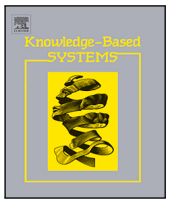
Knowledge-Based Systems 319 (2025) 113599

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Adaptive frequency-domain enhanced deep model driven by heterogeneous networks for medical image segmentation

Dong Liu [a,b,c,1], Jin Kuang [a,b,d,*,1]

[a] Hunan Engineering Research Center of Advanced Embedded Computing and Intelligent Medical Systems, Xiangnan University, Chenzhou, 423300, China
[b] School of Computer and Artificial Intelligence, Xiangnan University, Chenzhou, 423300, China
[c] Key Laboratory of Medical Imaging and Artificial Intelligence of Hunan Province, Xiangnan University, Chenzhou, 423300, China
[d] School of Geosciences, Yangtze University, Wuhan, 430100, China

ARTICLE INFO

Dataset link: https://datasets.simula.no/kvasir-seg/, https://scholar.cu.edu.eg/?q=afahmy/pages/dataset, https://challenge.isic-archive.com/data/#2017, https://acdc.creatis.insa-lyon.fr/description/databases.html, https://www.synapse.org/Synapse:syn3193805/wiki/217789, https://datasets.simula.no/kvasir-capsule-seg/, https://github.com/xbhlk/STU-Hospital.git, https://www.ub.edu/mnms/, https://github.com/promisedong/AFDSeg,

Keywords:
Frequency domain aware
Progressive feature coupling
Heterogeneous networks
Medical image segmentation
Prototype feature fusion

ABSTRACT

Accurate medical image segmentation necessitates precise localization of global structures and local boundaries due to the high variability in lesion shapes and sizes. However, existing models are limited by conventional spatiotemporal features and single-network architectures, which restrict the simultaneous captures of semantic information and boundary details, thereby challenging generalizable medical image segmentation. To overcome these limitations, we propose a heterogeneous network-driven adaptive frequency-domain enhanced deep model(AFDSeg). First, we introduce the Frequency Domain Adaptive High-Frequency Feature Selection(FAHS) module, which adaptively extracts high-frequency features to enhance contour and detail representation while integrating spatiotemporal and frequency-domain features for improved consistency. Additionally, Prototype-Guided Low-Frequency Feature Aware(PFLA) and Local High-Frequency Salient-Feature Denoising (LHSD) modules are developed, which extract discriminative low-frequency features while suppressing local noise in high-frequency components, thereby facilitating efficient multi-scale feature fusion. Furthermore, the Multi-Level Prototype Feature Refinement(MPFR) Module is introduced to align low- and high-dimensional features during decoding and enhance semantic consistency. Finally, a heterogeneous network framework capable of accommodating multiple network architecture for medical image segmentation is proposed. Our method achieves mDice scores of 93.91%, 88.64%, 91.27%, 90.74%, and 81.38% on the Kvasir-SEG, BUSI, ISIC-2017, ACDC, and Synapse datasets, respectively, and attains 92.09%, 93.50%, and 83.92% in cross-domain experiments on three unseen datasets (Kvasir Capsule-SEG, BUS42, and M&Ms). Our approach consistently outperforms state-of-the-art methods on both benchmark and cross-domain datasets. Extensive quantitative and qualitative experiments demonstrated that AFDSeg accurately segments global structures and local details while maintaining superior generalization, underscoring its clinical significance. The Code is available at https://github.com/promisedong/AFDSeg.

1. Introduction

Medical image segmentation is the process of partitioning medical images into meaningful regions, each corresponding to a specific anatomical structure or pathological feature, thereby providing accurate quantitative information and visualization support for clinical applications, such as precise diagnosis, treatment planning, surgical navigation, and organ reconstruction [1]. However, the rapid increase in medical image volumes in recent years, along with diverse image types, complex pathological features, blurred lesion boundaries, and noise, has introduced significant challenges to accurate segmentation [2,3].

In recent years, end-to-end deep learning methods have achieved remarkable performance in medical image segmentation [4–6]. These methods can be classified into three categories based on their network architecture. The first category includes CNN-based methods, which rely on convolution and pooling operations. Notable examples include DeepLabV3 [4], which employs dilated convolution, and the U-Net series [7], characterized by its U-shaped architecture; Spatial information embedding-boundary shape characteristics (SIE-BSC [8]) captures multimodal spatial information and boundary shape characteristics using a convolutional neural network. These models effectively

* Corresponding author.
E-mail addresses: liudong@xnu.edu.cn (D. Liu), gasquue@gmail.com (J. Kuang).
URLs: https://github.com/promisedong (D. Liu), https://github.com/gasking (J. Kuang).
[1] The two authors contribute equally to this work.

https://doi.org/10.1016/j.knosys.2025.113599
Received 4 February 2025; Received in revised form 6 April 2025; Accepted 18 April 2025
Available online 30 April 2025

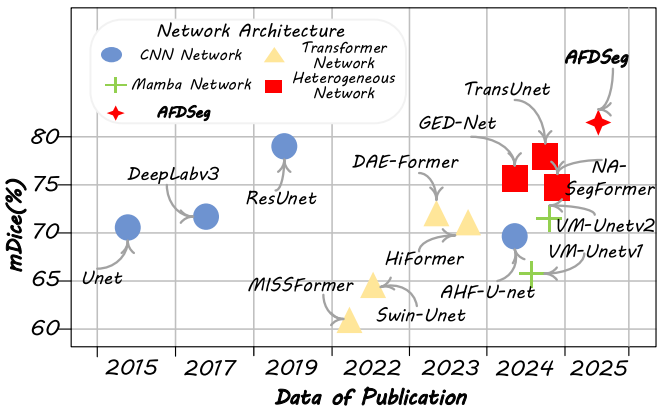


Fig. 1. Segmentation performance of various model architectures on the Synapse dataset [21]. Most previous methods rely on a single network architecture and overlook adaptive joint spatiotemporal and frequency-domain feature extraction for the progressive coupling of normal tissues and lesions (organs) in medical images. We revisit the role of the frequency domain in feature extraction and introduce AFDSeg, which outperforms previous state-of-the-art methods in segmentation performance.

extract local features but have limited global context awareness and a weaker ability to capture long-range dependencies. The second category consists of Transformer-based methods, such as SETR [9], Swin Transformer [10], and DAE-former [5], which utilize self-attention mechanisms. These models excel at capturing global dependencies and generalize effectively. However, pixel-level feature extraction and high computational resource requirements limit such methods. The third category includes the Mamba method based on the Selective State Space Model (SSM) [11], such as the VM-Unet [6] and VM-Unetv2 [12], which are computationally efficient and perform well with long sequences. However, due to its specialized scanning mode, it can exacerbate domain-specific biases, limiting its adaptability to new domains. Despite the promising segmentation performance of these methods, challenges remain in achieving precise medical image segmentation.

Recently, two emerging techniques have garnered significant attention in medical image segmentation. Developing heterogeneous deep learning frameworks that integrate the strengths of different architectures can enhance segmentation accuracy. For instance, the popular TransUnet [13] integrates Transformer with U-Net by embedding the self-attention mechanism of Transformer into U-Net's backbone. This combination leverages the advantages of both architectures, making it a powerful tool for medical image segmentation. Similarly, GED-Net [14] proposes a heterogeneous network architecture that combines a dual-branch network with a graph neural network, substantially improving the accuracy and robustness of lesion segmentation in ultrasound images. These approaches offer new insights into feature complementarity in heterogeneous networks, effectively boosting segmentation performance by utilizing the strengths of different frameworks. Contrarily, the frequency domain offers a novel analytical perspective to capture frequency characteristics that are challenging to extract using spatiotemporal domain features in deep learning frameworks [15–17]. This approach mitigates the impact of significant feature variations such as morphology, size, and color, thereby enhancing the generalization ability of deep learning networks. For instance, methods like GLFNet [18], GFUNet [19], and FreqMamba [20] extract frequency domain features within Transformer, CNN, and Mamba architectures, respectively, each demonstrating enhanced performance in related vision tasks.

However, despite advancements in recent studies, several challenges remain unaddressed. (1) Efficient feature interaction in heterogeneous networks: Although existing heterogeneous networks have improved feature fusion, the complementary enhancement effect across models remains insufficient. This limitation is particularly pronounced in the precise segmentation of complex anatomical structures, restricting their applicability to a limited range of lesions and leading to poor generalization performance. (2) Joint learning of temporal and frequency domain features: Although some studies have attempted to integrate frequency-domain and spatiotemporal-domain features within a single deep learning framework, most rely on simple feature concatenation or linear weighting strategies, exhibiting significant limitations in cross-domain feature interaction mechanisms. (3) Effective coupling of different frequency band features: Current frequency-domain methods primarily focus on extracting global high-frequency features while neglecting low-frequency components. This imbalance hinders the representation of fine-grained details, leading to blurred boundaries and topological distortions. To address these challenges, this paper plans to utilize frequency-domain features as a bridge for heterogeneous network learning and design a generalized heterogeneous network with adaptive feature enhancement across temporal and frequency domains, enabling precise segmentation of irregular lesions and organ boundary details.

This study introduces an adaptive frequency-domain enhanced deep model driven by heterogeneous networks (AFDSeg) for medical image segmentation. Experimental results (see Fig. 1) demonstrate that AFDSeg significantly outperforms existing methods, validating the effectiveness of our approach. AFDSeg adaptively learns both high-frequency and low-frequency features in the frequency domain while facilitating the integration of diverse heterogeneous networks for feature perception and interaction. First, we propose the Frequency domain Adaptive High-Frequency Feature Selection (FAHS) module, which selectively extracts high-frequency features and amplifies their most informative components to enhance feature representation. Next, we develop the Prototype-Guided Low-Frequency Feature Aware (PFLA) module to improve the extraction of low-frequency features and ensure their effective coupling with high-frequency features. To suppress the noise interference in shallow detail features, we introduce the Local High-Frequency Salient-Feature Denoising (LHSD) module to mitigate local noise in high-frequency features. We then develop the Multi-Level Prototype Feature Refinement (MPFR) module to align low-dimensional and high-dimensional semantic features. Finally, we propose a heterogeneous network framework that supports multiple deep learning architectures, embedding the aforementioned modules as plug-and-play components, providing a versatile heterogeneous network-driven frequency-domain enhanced learning network. With the proposed design, our AFDSeg model accurately captures both global boundary structures and local details. Extensive experiments on medical image segmentation across multiple modalities and disease types, including rigorous evaluations on five public datasets and three unseen cross-domain datasets, demonstrate the superior performance of the proposed method.

In summary, the main contributions of this paper are as follows:

(1) We propose a novel heterogeneous network-driven adaptive domain enhancement deep model, which enables efficient cross-model feature interaction, cross-temporal-frequency domain feature coupling, and adaptive frequency-domain feature selection within a dual-stream heterogeneous network composed of a local perception branch and a global context branch. These advancements significantly improve medical image segmentation performance.
(2) A multi-scale frequency band selection mechanism is introduced to enable the adaptive fusion of high-frequency fine-detail features and low-frequency structural components. Specifically, the FAHS module adaptively selects high-frequency domain features by progressively coupling temporal- and frequency-domain features, ensuring spatially invariant feature representation and enhancing precise lesion boundary localization. Meanwhile, the PFLA module amplifies pathological response signals in low-frequency features, significantly improving sensitivity in detecting diffuse lesions.

(3) The Hierarchical LHSD module is designed to create dedicated channels for fine-detail components and contextual features during the encoding stage, effectively reducing semantic interference from detailed features in the contextual representation. Additionally, the MPFR module ensures semantic alignment across diverse scales through a dynamic feature calibration mechanism, preserving structural integrity in feature representations.

2. Related work

2.1. Medical image segmentation

Recently, deep learning has significantly enhanced image feature extraction, through data-driven network architectures, achieving remarkable success in medical image segmentation. Based on their backbone architectures, medical image segmentation methods can be categorized into CNN-based, Transformer-based, Mamba-based, and Heterogeneous network-based approaches.

CNN-based. CNNs can automatically learn hierarchical features from image; remain robust to image noise, blurring, and contrast variations, and provide high segmentation accuracy, adaptability, and scalability [4]. Among them, the U-Net series [7] with its symmetric structure and skip connections, effectively overcomes the limitation of CNNs in capturing pixel-level contextual information, leading to widespread use in medical image segmentation and the development of numerous various variants [22–25]. For instance, Res-UNet [26] incorporates residual connections between convolutional blocks to ensure that lesions or tissues are preserved as the network deepens. AFC-Unet [24] integrates a Multi-Scale Fusion Attention Gate (MFAG) module and pyramid sampling to combine multi-scale features, reducing spatial information loss caused by traditional convolution features. SIG-Unet [25] introduces a shape-intensity attention block to refocus the decoder on shape and intensity features, minimizing texture bias and further improving the generalization of U-Net variants in medical image segmentation. AHF-Unet [27] incorporates spatial and channel-level dual attention mechanisms, along with hierarchical attention-enhanced skip connections, enhancing the model's ability to comprehensively understand images. These variant models have achieved performance improvements in different medical image fields through unique innovations, but still face challenges such as weak global information modeling and heavy reliance on large annotated datasets [28].

Transformer-based. The Transformer [29] replaces convolution with an attention mechanism as its core component. By assigning different weights to information across various spatial positions, the model enhances its ability to capture features in distinct subspaces. The Swin Transformer [10] utilizes shifted windows to independently compute self-attention within each local window, facilitating feature interaction across windows while reducing computational load and improving local feature extraction capabilities. Accordingly, SwinUnet [30] employs Swin Transformer blocks to construct an encoder, decoder, and bottleneck, enabling the capture of both local and global image features. SETR [9] designs three upsampling methods Naive upsampling, progressive upsampling, and multi-level feature aggregation within the encoder–decoder framework, enhancing segmentation performance. SDV-TUNet [31] incorporates sparse dynamic encoding and edge feature extraction into a Transformer-based backbone, thereby achieving outstanding performance in MRI brain tumor segmentation. MISSFormer [32] enhances medical image segmentation by designing an improved Transformer feature alignment module and a context bridging module, capturing more valuable dependencies and contextual information. DAE-former [5] constructs a dual-attention mechanism and a skip connection crossover module to model contextual features and effectively preserve low-level features, yielding finer segmentation results. NA-SegFormer [33] utilizes a neighborhood attention mechanism and overlapping patch fusion to improve edge segmentation accuracy within the Transformer framework, achieving significant results in colorectal polyp segmentation. In summary, Transformer excels in global context modeling, but its generated low-resolution features may lack fine-grained details, potentially leading to insufficient boundary localization accuracy.

