelet-based generative adversarial network, IEEE J. Biomed. Heal. Inform. 26 (7) (2022) 3080–3091, http://dx.doi.org/10.1109/jbhi.2022.3144628.

[59] R. Liu, Y. Liu, H. Wang, K. Hu, S. Du, A novel medical image fusion framework integrating multi-scale encoder-decoder with discrete wavelet decomposition, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 1961–1965, http://dx.doi.org/10.1109/icassp48485.2024.10446618.

[60] Z. Chao, X. Duan, S. Jia, X. Guo, H. Liu, F. Jia, Medical image fusion via discrete stationary wavelet transform and an enhanced radial basis function neural network, Appl. Soft Comput. 118 (2022) 108542, http://dx.doi.org/10.1016/j.asoc.2022.108542.

[61] 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.

[62] G. Kwon, E. Kim, S. Kim, S. Bak, M. Kim, J. Kim, Bag of tricks for developing diabetic retinopathy analysis framework to overcome data scarcity, in: MICCAI Challenge on Mitosis Domain Generalization, Springer, 2022, pp. 59–73, http://dx.doi.org/10.1007/978-3-031-33658-4_7.

[63] J.-H. Nam, N.S. Syazwany, S.J. Kim, S.-C. Lee, Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2024, pp. 11480–11491, http://dx.doi.org/10.1109/CVPR52733.2024.01091.