Mamba-based. The Mamba framework is built on the Structured State Space Model (SSM) [34], which efficiently handles long sequence data through selective scanning mechanisms and hardware-aware algorithms. In visual tasks, images need to be transformed into sequential data. VMamba [34] introduces a cross-scan module (CSM) and a four-way scanning mechanism to traverse the spatial domain, achieving linear complexity without sacrificing global receptive fields, thereby enhancing visual feature learning. VM-Unet [6] combines VMamba with U-Net to propose a U-shaped architecture for medical image segmentation. It replaces the convolutional modules in U-Net with the Visual State Space (VSS) block from VMamba, capturing context, and constructs an asymmetric encoder–decoder structure. VM-Unetv2 [12] further incorporates semantic and detail attention mechanisms to enhance the fusion of features at different levels. Overall, the Mamba model overcomes the limitations of traditional models with its excellent computational efficiency and flexibility, demonstrating significant potential in medical image segmentation.

Heterogeneous network-based. Recently, heterogeneous networks that integrate two or more distinct deep learning frameworks have gained significant attention [35]. For instance, as mentioned earlier, TransUnet [13], VM-Unet [12], and NA-SegFormer [33] have explored the use of dual-framework approaches to enhance segmentation performance. Specifically, heterogeneous network-driven approaches for medical image segmentation can be categorized into two types: (1) Feature fusion-based approaches. For example, Mamba-UNet [36] enhances multi-level feature modeling and integration by incorporating the VMamba module into the encoder–decoder architecture with skip connections. This design effectively captures long-range dependencies in medical images while maintaining computational efficiency. Similarly, Zhu et al. [37] integrate Swin Transformer and CNN, where the Transformer performs global semantic modeling and the CNN focuses on local edge detection. Multi-level feature fusion is further achieved through graph convolution. (2) Knowledge distillation-based approaches: These methods employ entropy or divergence measures for adaptive feature alignment in heterogeneous networks [38]. For instance, DW-KD [39] enhances the robustness of brain tumor segmentation by introducing a regularized cross-entropy loss with controlled noise in a knowledge distillation-based heterogeneous network. Meanwhile, AWM-KD [40] improves medical image segmentation by focusing on intermediate and high-level feature mapping relationships between multiple teacher and student models. Overall, although the aforementioned approaches have advanced multi-level feature fusion and feature alignment, further improvements are required to achieve more efficient feature interactions. Moreover, current heterogeneous network methods are primarily designed for specific application scenarios, and a general heterogeneous network-driven framework for universal medical image segmentation has yet to be developed. This study proposes a heterogeneous network-driven framework for medical image segmentation, laying the foundation for seamlessly integrating CNNs, Transformer, and Mamba frameworks across different medical applications to achieve superior segmentation accuracy.

2.2. Frequency domain feature learning

Existing neural networks continue to exhibit suboptimal performance in image segmentation tasks involving lesions and surrounding tissues with high similarity or organs with unclear boundaries. This occurs because they rely solely on spatio-temporal domain feature analysis, leading to the loss of object-level features between different structures. Conversely, frequency-domain feature analysis yields higher accuracy, as distinct object features are preserved in different frequency components. Recent studies have explored the integration

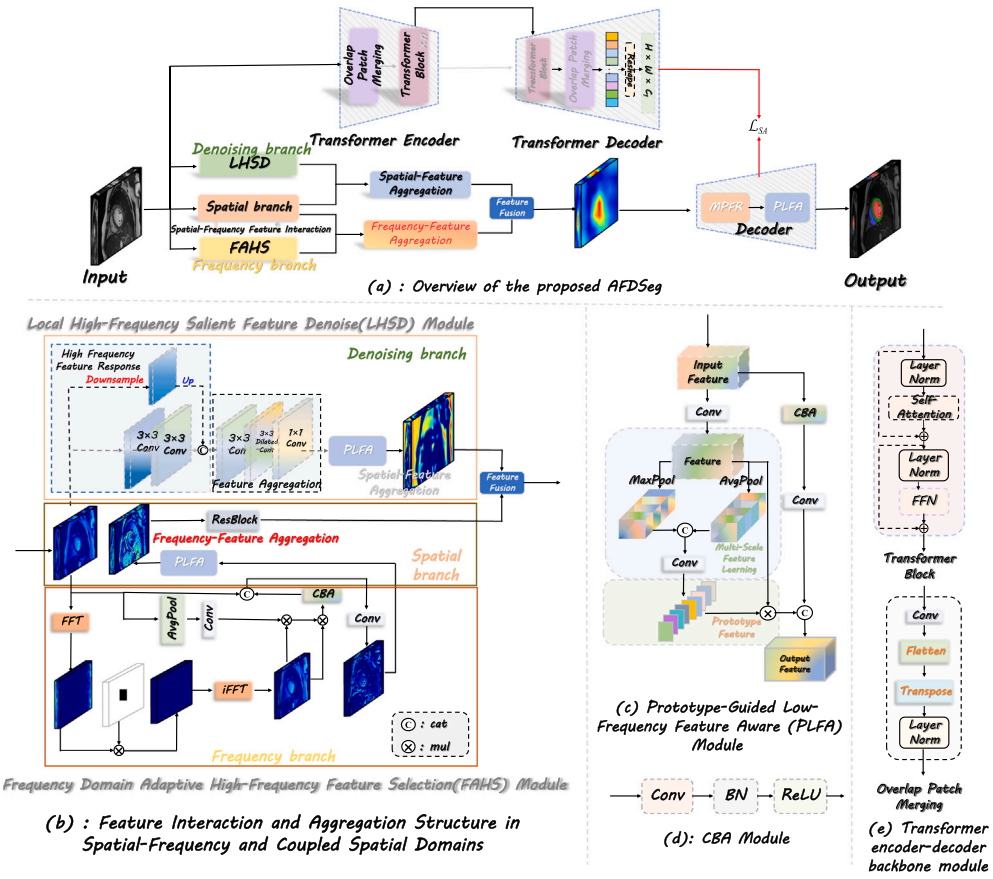


Fig. 2. (a) Overall architecture of the proposed AFDSeg; (b) The LHSD and FAHS modules correspond to the denoising branch and the frequency domain branch, respectively, while the PLFA module transforms the frequency domain into the spatial domain, representing the spatial domain branch. This process involves feature interaction and feature aggregation; (c–d) Prototype-Guided Low-Frequency Feature Aware (PLFA) module and CBA module; (e) Building Blocks of the Transformer Network.

of frequency domain feature learning into deep learning models. For example, FFC [17] leverages Fourier spectral theory to enable non-local receptive field feature extraction in deep learning models. GFNet [16] employs Fourier transforms and learnable global frequency-domain filters to balance computational complexity and classification accuracy. GLFNet [18] integrates the Transformer architecture with a global-local filter module to efficiently extract frequency-domain features, thereby accelerating feature extraction computations. Huang et al. [41] proposed the frequency domain attention (FDAM) workflow, which introduces a small number of parameters in CNNs to capture intra-class frequency relationships and suppress interference noise. Although existing studies offer solutions for frequency domain feature extraction, they still face limitations in medical image segmentation tasks under complex conditions. Therefore, GFUNet [19] introduced a global frequency domain architecture for medical image segmentation, combining Fourier transforms with the U-net structure to achieve powerful feature extraction while reducing computational complexity. FDFU-Net [42] designed a multi-scale frequency domain filter that combines frequency domain and spatio-temporal domain features to extract global and local features, effectively utilizing feature information from all channels. PSTNet [43] proposed a frequency characteristic attention module that integrates frequency cues from low-level features into feature representation, efficiently merging global and local features. This allows the model to accurately distinguish polyp tissues from surrounding normal tissues in RGB images with low contrast and blurred polyp boundaries.

These studies demonstrate that incorporating frequency domain features in image segmentation tasks can effectively enhance the feature extraction capabilities of networks, improve model generalization, and reduce interference between lesions and normal tissues in medical images. However, current research primarily focuses on integrating frequency domain features within a single disease type and a single network architecture. Further exploration is required to develop general-purpose medical image segmentation models and improve feature interactions across different network architectures. Accordingly, this study introduces an adaptive coupling mechanism for spatio-temporal and frequency domain features driven by heterogeneous networks, to enhance the ability of the model to capture both local and global features for medical image segmentation. Our approach addresses feature misalignment at the object level, enabling general-purpose medical image segmentation.

3. Methodology

3.1. Overview of the AFDSeg

Fig. 2 illustrates our proposed AFDSeg, a hybrid encoder–decoder architecture driven by heterogeneous CNN-Transformer components, where modular design enables flexible integration of diverse feature extractors. The CNN component in the feature encoding stage consists of three branches: (1) the denoising branch, (2) the spatial-domain branch, and (3) the frequency-domain branch. Through feature interaction and aggregation from multiple perspectives during encoding, it effectively addresses the challenges posed by highly irregular lesions and the poor segmentation accuracy of diffuse tissues in medical images.

Specifically, the denoising branch is composed of the Local High-Frequency Salient-Feature Denoising (LHSD) module, which aims to remove low-frequency noise from local high-frequency features. The spatial-domain branch consists of the Prototype-Guided Low-Frequency

Feature Aware (PLFA) module, which highlights the low-frequency components of high-frequency features by leveraging prototype features obtained through multi-level pooling operations. The frequency-domain branch is built upon the Frequency Domain Adaptive High-Frequency Feature Selection (FAHS) module, which utilizes a Fourier transform to convert spatial-domain features into frequency-domain features, enabling adaptive extraction of high-frequency textures and detailed features in the frequency domain.

The feature encoding layer is designed based on ResNet50 and integrated with the aforementioned modules to form the feature encoder. Additionally, in the decoding stage, we introduce the Multi-Level Prototype Feature Refinement (MPFR) module, which enables efficient perception and alignment of multi-scale feature components.

The Transformer part includes an encoder composed of Overlap Patch Merging and Transformer Blocks, while the decoder consists of Linear layers and convolutions.

It is particularly worth noting that the framework we propose is a universal heterogeneous network framework. Specifically, the core components of the framework in Fig. 2(a) are plug-and-play and can be freely combined with other networks such as Transformer and Mamba. We will conduct a detailed evaluation of their combinations in Section 4.6.

3.2. Prototype-guided low-frequency feature aware module

During the feature extraction process, existing models typically use convolutional operations for downsampling because convolution operations exhibit good spatial invariance, promoting effective feature associations between key points and neighboring features. However, neither skip nor residual connections in convolutional operations can capture global semantic information (low-frequency features). This study introduces representative global features using prototype features and develops a PLFA module to enable the model to focus on perceiving low-frequency feature information related to the segmentation target while retaining high-frequency features. Thus, the study resolves the challenge that convolution cannot effectively utilize low-frequency features. The PLFA module is shown in Fig. 2(c) and its specific operation is as follows:

(1) First, we divide the input feature $Z_1 \in \mathbb{R}^{h\times w\times c}$ into two branches. The first branch applies a 1×1 convolution operation to obtain $Z_2 \in \mathbb{R}^{h\times w\times c_1}$. Subsequently, Z_2 undergoes max pooling and average pooling respectively along the channel dimension. Through these different pooling operations, the network further enhances its perception ability of multi-scale low-frequency features. Then, the features are superimposed along the channel to obtain the feature $Z_3 \in \mathbb{R}^{1\times 1\times c_2}$. Next, a fully connected operation is utilized to construct the low-frequency prototype feature and further smooth the low-frequency feature to obtain the feature $P_1 \in \mathbb{R}^{1\times 1\times c_1}$. Finally, considering that multi-scale pooling operations may disrupt the spatial continuity and local feature expression to some extent, we perform a tensor dot product between Z_2 and P_1, resulting in a new feature $P_2 \in \mathbb{R}^{h\times w\times c_1}$ that retains continuous expression and can perceive the strength of low-frequency features.

$$Z_2 = Conv_{1\times1}(Z_1), \tag{1}$$

$$f_1 = Avgpool(Z_2), f_2 = Avgpool(Z_2), Z_3 = cat((f_1, f_2), axis = -1), \tag{2}$$

$$P_1 = Sigmoid(Transpose(Linear(Z_3))), P2 = Z_2 \otimes P_1. \tag{3}$$

(2) In the second branch, we perform local feature extraction on the input feature $Z_1 \in \mathbb{R}^{h\times w\times c}$ using the CBA (Conv + BN + ReLU) module and a 3×3 convolution to obtain $F \in \mathbb{R}^{h\times w\times c_1}$.

$$F = Conv_{3\times3}(CBA(Z_1)). \tag{4}$$

(3) Finally, we concatenate the features of P_2 and F to obtain $F_1 \in \mathbb{R}^{h\times w\times c}$, enabling effective feature aggregation of different modalities.

$$F_1 = cat((P_2, F), axis = -1). \tag{5}$$

Herein, $Sigmoid$ is an activation function mapping inputs to [0, 1], $Transpose$ refers to feature tensor transposition, $Linear$ denotes a fully connected layer, $Conv$ is the convolution operation, $Avgpool$ is the average pooling operation, $\otimes$ represents element-wise multiplication, and cat indicates feature concatenation along the specified axis.

Through the above operations, the perception of strongly correlated low-frequency features is further enhanced through prototype guidance, thereby improving the generalization ability of feature extraction in complex medical image segmentation scenarios.

3.3. Local high-frequency salient feature denoise module

Further denoising of easily interfering low-frequency features from shallow detailed features is a crucial issue that requires consideration in medical image segmentation. Existing methods primarily expanded feature perception by altering feature dimensions; however, high-frequency details are often lost during this process. To further remove noise interference from high-frequency features, we design the Local High-Frequency Salient-Feature Denoising (LHSD) Module to denoise low-frequency features interfering with high-frequency features through multi-stage convolution operations, as illustrated in Fig. 2(b).

First, to extract salient low-frequency features from the high-freq uency components of local features, we perform down-sampling on the input $F_1 \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c}$ to obtain $F_2 \in \mathbb{R}^{\frac{h}{s}\times\frac{w}{s}\times c}$. Then, we conduct two 3×3 convolution operations on F_1 without changing the feature resolution but only altering the feature channels to obtain $P \in \mathbb{R}^{h\times w\times c_1}$. Afterward, P and the upsampled F_2 are concatenated along the channel dimension to obtain $P_1 \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c_1}$. This simple sampling operation allows the model to further enhance local features at minimal cost while enabling effective interaction between local and global features. Subsequently, P_1 undergoes a 3×3 convolution, a 3×3 dilated convolution with a dilation rate of 2, and a 1×1 convolution in sequence, aiming to suppress feature noise and produce the refine feature representation, denoted as $feat \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c}$. Finally, the PLFA module is used to denoise the features via multi-scale prototype guidance. Eqs. (6)–(7) can be used to describe the above process.

$$\begin{aligned} F_2 &= Downsamole_s(F_1), \\ P &= Conv^2_{3\times3}(F_1), \\ P_1 &= Conv_{3\times3}(cat((Up(F_2), P), axis = -1)), \end{aligned} \tag{6}$$

$$feat = PLFA(Conv_{1\times1}(Conv_{(3\times3),d=2}(P_1))). \tag{7}$$

where $Downsample_j$, Up and $Conv^i_{(k\times k),d=f}$ represent downsampling, upsampling and convolution operations, respectively, j denotes the downsampling factor, i indicates the number of convolution operations, k refers to the convolution kernel size, and d represents the dilation rate.

3.4. Frequency domain adaptive high-frequency feature selection module

Compared to convolution and pooling operations, the frequency domain accurately reflects the distribution of low- and high-frequency features. The key step in medical image segmentation is to efficiently extract the most discriminative high-frequency features.

To fully leverage the progressive coupling of temporal and frequency domain features for feature extraction, we propose the FAHS module(see Fig. 2(b)), which effectively combines the spatial invariance of convolution and pooling in the spatiotemporal-domain to capture feature correlations. Additionally, it adaptively selects high-frequency features in the frequency domain to further enhance the effective components of the high-frequency features, thereby improving the ability to capture high-frequency details.

The FAHS module first applies the Discrete Fourier Transform (DFT) to map the input features $f_1 \in \mathbb{R}^{H\times W\times C}$ into the frequency domain:

$$F(c,k,l) = \frac{1}{H\times W}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f_1(h,w,c)e^{-2\pi j(kh+lw)} . \tag{8}$$

where $F \in \mathbb{R}^{H\times W\times C}$ represents the complex domain output from the DFT. H and W denote the height and width of the feature, respectively, while h, w and c correspond to the feature f_1's position in the feature map. The height and width in the frequency domain are defined by $|k|$ and $|l|$, where k is drawn from the set $\{0, \frac{1}{H}, \ldots, \frac{H-1}{H}\}$, and l is drawn from the set $\{0, \frac{1}{W}, \ldots, \frac{W-1}{W}\}$.

Further, the frequency domain feature F and the high-pass mask feature $mask \in \mathbb{R}^{H\times W\times C}$ are subjected to a tensor dot product operation to obtain the sharpened high-frequency feature $F \in \mathbb{R}^{H\times W\times C}$. Then, the inverse Fast Fourier Transform (iFFT) is applied to convert the frequency domain features back to the spatio-temporal domain. The above operations are expressed in Eqs. (9)–(11).

$$mask = Ones(H,W,C). \tag{9}$$

$$\begin{aligned}&\text{mask}\left[\left[\frac{H}{2}\right]-\text{filter window size}:\left[\frac{H}{2}\right]+\text{filter window size},\right.\\&\left.\left[\frac{W}{2}\right]-\text{filter window size}:\left[\frac{W}{2}\right]+\text{filter window size}\right]=0.\end{aligned} \tag{10}$$

$$F_1^{(h,w,c)} = iFFT(F^{(h,w,c)} \otimes mask^{(h,w,c)}). \tag{11}$$

Subsequently, through pooling and convolution operations, the feature space invariance is efficiently utilized to establish feature correlations in the time domain. A global denoising operation is applied to the input high-frequency feature f_1 to obtain the low-frequency, high-perception signal $f_2 \in \mathbb{R}^{1\times 1\times C}$ from the high-frequency features in the spatiotemporal-domain, as shown in Eq. (12).

$$f_2 = Conv_{1\times 1}(AvgPool(f_1)). \tag{12}$$

Herein, f_2, F_1 contain high-response low-frequency features and high-frequency salient feature information, respectively. To enable effective feature interaction across different modalities, an adaptive approach is adopted for the network to mine feature information. Specifically, we introduce learnable tensors $|\varsigma| = 0.995$, $|\alpha| = 1 - |\varsigma|$. In the feature mining process, $F_2 \in \mathbb{R}^{H\times W\times C}$ serves as the guiding feature, and after adaptive feature interaction, we obtain feature $P \in \mathbb{R}^{H\times W\times C}$, which possesses both global high-response and local high-frequency feature representation capabilities. Considering that with increasing network depth, some representative fine-grained high-frequency features such as details and textures may be easily lost, we employ skip connections to concatenate the input features f_1 and P along the channel dimension. Subsequently, convolution operations are applied to obtain the adaptive high-frequency features $P_1 \in \mathbb{R}^{H\times W\times C}$, as shown in Eq. (13).

$$\begin{aligned}F_2 &= \varsigma * (F_1 \otimes f_2),\\ P &= F_2 + \alpha * F_1,\\ P_1 &= Conv_{1\times 1}(cat(f_1, CBA(P)), axis = -1).\end{aligned} \tag{13}$$

Herein, $\otimes$ denotes the tensor dot product operation, and * represents multiplication by a scalar.

To provide a more detailed description of the execution process of the above modules, we present the following pseudocode (see Algorithm 1):

Algorithm 1 Frequency Domain Adaptive High-Frequency Feature Selection Module

Input: Spatial domain feature $f_1 \in \mathbb{R}^{H\times W\times C}$, Learnable parameters α, ζ, Frequency domain filtering window size $patch\_h$.
Output: Coupling high-frequency features $P_1 \in \mathbb{R}^{H\times W\times C}$.

```
Get spatial domain low-frequency feature capture: f_2 ← Conv_{1×1}(AvgPool(f_1))
Get features transformed (based on Eq. (8)) F(c,k,l) from the spatial domain via Fourier transform (FFT).
Get high-frequency filtering window mask generated: mask ← Get_highfilter_mask(h,w,c).
for h = ⌈H/2⌉ − patch_h to ⌈H/2⌉ + patch_h do
    for w = ⌈W/2⌉ − patch_h to ⌈W/2⌉ + patch_h do
        mask[h,w] ← 0
    end for
end for
Apply iFFT to convert frequency domain features to spatial domain feature: F_1^{(h,w,c)} ← iFFT(F^{(h,w,c)} ⊗ mask^{(h,w,c)})
Adaptive spatial-frequency domain feature coupling: F_2 ← ζ * (F_1 ⊗ f_2)
Progressive feature fusion: P = F_2 + α * F_1
Residual enhancement of coupled features: P_1 ← Conv_{1×1}(cat(f_1, CBA(P)), axis = −1)
return P_1
```

3.5. Multi-level prototype feature refinement module

Multi-scale features exhibit inconsistent semantic feature expressions, which present challenges for feature alignment. In the learning of frequency domain features, the interference of certain high-frequency feature components may cause category shifts among features, thereby influencing segmentation performance [15]. Further investigation reveals that the deep semantic features of low resolution are primarily dominated by low-frequency feature signals, while the detailed features of high resolution comprise high-frequency signals. Expressing the high-response features in the low-frequency domain can effectively mitigate the offset problem in the process of feature alignment.

Accordingly, we design the MPFR(see Fig. 3) to use prototype features for the correction and alignment of multi-scale features. The specific operations are as follows.

(1) A top-down feature alignment approach is employed. For the input high-resolution detail features $X \in \mathbb{R}^{h\times w\times c_1}$ and low-resolution semantic features $Y \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c_2}$, Y is first upsampled and then concatenated with X along the channel dimension. Subsequently, feature alignment and refinement are performed using a 3×3 convolution and an activation function to obtain $f \in \mathbb{R}^{h\times w\times c}$, as shown in Eq. (14).

$$f = ReLU(Conv_{3\times 3}(cat(X, Up(Y)), axis = -1)). \tag{14}$$

(2) Then, f is successively subjected to average pooling, a 3×3 convolution, and an activation function to obtain the low-frequency prototype feature $p \in \mathbb{R}^{1\times 1\times c}$, as shown in Eq. (15).

$$p = ReLU(Conv_{3\times 3}(AvgPool(f))). \tag{15}$$

(3) Further, we perform global-scale low-frequency semantic information interaction on p: First, the prototype feature p is subjected to feature probability mapping, expressing the prominent part of the low-frequency feature information to obtain $p_1 \in \mathbb{R}^{h\times w\times c}$. Then, the corrected feature f is tensor-multiplied with p_1 to explore the semantic consistency between different modal features, resulting in a more refined $p_2 \in \mathbb{R}^{h\times w\times c}$. Finally, P_2 and f are integrated and fused to obtain the decoded feature $F_{out} \in \mathbb{R}^{h\times w\times c}$, as shown in Eq. (16).

$$\begin{aligned}p_1^{(i,j,c)} &= Sigmoid((p^{(i,j,c)})),\\ &= \frac{1}{1+e^{-p_{(i,j,c)}}}\\ p_2^{(i,j,c)} &= f^{(i,j,c)} \otimes p_1^{(i,j,c)},\\ F_{out}^{(i,j,c)} &= p_2^{(i,j,c)} \oplus f^{(i,j,c)}.\end{aligned} \tag{16}$$

where i and j are the coordinate mappings of the feature points of f, and c is the feature category information, $\oplus$ represents the tensor plus operation, $\otimes$ represents element-wise multiplication.

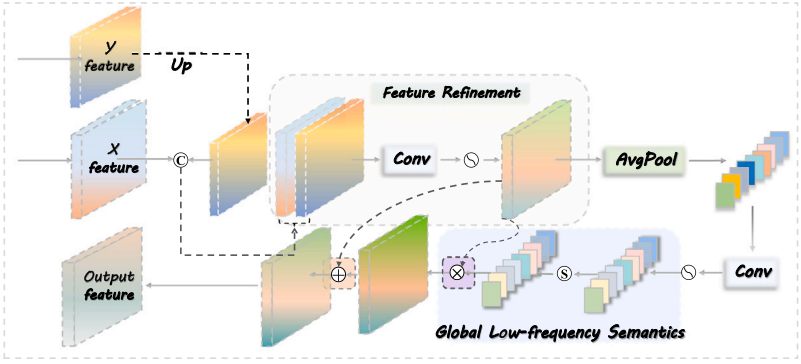


Fig. 3. Multi-Level Prototype Feature Refinement (MPFR) module.

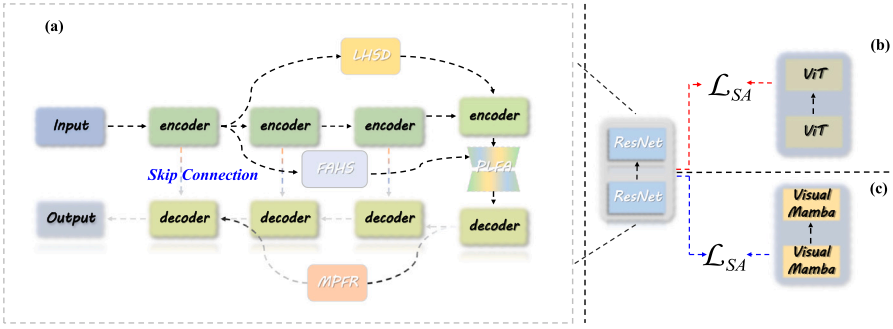


Fig. 4. General heterogeneous network architecture. (a + b) Our proposed the heterogeneous network-driven architecture; (b) CNN-Transformer network; (c) CNN-Mamba network.

3.6. Architecture of heterogeneous networks

To fully leverage the advantages of different network architectures, a universal heterogeneous network framework is proposed for medical image segmentation, as shown in Fig. 4. Specifically, we use the CNN architecture as a local feature extractor and employ the Transformer/Mamba architectures for capturing long-range contextual semantic features. In the decoding layer, we perform deep soft supervision learning between the heterogeneous networks features and compute the $\mathcal{L}_{SA}$ (as detailed in Section 3.7). Among them, multiple networks can be selected for CNN, Transformer, and Mamba architectures, including ResNet50 and ResNet101 for CNN, MiT-B1 and MiT-B5 for Transformer, while Visual-Mambav1 and Visual-Mambav2 for Mamba. We will conduct a detailed evaluation of the different combinations of these heterogeneous networks in the ablation study. Notably, our heterogeneous network structure and the proposed frequency domain feature learning module follow the plug-and-play model.

3.7. Loss function

In this section, we elaborate on the loss functions utilized by AFDSeg during the training process. For the classification loss, we employ $\mathcal{L}_{dice}$ and $\mathcal{L}_{ce}$ for pixel regression discrimination. For feature soft supervision between heterogeneous models, we use $\mathcal{L}_{SA}$ for feature alignment.

$$\mathcal{L}_{dice}(p_i, t_i) = 1 - \frac{2(\sum_{n=1}^{N}\sum_{j=0}^{J}\sum_{c=0}^{C} p_i^{(n,j,c)} t_i^{(n,j,c)}) + \varepsilon}{\sum_{n=1}^{N}\sum_{j=0}^{J}\sum_{c=0}^{C}(p_i^{(n,j,c)} + t_i^{(n,j,c)}) - 2(\sum_{n=1}^{N}\sum_{j=0}^{J}\sum_{c=0}^{C}(p_i^{(n,j,c)} t_i^{(n,j,c)})) + \varepsilon},$$
$$\mathcal{L}_{ce}(p_i, r_i) = -\sum_{c=1}^{C-1}\sum_{h,w}^{H-1,W-1} r_i^{(h,w)} log(p_i^{(h,w,c)}). \quad (17)$$

where n and j represent the value of the jth element in the nth batch of data. To prevent division by zero, ε is introduced as a smoothing factor. $p_i \in \mathbb{R}^{H\times W\times C}$ denotes the predicted class probability, $t_i \in \mathbb{R}^{H\times W\times C}$ represents the true class label, and $r_i \in \mathbb{R}^{H\times W}$ indicates the true classification One-Hot label.

For the soft supervision between features, we use $\mathcal{L}_{SA}$, as shown below.

$$\mathcal{L}_{SA}(p_i^{(h,w,c)}, q_i^{(h,w,c)}) = \frac{1}{B}\sum_{b=0}^{B-1}\sum_{h,w=0}^{H-1,W-1}(p_i^{(h,w,c)} - q_i^{(h,w,c)})^2. \quad (18)$$

where B represents the batch size, i denotes the feature layers at different scales, and h, w, and c represent the feature resolution in terms of height, width, and channels, respectively.

Finally, our loss function can be represented as $\mathcal{L}_{afd}$.

$$\mathcal{L}_{afd}(\theta) = \underset{\theta}{argmin}(exp(\mathcal{L}_{dice}(\theta) + \mathcal{L}_{ce}(\theta)) + \mathcal{L}_{SA}(\theta)). \quad (19)$$

Herein, θ denotes the network parameters optimized by the loss function. Overall, deep feature soft supervision between heterogeneous networks effectively facilitates the interaction between their features. We apply an exponential function to scale the classification loss, enabling the network to address the pixel distribution imbalance between positive and negative samples and promote weak gradients from fewer pixels to positively contribute to the classification gradients. For an ablation study, we conducted a more detailed evaluation of the combinations of loss functions.

4. Experiments and analysis

4.1. Experimental datasets

In this study, we conducted extensive experiments on five publicly available datasets involving different medical imaging modalities and disease types, i.e., the Kvasir-SEG dataset [44] for polyp segmentation in colonoscopy images, the BUSI dataset [45] for breast tumor segmentation in ultrasound images, the ISIC 2017 dataset [46] for melanoma segmentation in dermoscopic images, the ACDC [47] for cardiac segmentation in MRI, and the Synapse dataset [21] for abdominal

Table 1
Description of training samples in the datasets and explanation of relevant parameters during model training.

DatasetName	Train		Test	Data augmentation	Max epoch	Division mode
Kvasir-SEG	800		200	HFlip, VFlip, RandomRotate	1000	Random
ISIC 2017	2000		750			
BUSI	Benign	349	88		2000	
	Malignant	168	42			
	Mixed	517	130			
ACDC	1902		1076	HFkip, VFlip, RandomHue, RandomSaturation; RandomBlur, RandomBrightness	700	Official division
Synapse	1658		553		800	

multi-organ segmentation in CT images. Among them, the Kvasir-SEG, BUSI, and ISIC 2017 are binary segmentation tasks, while ACDC and Synapse are multi-class segmentation tasks. The following are detailed descriptions of each dataset.

1. **Kvasir-SEG:** The Kvasir-SEG is a pixel-level colorectal polyp endoscopic dataset, consisting of 1000 gastrointestinal polyp images along with their corresponding segmentation masks. These images have been annotated and verified by experienced gastroenterologists from Vestre Viken Health Trust in Norway.
2. **BUSI:** BUSI is a breast tumor ultrasound image dataset, consisting of 780 images from 600 female patients aged 25 to 75, collected using the LOGIQ E9 Agile ultrasound system. Each image has an average size of 500 × 500 pixels. The images are classified into three categories: normal, benign, and malignant. In this paper, we perform binary segmentation tasks for the benign, malignant, and mixed benign and malignant breast tumors in the BUSI dataset respectively.
3. **ISIC 2017:** The ISIC 2017 dataset is a large-scale dermoscopic image collection released by ISIC, aimed at melanoma detection on the skin. It consists of 2750 dermoscopic images, with 2000 images allocated for training, 150 for validation, and 600 for testing. Each image is annotated with binary masks by professional dermatologists. To assess the performance of our method, we combine the validation and test sets during the evaluation phase, resulting in a total of 750 test images. This introduces a certain level of challenge for our approach.
4. **ACDC:** The ACDC dataset originates from the MICCAI 2017 Automated Cardiac Diagnosis Challenge, which focuses on the segmentation of the left ventricle (LV), right ventricle (RV), and myocardium (MYO) in the diastolic and systolic phases of cardiac MRI. Accurate segmentation of these cardiac structures is crucial for evaluating heart function. Specifically, the ACDC dataset includes data from 150 patients, with a total of 2978 image slices.
5. **Synapse:** The Synapse dataset comes from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge and consists of 30 abdominal CT scans, comprising a total of 3779 axial contrast-enhanced clinical abdominal CT images. Following TransUnet [13], the dataset is split into 18 scans for training and 12 scans for testing. We resize the resolution of each slice image to 224 × 224. The segmentation involves 8 abdominal organs, i.e., Aorta, Stomach, Left Kidney, Right Kidney, Gallbladder, Pancreas, Liver, and Spleen.

4.2. Implementation details

The hardware environment for this study consists of a computer configured with an Intel Xeon Gold 6326 CPU and NVIDIA A100 GPU. The software environment is developed on an Ubuntu 20.04.4 LTS system, featuring Python 3.9.13 and the PyTorch 2.5.1+cu124 deep learning library. The AdamW optimizer is used with a cosine annealing learning rate schedule. The minimum learning rate is set to 1e−6, the maximum learning rate to 2.5e−4, and the weight decay rate is 1e−3. During network training, all images are resized to 224 × 224. The number of training and testing samples for each dataset, along with the data augmentation methods and relevant parameters, are listed in Table 1.

4.3. Evaluation metrics

For experimental evaluation, we use dice coefficient (Dice), Intersection over union (IoU), sensitivity (SEN, i.e, Recall), specificity (SPN), overall pixel Accuracy (ACC), and mean class pixel Accuracy (mCPA) as the primary evaluation metrics. They are defined as:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{20}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{21}$$

$$Sen(Recall) = \frac{TP}{TP + FN} \tag{22}$$

$$SPE = \frac{TN}{TN + FP} \tag{23}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{24}$$

$$CPA = \frac{TP}{TP + FP} \tag{25}$$

where true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

4.4. Comparison with state-of-the art methods

For a comprehensive comparison, fourteen state-of-the-art methods are selected for comparison, including CNN-based (i.e., Unet [7], DeepLabv3 [4], ResUnet [26], AHF-Unet [27]), Transformer-based (i.e., SETR [9], MISSFormer [32], Swin-Unet [30], DAE-Former [5], HiFormer [48]), Mamba-based (i.e.,VM-Unetv1 [6], VM-Unetv2 [12]) and heterogeneous networks based methods (i.e., NA-SegFormer [33], TransUnet [13], GED-Net [14]). A majority of these methods have been published within the last three years. Especially, over half of these methods were published within the last two years and all achieved the state-of-the-art in image segmentation upon publication. To ensure fairness in the experiment, we reproduced all comparison methods based on the source code provided in the original paper, and the training parameters for all networks were consistent with those in Section 4.2.

4.4.1. Quantitative results

(1) Results of the Kvasir-SEG Dataset:

The quantitative comparison results of the polyp segmentation are presented in Table 2. Notably: (1) Our AFDSeg attained the optimal results in seven of the eight metrics and surpasses all comparison methods in comprehensive metrics such as mCPA, ACC, mIoU, and mDice. (2) Models based on pure convolutional neural networks (e.g., ResUnet, DeepLabv3) achieved competitive performance in polyp segmentation. Particularly, the recent AHF-Unet method achieved the second-best results for SPE and mCPA. This, designing effective convolutional blocks and multi-scale feature learning structures can significantly enhance

Table 2
Quantitative comparison with state-of-the-art methods on the Kvasir-SEG dataset. The optimal results are marked as black bold while the suboptimal results are underlined '–'.

Method	Venue	*IoU*		*Spe*	*Sen*	*Dice*		*mCPA*	*ACC*	*mIoU*	*mDice*
		Background	Polyp	Background	Polyp	Background	Polyp				
Unet	MICCAI 2015	94.88	74.81	96.48	90.18	97.37	85.59	89.87	95.56	84.85	91.48
DeepLabv3	CVPR 2017	95.91	79.24	96.85	**94.17**	97.91	88.42	91.17	96.46	87.58	93.17
ResUnet(ResNet50)	ISM 2019	95.94	79.82	97.22	92.42	97.93	88.78	92.03	96.50	87.88	93.36
SETR	CVPR 2021	88.34	44.06	91.22	74.43	93.81	61.17	74.23	89.32	66.20	77.49
MISSFormer	TMI 2022	94.33	72.69	96.36	87.76	97.08	84.18	89.35	95.08	83.51	90.63
Swin-Unet	ECCV 2022	91.33	58.89	94.04	81.18	95.47	74.13	82.57	92.29	75.11	84.80
DAE-Former	MICCAI 2023	94.97	75.49	96.71	89.60	97.42	86.03	90.44	95.65	85.23	91.73
HiFormer	WACV 2023	96.19	80.99	97.36	93.16	98.06	89.50	92.45	96.73	88.59	93.78
TransUnet	Medical Image Analysis 2024	96.17	80.71	97.21	93.73	98.05	89.32	92.11	96.70	88.44	93.69
NA-SegFormer	Scientific Reports 2024	96.06	80.31	97.23	93.00	97.99	89.08	92.11	96.60	88.18	93.54
AHF-U-net	Information Fusion 2024	95.73	79.28	97.45	90.24	97.82	88.44	92.45	96.33	87.50	93.13
GED-Net	ESWA 2024	95.24	77.03	97.12	89.18	97.56	87.02	91.49	95.89	86.13	92.29
VM-Unetv1	Arxiv 2024	94.64	74.34	96.75	87.68	97.24	85.28	90.38	95.36	84.49	91.26
VM-Unetv2	ISBRA 2024	94.56	74.06	96.74	87.36	97.21	85.10	90.32	95.29	84.31	91.16
Ours	–	**96.23**	**81.39**	**97.59**	92.22	**98.08**	**89.74**	**92.99**	**96.76**	**88.81**	**93.91**

Table 3
Quantitative comparison with state-of-the-art methods on BUSI datasets. The optimal result are marked as black bold while the suboptimal results are underlined '–'.

Method	Venue	*mCPA*			*ACC*			*mIoU*			*mDice*		
		Benign	Malignant	Mixed	Benign	Malignant	Mixed	Benign	Malignant	Mixed	Benign	Malignant	Mixed
Unet	MICCAI 2015	79.40	83.64	84.53	96.87	93.26	95.73	74.93	77.05	77.73	83.87	86.18	86.34
DeepLabv3	CVPR 2017	87.56	82.62	85.90	97.16	91.77	95.94	79.21	73.73	78.90	87.29	83.77	87.22
ResUnet(ResNet50)	ISM 2019	88.48	88.93	88.14	97.61	91.67	95.99	81.79	75.87	79.80	89.16	85.50	87.90
SETR	CVPR 2021	–	83.07	–	–	91.47	–	–	73.39	–	–	83.54	–
MISSFormer	TMI 2022	72.53	84.33	81.87	96.05	92.29	95.55	69.39	75.40	76.15	77.97	85.03	85.11
Swin-Unet	ECCV 2022	82.07	86.36	87.18	97.20	92.72	96.31	77.55	76.98	80.59	84.47	86.21	88.44
DAE-Former	MICCAI 2023	78.97	80.84	82.93	95.58	89.10	95.12	70.09	68.68	75.30	79.75	79.39	84.48
HiFormer	WACV 2023	87.71	88.11	86.08	97.66	93.40	95.97	81.83	78.96	79.05	89.18	87.60	87.33
TransUnet	Medical Image Analysis 2024	88.36	**89.30**	88.21	97.48	92.50	96.17	81.03	77.48	80.37	88.62	86.62	88.29
NA-SegFormer	Scientific Reports 2024	83.54	86.03	86.75	97.29	92.94	95.98	78.53	77.29	79.31	86.74	86.41	87.53
AHF-U-net	Information Fusion 2024	81.16	84.96	86.03	96.65	92.48	96.23	74.80	75.99	79.91	83.79	85.47	87.95
GED-Net	ESWA 2024	82.84	86.30	88.18	97.17	92.54	95.86	77.71	76.61	79.42	86.11	85.95	87.62
VM-Unetv1	Arxiv 2024	79.84	82.54	80.39	96.28	92.77	95.32	72.81	75.60	74.82	82.12	85.12	84.05
VM-Unetv2	ISBRA 2024	79.52	85.94	82.01	96.86	92.04	95.98	74.97	75.51	77.57	83.90	85.16	86.18
Ours	–	**91.23**	87.47	**88.39**	**97.75**	**93.76**	**96.28**	**83.27**	**79.51**	**80.86**	**90.20**	**87.97**	**88.64**

segmentation performance. (3) Among Transformer-based methods, HiFormer demonstrated superior performance, achieving sub-optimal results for five metrics. Thus, the efficacy of the Transformer is validated in terms of global modeling capabilities. In addition, TransUnet, which integrates Transformer into Unet, has also attracted considerable attention. It significantly improved the baseline and achieved the second-best experimental results in both SEN and ACC metrics. Therefore, designing effective heterogeneous networks can further enhance model performance. (4) Mamba-based methods, such as VM-Unet and VM-Unetv2, exhibit average performance in polyp segmentation tasks, attributed to their limited attention to the overall structure of the polyps. Overall, our proposed approach advanced polyp segmentation to the state-of-the-art by effectively learning frequency-domain features within a heterogeneous network.

(2) Results of the BUSI Dataset:

The quantitative comparison results of the BUSI dataset for breast tumor segmentation are presented in Table 3. It can be observed that: (1) In breast tumor segmentation tasks, whether benign, malignant, or mixed benign-malignant, our AFDSeg achieves the best performance across the majority of metrics. Notably, in the benign tumor segmentation task, we obtained the best performance for all mean metrics, surpassing the second-best method, HiFormer, by 1.44%, 1.02%, and 3.52% for mIoU, mDice, and mCPA, respectively. (2) In the case of scarce malignant breast cancer data and highly variable lesion regions, our method still achieved the best performance in 3 out of 4 representative mean metrics, with an ACC of 93.76%, mIoU of 79.51%, and mDice of 87.97%; (3) In the mixed training of benign and malignant breast cancer, our method demonstrated superior segmentation performance. Despite inconsistent lesion characteristics, our method achieved the best results for 4 average metrics. Notably, the mIoU of 80.86% outperformed that of the DAE-Former by 5.56%, highlighting the strength of our heterogeneous network in handling fine-grained edges and high-frequency features.

(3) Results of the ISIC 2017 Dataset:

A comparison of the experimental results on the ISIC 2017 dataset is presented in Table 4. Our method achieved the best performance on four key segmentation IoU (melanoma), Dice (melanoma), mIoU, and mDice with values of 75.89%, 86.29%, 84.33%, and 91.27%, respectively, while obtaining the second-best ACC (94.11%). This further demonstrated the effectiveness of our method in large-scale medical image segmentation. A noteworthy observation is that TransUNet achieves the highest ACC but ranks second in mIoU and mDice. This can be attributed to its use of strong pre-trained weights (ImageNet-21k + ImageNet2012_R50 + ViT-B_16).

(4) Results of the ACDC Dataset:

ACDC is a multi-class cardiac segmentation task (segmenting LV, MYO, and RV). Therefore, the experimental results from both the category-level and global average perspectives were analyzed, as shown in Table 5. For the IoU and Dice metrics, our AFDSeg method achieved the best results in all three categories (LV, MYO, and RV), indicating excellence in mIoU and mDice metrices. Therefore, the proposed method achieved superior performance in medical image segmentation tasks. Regarding the Recall metric, AFDSeg achieved the highest score in MYO segmentation and maintained the top four rankings in LV and RV segmentation. Although the Unet method performed well in the SEN metric, its mIoU and Dice scores were relatively low, indicating over-segmentation. Notably, in the cardiac segmentation task, the target heart tissue was typically small with indistinct boundaries, especially as the MYO was closely connected to the LV and RV, making the boundary unclear. Compared to the latest studies (including TransUnet,

Table 4
Quantitative comparison with state-of-the-art methods on the ISIC 2017 dataset. The optimal results are marked as black bold while the suboptimal results are underlined '–'.

Method	Venue	*IoU*		*Spe*	*Sen*	*Dice*		*mCPA*	*ACC*	*mIoU*	*mDice*
		Background	Melanoma	Background	Melanoma	Background	Melanoma				
Unet	MICCAI 2015	91.81	72.29	93.26	93.18	95.73	83.92	87.32	93.25	82.05	89.83
DeepLabv3	CVPR 2017	92.55	75.56	94.53	91.67	96.13	86.08	89.46	93.95	84.06	91.11
ResUnet(ResNet50)	ISM 2019	92.39	74.16	93.65	94.19	96.04	85.17	88.14	93.75	83.28	90.61
SETR	CVPR 2021	89.44	67.34	93.41	83.67	94.42	80.48	86.50	91.33	78.39	87.45
MISSFormer	TMI 2022	91.11	70.28	92.97	91.32	95.39	82.55	86.58	92.65	80.70	88.97
Swin-Unet	ECCV 2022	92.00	73.00	93.48	91.19	95.83	84.40	87.71	93.42	82.50	90.12
DAE-Former	MICCAI 2023	91.49	71.21	93.01	92.80	95.56	83.19	86.81	92.97	81.35	89.38
HiFormer	WACV 2023	91.34	69.64	92.14	**95.79**	95.47	82.10	85.44	92.77	80.49	88.79
TransUnet	Medical Image Analysis 2024	**92.83**	75.62	94.01	94.66	**96.28**	86.12	88.83	**94.14**	84.23	91.20
NA-SegFormer	Scientific Reports 2024	92.53	75.64	**94.67**	91.11	96.12	86.13	**89.64**	93.93	84.08	91.13
AHF-U-net	Information Fusion 2024	91.64	71.76	93.16	92.83	95.64	83.56	87.10	93.10	81.70	89.60
GED-Net	ESWA 2024	92.12	73.28	93.45	93.83	95.90	84.48	87.73	93.52	82.70	90.19
VM-Unetv1	Arxiv 2024	92.09	73.79	93.95	91.85	95.88	84.92	88.43	93.53	82.94	90.40
VM-Unetv2	ISBRA 2024	91.95	72.95	93.52	92.81	95.81	84.36	87.76	93.39	82.45	90.09
Ours	–	92.77	**75.89**	94.34	93.16	96.25	**86.29**	89.30	94.11	**84.33**	**91.27**

Table 5
Quantitative comparison with state-of-the-art methods on ACDC dataset. The optimal results are marked as black bold while the suboptimal results are underlined "–".

Method	Venue	*IoU*			*Recall*			*Dice*			*mIoU*	*mDice*
		RV	MYO	LV	RV	MYO	LV	RV	MYO	LV		
Unet	MICCAI 2015	78.89	69.98	86.07	**95.85**	81.48	**96.10**	88.20	82.34	92.51	78.31	87.68
DeepLabv3	CVPR 2017	75.82	60.67	82.08	92.26	71.27	91.27	86.25	75.52	90.16	72.86	83.98
ResUnet(ResNet50)	ISM 2019	83.28	72.13	87.19	89.50	78.04	91.22	90.88	83.81	93.16	80.87	89.28
MISSFormer	TMI 2022	76.55	64.30	85.58	92.74	82.20	92.05	86.72	78.27	92.23	75.48	85.74
Swin-Unet	ECCV 2022	81.80	68.59	86.24	89.59	75.95	90.18	89.99	81.37	92.61	78.88	87.99
DAE-Former	MICCAI 2023	82.28	70.35	87.89	88.25	76.33	92.34	90.28	82.59	93.55	80.17	88.81
HiFormer	WACV 2023	81.27	69.94	87.57	93.41	80.23	93.51	89.67	82.31	93.37	79.59	88.45
TransUnet	Medical Image Analysis 2024	80.65	73.53	88.53	94.11	80.11	95.92	89.29	84.74	93.92	80.90	89.32
NA-SegFormer	Scientific Reports 2024	82.54	72.06	88.39	92.42	78.52	93.11	90.44	83.76	93.84	81.00	89.35
AHF-U-net	Information Fusion 2024	82.60	74.19	88.17	91.04	83.75	93.63	90.47	85.18	93.71	81.65	89.79
GED-Net	ESWA 2024	81.99	71.16	87.52	90.83	79.74	93.98	90.10	83.15	93.35	80.22	88.87
VM-Unetv1	Arxiv 2024	79.97	67.01	86.16	91.03	79.16	90.99	88.87	80.25	92.57	77.71	87.23
VM-Unetv2	ISBRA 2024	75.48	61.22	85.05	92.83	79.30	91.74	86.03	75.94	91.92	73.92	84.63
Ours	–	**84.61**	**75.32**	**89.80**	93.35	**84.42**	94.95	**91.66**	**85.93**	**94.62**	**83.24**	**90.74**

Table 6
Quantitative comparison with state-of-the-art methods on Synapse dataset. The optimal results are marked as black bold while the suboptimal results are underlined "–".

Method	Venue	*IoU*								*Dice*								*mIoU*	*mDice*
		Aorta	Gallbladder	LV	RV	Liver	Pancreas	Spleen	Stomach	Aorta	Gallbladder	LV	RV	Liver	Pancreas	Spleen	Stomach		
Unet	MICCAI 2015	64.30	47.37	58.30	57.32	88.04	28.42	63.89	56.97	78.27	64.28	73.66	72.87	93.64	44.26	77.97	72.59	58.08	72.19
ResUnet(ResNet50)	CVPR 2017	76.52	31.73	81.10	74.62	91.10	**38.80**	**87.15**	**69.53**	86.70	48.17	89.56	85.47	95.34	55.91	**93.13**	**82.03**	68.82	79.54
DeepLabv3	CVPR 2017	56.84	37.72	71.10	70.61	87.71	30.25	70.39	57.49	72.48	54.78	83.11	82.77	93.45	46.45	82.62	73.01	60.26	73.58
MISSFormer	TMI 2022	50.61	26.84	50.58	42.29	85.94	7.80	61.43	49.18	67.21	42.32	67.18	59.44	92.44	14.47	76.11	65.94	46.83	60.64
Swin-Unet	ECCV 2022	59.11	41.73	55.43	42.62	85.16	11.37	64.94	47.21	74.30	58.89	71.32	59.77	91.98	20.42	78.75	64.14	50.95	64.95
DAE-Former	MICCAI 2023	65.93	34.62	66.64	58.39	86.94	27.12	78.82	51.36	79.47	51.44	79.98	73.73	92.01	42.66	88.16	67.86	58.73	71.91
HiFormer	WACV 2023	67.26	26.29	65.69	66.12	88.14	19.77	77.26	61.14	80.43	41.63	79.29	79.61	93.69	33.02	87.17	75.89	58.96	71.34
TransUnet	Medical Image Analysis 2024	75.69	40.89	71.74	71.22	91.01	32.90	82.52	60.51	86.16	58.05	83.55	83.19	95.29	49.51	90.42	75.40	65.81	77.70
NA-SegFormer	Scientific Reports 2024	**79.32**	31.16	73.21	67.34	88.12	30.06	72.08	58.82	**88.47**	47.52	84.53	80.48	93.68	46.22	83.78	74.07	62.51	74.84
AHF-U-net	Information Fusion 2024	65.01	30.35	64.61	55.32	89.40	10.97	76.48	51.65	78.80	46.57	78.50	71.23	94.40	19.77	86.67	68.12	55.47	68.01
GED-Net	ESWA 2024	68.78	30.12	75.98	69.11	90.44	26.44	80.45	62.82	81.50	46.29	86.35	81.73	94.98	41.82	89.17	77.16	63.02	74.88
VM-Unetv1	Arxiv 2024	64.50	17.36	57.15	51.23	89.29	18.07	77.92	59.46	78.42	29.58	72.73	67.75	94.34	30.61	87.59	74.58	54.37	66.95
VM-Unetv2	ISBRA 2024	60.71	26.28	73.09	62.47	87.43	23.16	77.67	57.99	75.55	41.62	84.45	76.90	93.29	37.61	87.43	73.41	58.60	71.28
Ours	–	77.64	**53.99**	**84.54**	**79.80**	**91.48**	31.67	83.05	64.30	87.41	**70.12**	**91.62**	**88.77**	**95.55**	**48.11**	90.74	78.72	**70.81**	**81.38**

NA-SegFormer, AHF-U-net, GED-Net, VM-Unetv1, and VM-Unetv2), our AFDSeg model achieved an average improvement of 4.0% in mIoU and 2.54% in mDice. Moreover, the proposed method outperformed all other methods across all metrics in the most challenging MYO category, demonstrating that AFDSeg effectively addressed the issues of small targets and blurry boundaries.

(5) Results of the Synapse Dataset:

A comparison experimental results of the Synapse dataset is presented in Table 6. Our AFDSeg method achieved the best results in an abdominal multi-organ segmentation task. Specifically, compared with classical methods such as Unet, DeepLabV2, and MISSFormer, AFDSeg demonstrated significant improvements across all organ segmentation tasks. Among all methods, the proposed method achieved the best performance in both mIoU and mDice, with improvements of 1.99% and 1.84%, respectively, over the second-best method. In terms of the IoU metric, AFDSeg achieved the best performance in four organ segmentations and the second-best performance in the other three segmentations. The Dice metric, achieved the best performance in five organ segmentations and the second-best performance in the remaining three. Thus, the superior performance for our method for multi-segmentation tasks was validated.

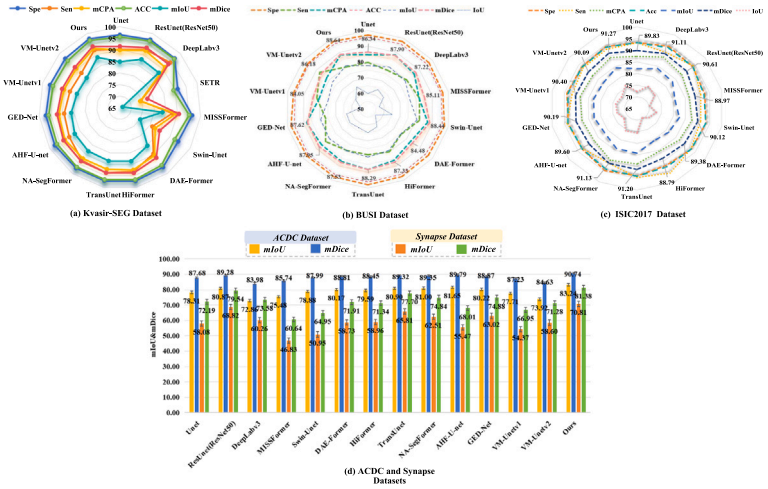


Fig. 5. Visualization of the quantitative comparison results between AFDSeg and state-of-the-art methods across five datasets.

In summary, the visualized quantitative comparison results (Fig. 5) demonstrated that our AFDSeg method achieved outstanding performance across the five datasets involving different modalities and disease types.

4.4.2. Qualitative results

For an intuitive experimental comparison, the qualitative results of our AFDSeg and state-of-the-art methods for binary segmentation tasks (conducted on the Kvasir-SEG, the BUSI, and the ISIC 2017 datasets) and multi-segmentation tasks (conducted on the ACDC and Synapse datasets) were analyzed, as shown in Figs. 6 and 7, respectively. As illustrated in Fig. 6, our AFDSeg achieved more accurate segmentation results than other methods in binary segmentation tasks. Even for challenging small targets, our method generated segmentation results that were closer to the ground truth, as seen in the fourth row of Fig. 6(a), and demonstrated superior performance for binary segmentation of breast lesions. The proposed method effectively captured the true breast boundaries by segmenting benign or malignant lesions. Notably, in the mixed benign and malignant experiments (Fig. 6(b)), our method accurately fitted the boundaries of highly variable and irregular benign and malignant lesions, while significantly reducing false negatives compared to other methods. Additionally, on the ISIC 2017 dataset, our method more effectively captures highly irregular melanoma contours compared to other approaches, further demonstrating its robustness (see Fig. 6(c)).

As depicted in Fig. 7, in multi-class segmentation tasks, existing methods exhibited numerous false positives and false negatives, particularly in multi-organ tasks, where the overall structure could be partially missing or the boundaries were inaccurately segmented. Conversely, the proposed method not only maintained high segmentation accuracy for both large and small targets, but also improved the completeness of the overall structure in medical image segmentation. These results confirmed that our approach, driven by heterogeneous networks and adaptive frequency-domain learning, effectively enhanced the performance of medical image segmentation.

4.5. Cross-domain segmentation for generalization performance analysis

To evaluate the generalizability of the proposed method, cross-domain segmentation experiments were conducted using three unseen datasets:

(1) Kvasir Capsule-SEG dataset [49]: It includes 55 endoscopic images of intestinal polyps and their corresponding segmentation masks. It is used for cross-domain polyp segmentation, where models are trained on the Kvasir-SEG and tested on Kvasir Capsule-SEG, denoted as Kvasir-SEG → Kvasir Capsule-SEG.

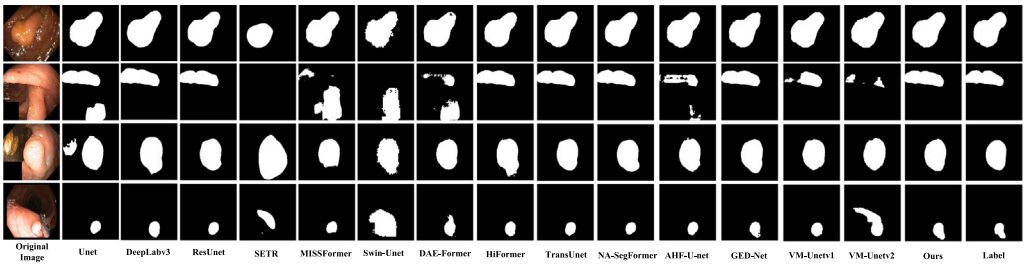
(2) BUS42 dataset [50]: It consists of 42 ultrasound images provided by the Imaging Department of the First Affiliated Hospital of Shantou University. As BUS42 does not distinguish between benign and malignant tumors, we merged the benign and malignant tumor subsets from the BUSI dataset to form a training set for this study. This setup is used to assess the generalization performance of different models on BUS42, denoted as BUSI → BUS42.

(3) M&Ms dataset [51]: This dataset presents a highly challenging cardiac magnetic resonance (CMR) segmentation task owing to its multi-center, multi-device, and multi-disease nature. For this study, we used its test set, comprising 4868 images, for cross-domain cardiac segmentation, where models were trained on ACDC and tested on M&Ms, denoted as ACDC → M&Ms.

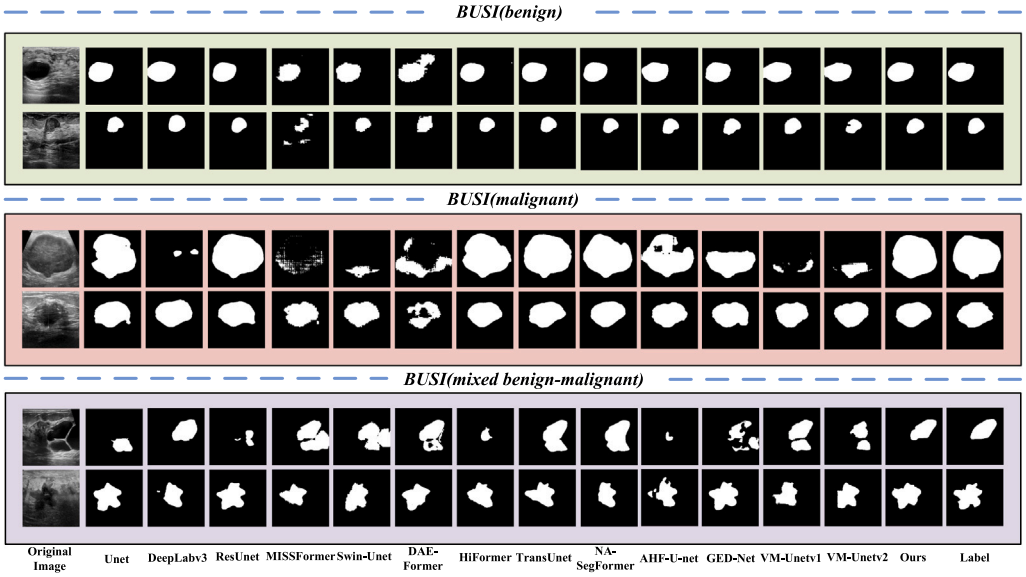
The quantitative results of the three cross-domain segmentation experiments are presented in Tables 7, 8, and 9, with the qualitative visual comparisons are shown in Figs. 8, 9, and 10, respectively.

The experimental results demonstrated that our method achieved the highest mIoU and mDice across the aforementioned three cross-domain segmentation tasks: Kvasir-SEG → Kvasir Capsule-SEG, BUSI → BUS42, and ACDC → M&Ms. Notably, our method outperformed the second-best approach considering mDice by 8.69%, 0.52%, and 0.59%, respectively, validating the superior generalization ability of the proposed method. As shown in Figs. 8 and 9, our method consistently achieves more anatomically coherent, complete, and continuous segmentation results, both in global structure representation and local detail preservation. To further demonstrate the effectiveness of our method in multi-class cross-domain segmentation, we utilized entropy maps [52] and t-SNE maps [53] in the ACDC→ M&Ms task to visualize prediction uncertainty and feature space distribution. As shown in Fig. 10, our method exhibited a more concentrated entropy distribution along clear and complete boundaries across all entropy maps, indicating higher confidence and better certainty in segmentation predictions. Additionally, in all t-SNE maps, the proposed method demonstrates well-defined and distinctly clustered categories, suggesting superior feature extraction and class discrimination capabilities. Moreover, the performance improvement of the proposed method in cross-domain segmentation of endoscopic polyp images was more significant than in breast tumor ultrasound and cardiac MRI segmentation tasks because our adaptive domain enhancement strategy was particularly effective in capturing the frequency characteristics of endoscopic polyp images.

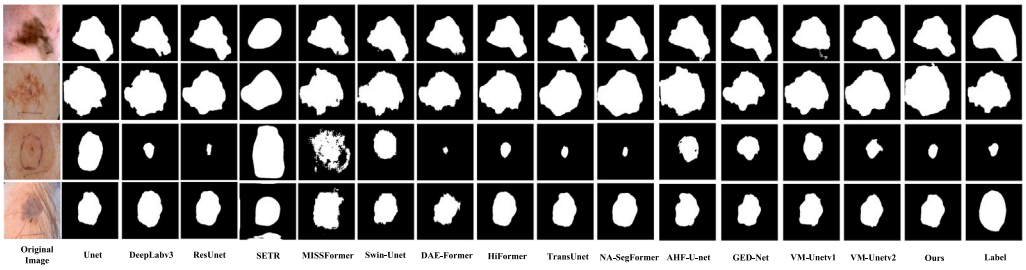
Overall, these rigorous experiments fully validated the strong generalization capability of the proposed method.



(a) Visual comparison of segmentation results on Kvasir-SEG dataset.



(b) Visual comparison of segmentation results for tumor segmentation on BUSI dataset.



(c) Visual comparison of segmentation results for melanoma segmentation on ISIC 2017 dataset.

Fig. 6. Visual comparison between AFDSeg and state-of-the-art methods for binary segmentation tasks on the Kvasir-SEG [44], BUSI [45] and ISIC 2017 datasets [46].

4.6. Ablation study

To validate the effectiveness of AFDSeg, we conducted a series of ablation experiments, including: (1) module ablation studies; (2) the impact of pre-training and non-pre-training on network performance; (3) module parameter analysis; (4) evaluation of the impact of different heterogeneous network architectures on medical image segmentation performance; (5) effect analysis of feature distillation alignment loss between heterogeneous networks; (6) ablation studies on the filter window size in the FAHS module to investigate its influence on feature response patterns.

Ablation study for proposed modules: As shown in Table 10, our progressive module fusion achieves significant improvements in key metrics, validating the effectiveness of our design. The proposed method achieves the best optimum performance across five key metrics (IoU, Dice, mIoU, and mDice). Specifically, our method attained a mIoU of 83.24% and mDice of 90.74%, representing improvements of 2.37% and 1.46%, respectively, compared to L1. Notably, the heterogeneous network for feature distillation (L2) enhances mDice by 0.39% and mIoU by 0.58%, indicating improved long-range feature modeling. The LHSD module (L3) significantly boosts Recall, effectively suppressing low-frequency noise in high-frequency features and enhancing foreground feature discrimination. The FAHS module (L5) further improves Recall and reduces false positives by adaptively coupling spatial and frequency domain features, enabling better multi-view feature interaction. As modules are progressively combined (L6–L8), Table 10 shows consistent improvements in mIoU, mDice, and Recall.

Additionally, Section 4.7 analyzes the computational complexity and mDice across five public datasets.

For qualitative analysis, we selected five cardiac cross-sectional image covering multiple views, various shapes, irregular structures. As illustrated in Fig. 11, L1 primarily focused on the internal regions of the

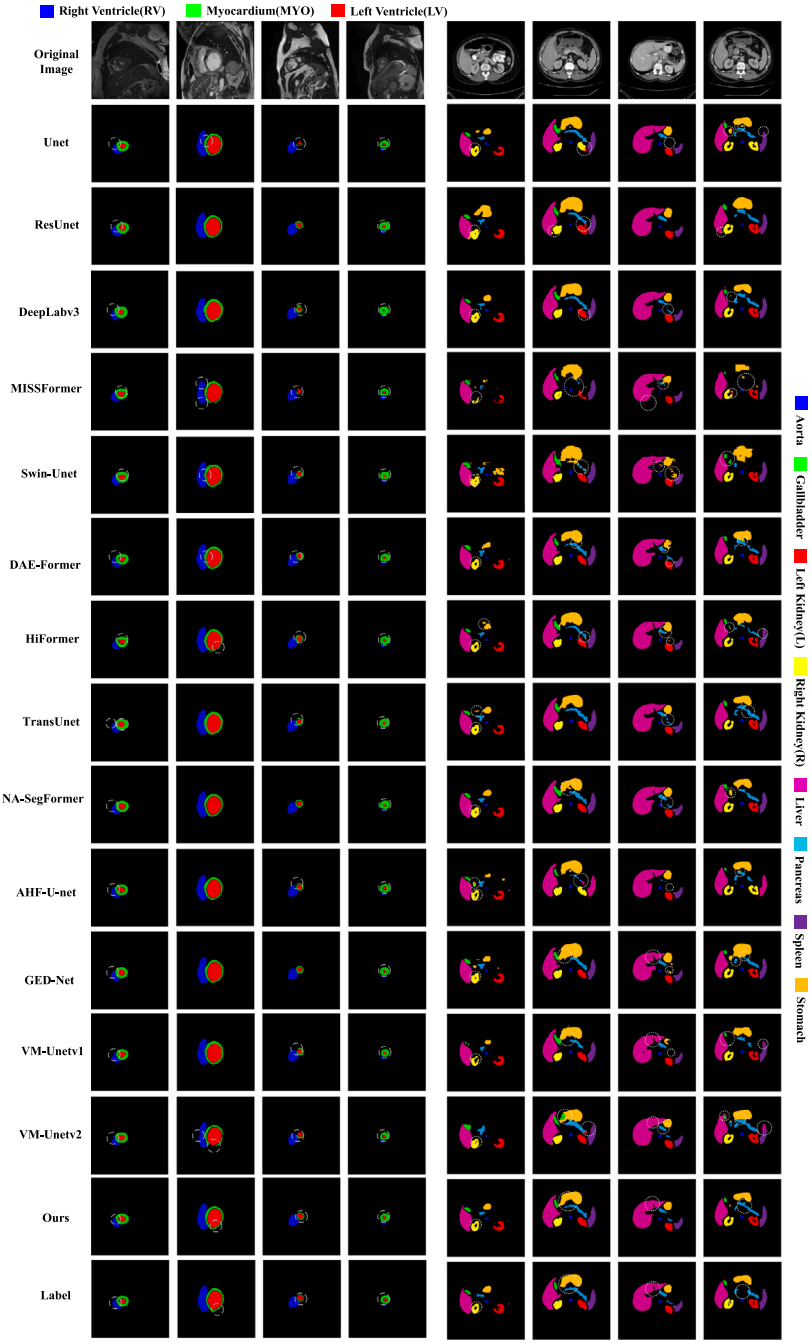


Fig. 7. Visual comparison between AFDSeg and state-of-the-art methods for multi-class segmentation tasks on ACDC [47] and Synapse datasets [21] (**Left:** Cardiac Segmentation; **Right:** Abdominal multi-organ Segmentation).

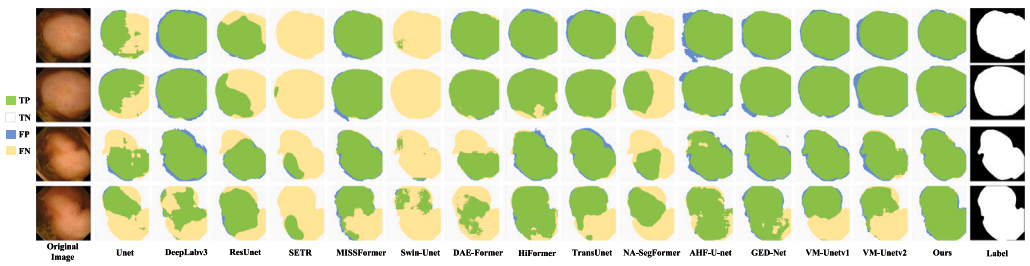


Fig. 8. Visual comparison between AFDSeg and state-of-the-art methods for cross-domain segmentation Kvasir-SEG → Kvasir Capsule-SEG.

Table 7
Cross-domain quantitative analysis (training on the Kvasir-SEG Dataset, model generalizability testing on the Kvasir Capsule-SEG dataset). The optimal results are marked as black bold while the suboptimal results are underlined "–".

Method	Venue	*CPA*		*IoU*		*Spe*	*Sen*	*Dice*		*mCPA*	*ACC*	*mIoU*	*mDice*
		Background	Polyp	Background	Polyp	Background	Polyp	Background	Polyp				
Unet	MICCAI 2015	97.84	43.20	50.02	42.65	50.58	97.11	66.68	59.80	70.51	63.56	46.33	63.24
DeepLabv3	CVPR 2017	92.56	72.70	63.42	69.62	66.83	94.27	77.62	82.09	82.63	80.10	66.52	79.86
ResUnet(ResNet50)	ISM 2019	96.11	48.76	51.61	47.66	52.71	95.47	68.08	64.55	72.44	66.41	49.63	66.32
SETR	CVPR 2021	98.15	25.07	43.41	24.80	43.77	95.79	60.54	39.74	61.60	52.31	34.10	50.14
MISSFormer	TMI 2022	93.35	70.55	62.41	67.87	65.32	94.69	76.86	80.86	81.95	79.05	64.14	78.86
Swin-Unet	ECCV 2022	**99.59**	10.78	39.81	10.76	39.88	97.79	56.95	19.43	55.19	43.88	25.29	38.19
DAE-Former	MICCAI 2023	98.23	52.35	54.51	51.81	55.06	**98.03**	70.56	68.26	75.29	69.45	53.16	69.41
HiFormer	WACV 2023	94.47	66.23	60.24	64.12	62.44	95.28	75.19	78.14	80.35	76.75	62.18	76.67
TransUnet	Medical Image Analysis 2024	96.02	59.65	57.25	58.27	58.64	96.18	72.82	73.63	77.84	72.23	57.76	73.23
NA-SegFormer	Scientific Reports 2024	97.91	45.44	51.05	44.89	51.61	97.35	67.59	61.96	71.68	65.00	47.97	64.78
AHF-U-net	Information Fusion 2024	87.96	76.70	63.19	71.58	69.17	91.47	77.44	83.44	82.33	80.90	67.38	80.44
GED-Net	ESWA 2024	92.95	78.27	68.07	75.13	71.77	94.92	81.00	85.80	85.61	83.74	71.60	83.40
VM-Unetv1	Arxiv 2024	90.71	75.96	64.58	71.98	69.15	93.23	78.48	83.71	83.33	81.46	68.28	81.10
VM-Unetv2	ISBRA 2024	93.40	56.44	53.89	54.31	56.03	93.50	70.04	70.39	74.92	70.22	54.10	70.22
Ours	–	91.34	**93.29**	**82.06**	**88.72**	**89.00**	94.77	**90.15**	**94.02**	**92.31**	**92.56**	**85.39**	**92.09**

Table 8
Cross-domain quantitative analysis (training on the BUSI Dataset, model generalizability testing on the BUS42 dataset). The optimal results are marked as black bold while the suboptimal results are underlined "–".

Method	Venue	*CPA*		*IoU*		*Spe*	*Sen*	*Dice*		*mCPA*	*ACC*	*mIoU*	*mDice*
		Background	BUSI	Background	BUSI	Background	BUSI	Background	BUSI				
Unet	MICCAI 2015	98.33	78.14	95.33	70.01	96.90	87.07	97.61	82.36	88.23	95.79	82.67	89.99
DeepLabv3	CVPR 2017	96.56	88.36	94.97	71.33	98.29	78.72	97.42	83.26	92.46	95.53	83.15	90.34
ResUnet(ResNet50)	ISM 2019	98.01	85.02	95.94	74.71	97.85	86.04	97.93	85.53	91.52	96.38	85.33	91.73
MISSFormer	TMI 2022	98.16	76.88	95.00	68.16	96.72	85.74	97.43	81.07	87.52	95.48	81.58	89.25
Swin-Unet	ECCV 2022	98.57	81.39	96.00	74.03	97.35	89.12	97.96	85.08	89.98	96.41	85.01	91.52
DAE-Former	MICCAI 2023	98.72	78.86	95.81	72.43	97.01	89.88	97.86	84.01	88.79	96.22	84.12	90.94
HiFormer	WACV 2023	97.47	86.20	95.57	73.31	98.22	83.06	97.73	84.60	91.84	96.05	84.44	91.17
TransUnet	Medical Image Analysis 2024	98.30	82.37	95.88	73.67	97.49	87.47	97.90	84.84	90.34	96.31	84.78	91.37
NA-SegFormer	Scientific Reports 2024	87.72	77.10	94.60	66.57	96.74	82.98	97.23	79.93	87.41	95.13	80.59	88.58
AHF-U-net	Information Fusion 2024	98.49	81.75	95.97	73.98	97.40	88.61	97.94	85.04	90.12	96.38	84.97	91.49
GED-Net	ESWA 2024	98.58	80.56	95.89	73.31	97.24	89.07	97.90	84.60	89.57	96.31	84.60	91.25
VM-Unetv1	Arxiv 2024	98.47	79.39	95.63	71.76	97.07	88.19	97.77	83.56	88.93	96.07	83.70	90.67
VM-Unetv2	ISBRA 2024	**99.02**	83.28	**96.69**	77.98	97.63	**92.45**	**98.32**	87.63	91.15	**97.04**	87.34	92.98
Ours	–	97.74	**92.21**	96.65	**79.69**	**98.86**	85.45	98.30	**88.70**	**94.97**	**97.04**	**88.17**	**93.50**

Table 9
Cross-domain quantitative analysis (training on the ACDC dataset, model generalizability testing on the M&MS dataset). The optimal results are marked as black bold while the suboptimal results are underlined "–".

Method	Venue	*IoU*			*Recall*			*Dice*			*mIoU*	*mDice*
		RV	MYO	LV	RV	MYO	LV	RV	MYO	LV		
Unet	MICCAI 2015	65.38	59.85	72.62	**91.78**	75.10	**96.00**	79.07	74.88	84.14	65.95	79.36
DeepLabv3	CVPR 2017	66.78	53.94	68.80	91.11	65.83	90.82	80.08	70.08	81.52	63.17	77.23
ResUnet(ResNet50)	ISM 2019	73.91	63.38	73.87	85.41	72.35	91.49	85.00	77.59	84.97	70.39	82.52
MISSFormer	TMI 2022	61.45	50.04	70.51	88.51	**77.48**	94.64	76.12	66.70	82.70	60.67	75.17
Swin-Unet	ECCV 2022	70.67	60.32	76.59	82.55	70.18	91.52	82.82	75.25	86.75	69.19	81.61
DAE-Former	MICCAI 2023	71.80	61.77	77.85	82.12	69.82	93.16	83.59	76.37	87.55	70.47	82.50
HiFormer	WACV 2023	72.32	60.83	76.65	91.09	73.34	93.32	83.94	75.64	86.78	69.93	82.12
TransUnet	Medical Image Analysis 2024	70.77	**65.30**	72.87	93.34	73.45	94.87	82.88	79.01	84.31	69.65	82.07
NA-SegFormer	Scientific Reports 2024	72.18	63.19	**77.88**	90.19	72.81	94.20	83.84	77.45	**87.57**	71.08	82.95
AHF-U-net	Information Fusion 2024	73.41	63.93	77.50	87.53	76.35	93.28	84.66	77.99	87.33	71.61	83.33
GED-Net	ESWA 2024	72.44	61.12	75.50	86.80	71.25	95.72	84.02	75.87	86.04	69.69	81.98
VM-Unetv1	Arxiv 2024	69.84	56.37	75.39	87.63	72.75	93.00	82.24	72.10	85.97	67.20	80.10
VM-Unetv2	ISBRA 2024	64.12	51.99	74.69	89.87	73.33	93.29	78.13	68.41	85.51	63.60	77.35
Ours	–	**76.57**	64.83	76.00	89.74	77.11	94.23	**86.73**	**78.66**	86.37	**72.47**	**83.92**

heart while lacking attention to its boundary regions. Contrarily, L3 and L5, incorporating LHSD and FAHS, respectively, enabled the network not only to enhance its focus on the heart's internal structures but also better delineateits contour boundaries . Our proposed model (L8) further refined this process by accurately capturing responses within the internal regions of the heart while simultaneously guiding the network to emphasize the contours of the heart, ultimately achieving precise segmentation.

Ablation study for the impact of pretrained and non-pretrained weights on model performance: Additionally, to further investigate the impact of pretrained weights on the proposed method, we conduct an ablation study by removing the pretrained weights. As shown in Table 11: (1) The mIoU and mDice achieved by our method without pretrained weights are 82.92% and 90.55%, respectively, which exhibit negligible differences compared to the results obtained with pretrained weights. This finding indicates that our AFDSeg is largely independent

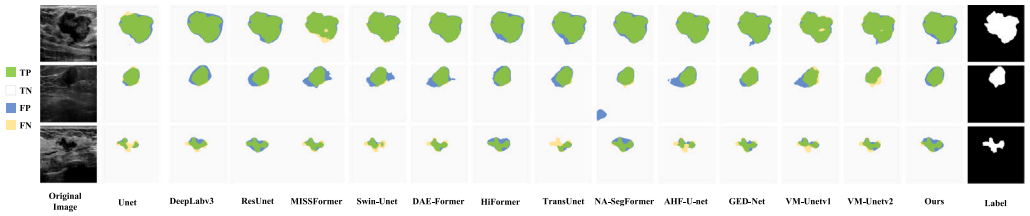


Fig. 9. Visual comparison between AFDSeg and state-of-the-art methods for cross-domain segmentation BUSI → BUS42.

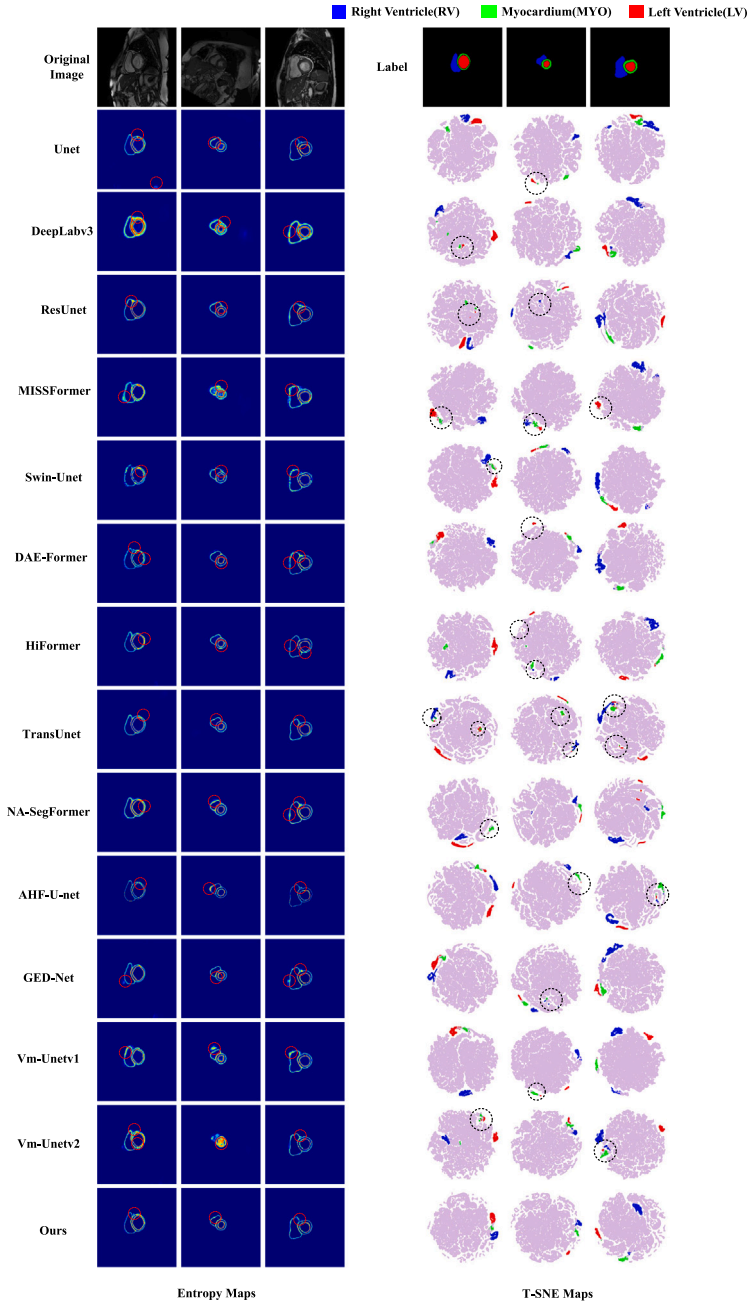


Fig. 10. Visual comparison between AFDSeg and state-of-the-art methods for cross-domain segmentation ACDC → M&Ms. **Left:** Entropy map [52] shows a better certainty for the predictions of our AFDSeg. **Right:** T-SNE [53] shows a better category separability for our AFDSeg.

of pre-trained weights. (2) Moreover, with the effective contribution of each module, AFDSeg can directly fit the data distribution based on the training samples, leading to a steady improvement in segmentation performance. This further highlights the robustness of our approach.

Ablation study for module parameter analysis: Finally, we report the parameter count and computational complexity of the proposed method. With an input resolution of 224 × 224 × 3, our model has 117.248M parameters and a complexity of 46.615G, primarily due to

Table 10
Ablation study of each module on the ACDC dataset(using pretrained model parameters). The optimal results are marked as black bold while the suboptimal results are underlined "–". L1 denotes a single-branch CNN (ResNet50), while L2 represents a heterogeneous network employing a dual-branch distillation framework that integrates CNN and Transformer. L3–L8 correspond to methods augmented with different additional modules.

Method	Distill	LHSD	PLFA	FAHS	MPFR	*IoU*			*Recall*			*Dice*			*mIoU*	*mDice*
						RV	MYO	LV	RV	MYO	LV	RV	MYO	LV		
L1(Baseline)	×	×	×	×	×	83.28	72.13	87.19	89.50	78.04	91.22	90.88	83.81	93.16	80.87	89.28
L2	✓					83.71	73.61	87.04	92.91	80.25	95.17	91.13	84.80	93.07	81.45	89.67
L3	✓	✓				82.16	73.00	88.43	94.29	84.22	94.40	90.21	84.39	93.86	81.20	89.49
L4	✓		✓			80.81	71.59	85.95	93.88	79.91	**95.87**	89.38	83.44	92.45	79.45	88.42
L5	✓			✓		77.48	72.71	85.21	**95.28**	83.00	94.59	87.31	84.20	92.02	78.47	87.84
L6	✓	✓	✓			81.00	70.92	85.73	91.23	75.69	95.00	89.50	82.99	92.32	79.22	88.27
L7	✓	✓	✓	✓		83.36	73.98	87.21	91.97	79.87	91.60	90.92	85.04	93.17	81.52	89.71
L8:Our	✓	✓	✓	✓	✓	**84.61**	**75.32**	**89.80**	93.35	**84.42**	94.95	**91.66**	**85.93**	**94.62**	**83.24**	**90.74**

Table 11
Ablation study of each module on the ACDC dataset(does not use pretrained model parameters). The optimal results are marked as black bold while the suboptimal results are underlined "–". L1 denotes a single-branch CNN (ResNet50), while L2 represents a heterogeneous network employing a dual-branch distillation framework that integrates CNN and Transformer. L3–L8 correspond to methods augmented with different additional modules.

Method	Distill	LHSD	PLFA	FAHS	MPFR	*IoU*			*Recall*			*Dice*			*mIoU*	*mDice*
						RV	MYO	LV	RV	MYO	LV	RV	MYO	LV		
L1(Baseline)	×	×	×	×	×	81.26	68.73	86.74	93.41	83.71	95.39	89.66	81.47	92.90	78.91	88.01
L2	✓					79.20	68.56	86.53	92.91	80.25	95.17	91.13	84.80	93.07	78.10	89.67
L3	✓	✓				78.59	68.92	86.59	93.87	83.51	93.66	88.63	81.60	92.81	78.03	87.68
L4	✓		✓			78.94	68.67	83.43	**94.05**	77.16	**95.74**	88.23	81.43	90.96	77.01	86.87
L5	✓			✓		80.87	69.05	86.94	92.10	**84.63**	91.67	89.43	81.69	93.01	78.95	88.04
L6	✓	✓	✓			80.21	69.85	87.01	91.24	80.32	92.69	89.02	82.25	93.06	79.02	88.11
L7	✓	✓	✓	✓		81.41	70.65	86.98	91.90	76.43	94.43	89.75	82.80	93.03	79.68	88.53
L8:Our	✓	✓	✓	✓	✓	**84.79**	**75.10**	**88.86**	93.42	84.02	93.73	**91.77**	**85.78**	**94.10**	**82.92**	**90.55**

Table 12
The number of parameters and computational complexity in the proposed module (MPFR receives low-resolution features (*y_feature*) and high-resolution features (*x_feature*) for multi-scale feature aggregation).

Module	Params(M)	FLOPs(G)	Resolutions
LHSD	0.617	9.957	
PLFA	0.063	0.886	$1 \times 64 \times 128 \times 128$
FAHS	0.017	0.207	
MPFR	0.369	6.063	*x_feature:* $(1 \times 64 \times 128 \times 128)$ *y_feature:* $(1 \times 128 \times 64 \times 64)$

the high parameter count of the baseline model and the additional plug-and-play modules in our heterogeneous network. To fairly assess module efficiency, we compute their FLOPs separately. As shown in Table 12, our plug-and-play modules have low parameter counts and computational complexity, enabling efficient expansion for lightweight medical image segmentation while balancing performance and computational cost.

Ablation study for heterogeneous networks: As seen in Table 13, with different heterogeneous network combinations, the overall mDice performance exhibited a slight decline compared to M1, which achieved an mDice of 90.74%. When the Transformer component was replaced with MiT-B5 while maintaining the CNN component as ResNet50 (i.e., M2), the mDice decreases by 1.34%. Further analysis in Fig. 12 reveals that M2 exhibited excessive coupling with the red region features, leading to an increased prediction entropy and degraded segmentation performance. Subsequently, replacing the Mamba component in the heterogeneous network with Mambav1 or Mambav2 resulted in a slight improvement in mDice compared with M2. This can be attributed to Mamba's ability to integrate both global and local features. However, as shown in Fig. 12(M5–M6), the segmentation performance remained suboptimal for small targets. Further, replacing the CNN component with ResNet101 and replacing the Transformer and Mamba components with MiT-B1, MiT-B5, Mambav1, and Mambav2 led to a performance decline compared with M1. As shown in Fig. 12(M3–M4, M7–M8), this can be attributed to interference from global features and insufficient responsiveness to local features resulted in the decline due to the increased difficulty of feature alignment introduced by ResNet101 and partly due to challenges in adapting to the ACDC dataset. Overall, the differences among heterogeneous network combinations on the ACDC dataset were relatively small. Compared with a single network, each combination achieves competitive results, demonstrating that our framework effectively enabled feature coupling and interaction between heterogeneous networks, thereby offering selection flexibility in the optimal combinations for different medical imaging scenarios.

Ablation study for feature distillation alignment loss: To further investigate feature alignment in heterogeneous networks, we used L1 loss, BCE loss, and $\mathcal{L}_{SA}$ to align features across different network architectures. As shown in Table 14, S3 achieved the highest mDice of 90.74%. S1 and S2 exhibited reductions of 2.53% and 2.85%, respectively, compared to S3 because L1 loss produced significantly small gradients, leading to insufficient feature alignment, whereas BCE loss coupled features from different architectures excessively, resulting in over-alignment. As illustrated in Fig. 13, S3 effectively captured both the internal and contour boundary features of the heart, whereas other methods exhibited inconsistencies in the boundary and small target features, adversely impacting the precise segmentation of medical images. In conclusion, S3 achieved a balanced performance in feature alignment for heterogeneous networks.

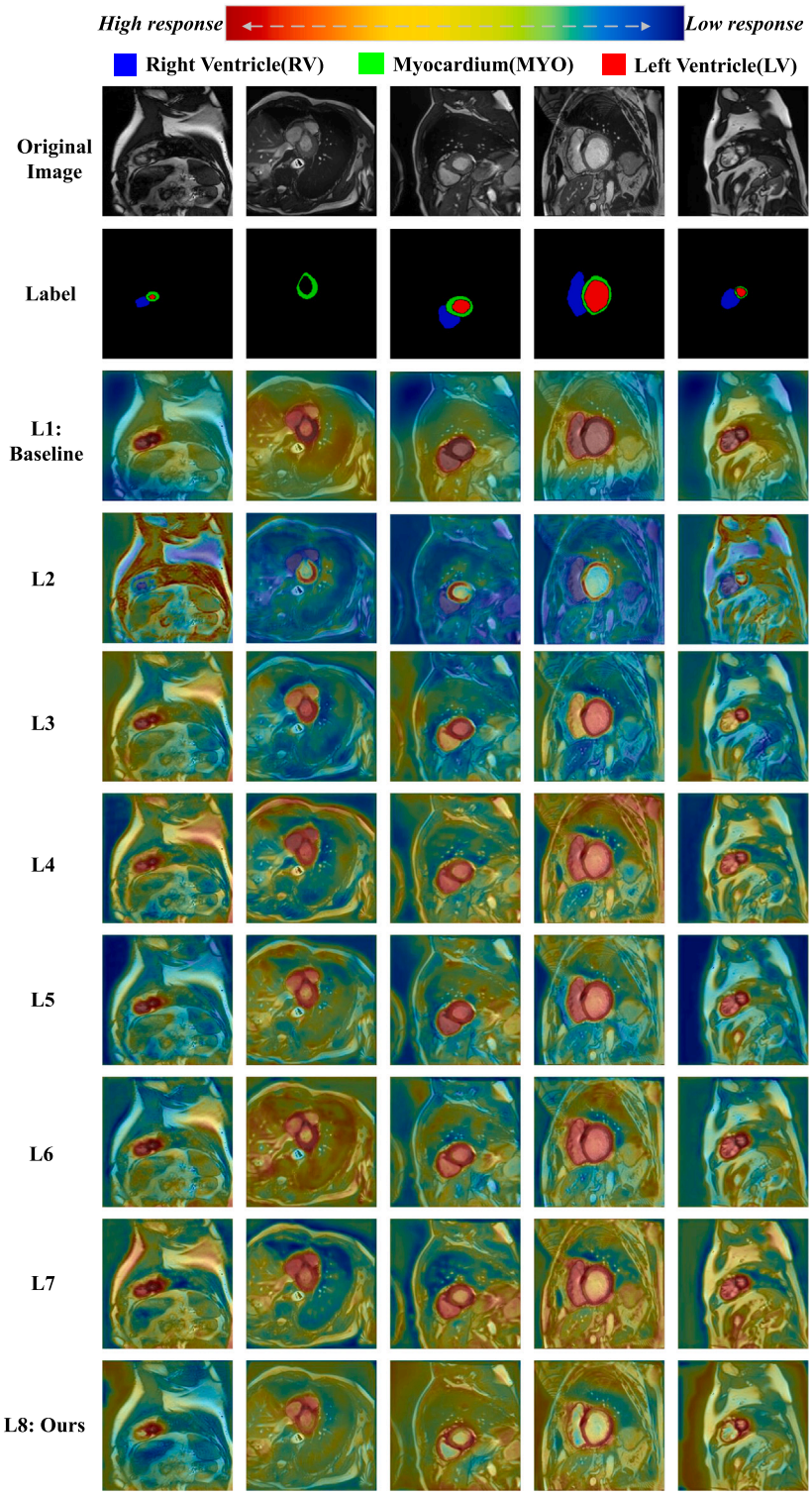


Fig. 11. Visuals of the attention regions of different modules on features using Grad-CAM [54].

Table 13
Results of ablation study of heterogeneous network on the ACDC dataset. The optimal results are marked as black bold while the suboptimal results are underlined "–".

Method	ResNet50	ResNet101	MiT-B1	MiT-B5	Mambav1	Mambav2	*IoU*			*Recall*			*Dice*			*mIoU*	*mDice*
							RV	MYO	LV	RV	MYO	LV	RV	MYO	LV		
M1:Ours	✓		✓				**84.61**	**75.32**	**89.80**	93.35	**84.42**	94.95	**91.66**	**85.93**	**94.62**	**83.24**	**90.74**
M2	✓			✓			82.45	73.14	87.48	92.96	81.89	94.75	90.38	84.49	93.32	81.02	89.40
M3		✓	✓				81.22	71.97	89.59	92.01	83.57	**95.75**	89.63	83.70	93.39	80.93	88.91
M4		✓		✓			81.53	71.52	87.26	91.72	81.97	92.55	89.83	83.40	93.20	80.10	88.81
M5	✓				✓		82.81	73.69	88.01	92.98	84.34	94.10	90.59	84.85	93.62	81.50	89.69
M6	✓					✓	82.07	73.64	88.25	**94.23**	82.33	95.34	90.15	84.82	93.76	81.32	89.58
M7		✓			✓		81.11	71.03	87.52	91.46	82.14	93.29	89.57	83.06	93.34	79.89	88.66
M8		✓				✓	83.11	72.66	88.72	93.61	81.14	84.52	90.78	84.16	94.02	81.50	89.65

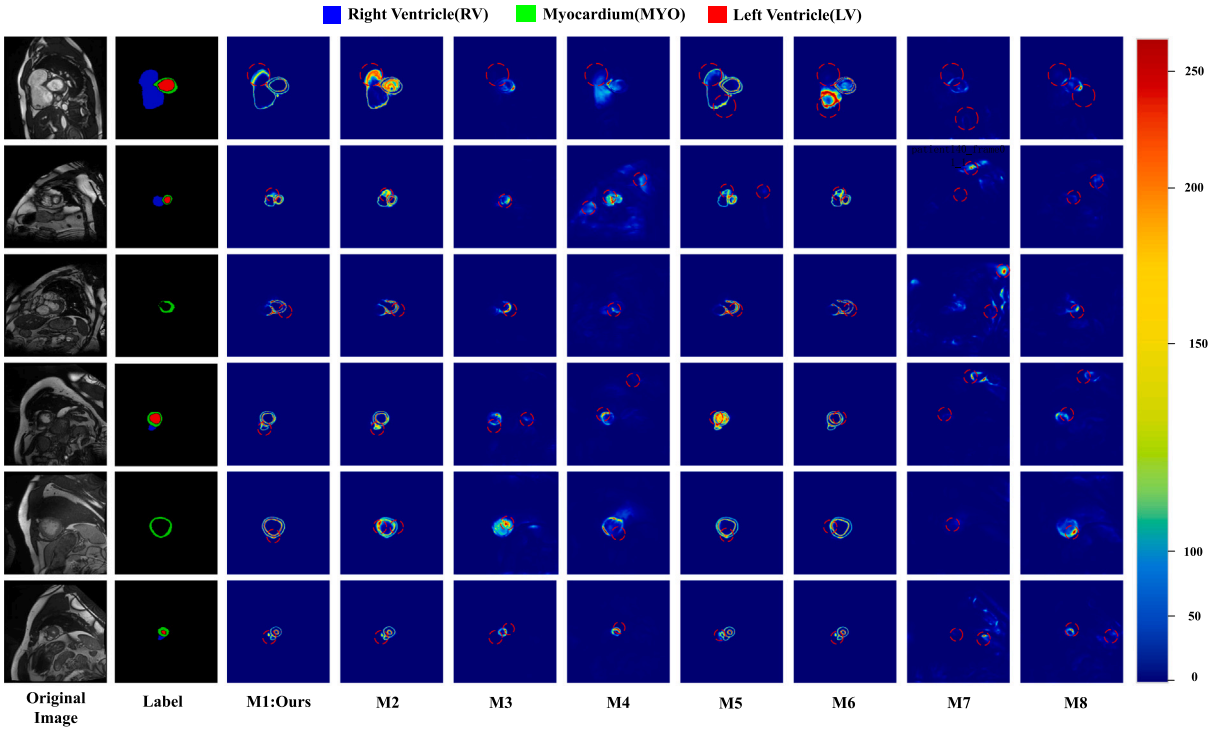


Fig. 12. Visualize the recognition accuracy of different heterogeneous networks on medical image features (red circles indicate regions with stronger feature coupling, where other models fail to perform fine-grained segmentation).

Table 14
Results of ablation study of feature distillation alignment loss on the ACDC dataset. The optimal results are marked as black bold.

Method	L1	BCE	$\mathcal{L}_{SA}$	*IoU*			*Recall*			*Dice*			*mIoU*	*mDice*
				RV	MYO	LV	RV	MYO	LV	RV	MYO	LV		
S1	✓			81.11	70.05	86.34	91.72	77.13	89.45	89.57	82.39	92.67	79.17	88.21
S2		✓		81.30	67.51	87.56	92.14	71.14	94.15	89.69	80.60	93.37	78.79	87.89
S3:(Ours)			✓	**84.61**	**75.32**	**89.80**	**93.35**	**84.42**	**94.95**	**91.66**	**85.93**	**94.62**	**83.24**	**90.74**

Ablation study for a filter window size of the FAHS module: As shown in Table 15, selecting moderate filter window sizes 10 in the first stage and 8 in the second stage yielded the highest value of mDice. As illustrated in Fig. 14, a moderate filter window size enabled a more efficient extraction of discriminative high-frequency cardiac features. Overall, regardless of the parameter combination, the proposed method consistently achieved competitive segmentation results, demonstrating its robustness.

4.7. Computational complexity analysis of the models

To further analyze the computational complexity, we used FLOPs as the evaluation metric and visualized the trade-off between FLOPs and mDice for all comparison methods, as shown in Fig. 15. The proposed method exhibited relatively high FLOPs, second only to AHF-UNet. The result was attributable to two main factors: first, the considered heterogeneous network incorporated multiple feature fusion and interaction modules; second, our frequency-domain feature learning introduced additional computationally intensive operations, such as frequency domain transformations, which increased the overall computational load. However, compared to state-of-the-art methods, our approach achieved significant improvements in segmentation accuracy, attaining the highest value of mDice and mIoU across all five datasets, despite the higher FLOPs. The trade-off between increased computational cost and substantial accuracy gains was justified, particularly in precision medicine applications where accuracy is paramount. Overall, the computational complexity of our remained acceptable among all methods.

Table 15
Results of ablation study for the size of the filtering window of the FAHS module on the ACDC dataset. The optimal results are marked as black bold. (Size: Filter Window Size).

Method	Size		*IoU*			*Recall*			*Dice*			*mIoU*	*mDice*
	2↓	4↓	RV	MYO	LV	RV	MYO	LV	RV	MYO	LV		
L1	5	2	82.46	71.93	87.41	90.20	**85.83**	91.64	90.38	83.67	93.28	80.60	89.11
L2	10	2	83.31	73.17	88.47	90.82	84.51	93.81	80.89	84.51	93.88	81.65	86.43
L3	10	5	80.16	71.13	85.71	88.91	79.16	91.25	88.99	83.13	92.31	79.00	88.14
L4:(Ours)	10	8	**84.61**	**75.32**	**89.80**	**93.35**	84.42	**94.95**	**91.66**	**85.93**	**94.62**	**83.24**	**90.74**
L5	15	8	82.85	72.53	87.83	87.87	79.94	91.78	90.62	84.08	93.52	81.07	89.41
L6	15	10	81.33	71.59	86.38	89.91	82.61	91.31	89.71	83.44	92.69	79.77	88.61
L7	20	10	82.73	72.80	88.05	90.36	82.78	95.00	90.55	84.26	93.65	81.19	89.49
L8	20	15	82.31	72.56	87.95	90.51	82.68	94.83	90.29	84.10	93.59	80.94	89.33

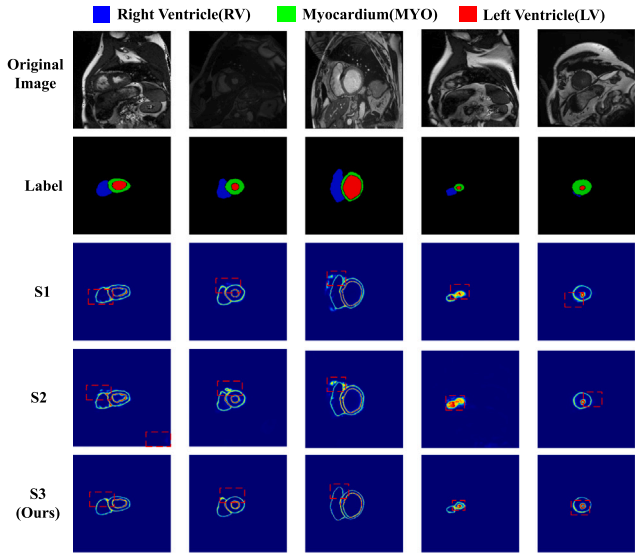


Fig. 13. Visualize the recognition accuracy of distillation alignment loss on medical image features (red rectangles indicate regions with stronger feature coupling, where other models fail to perform fine-grained segmentation).

With its high accuracy, the proposed method offers substantial potential for clinical applications.

5. Limitation analysis

Extensive experiments were performed to establish the superiority of our approach; however, it presents certain limitations. First, the model has a relatively large number of parameters, owing to its design as a general heterogeneous network framework for medical image segmentation. Although the dual-branch architecture enables cross-domain feature interaction, it introduces additional computational costs compared to single-branch models. Second, the performance of the frequency-domain enhancement mechanism is contingent upon the quality of the input images. For highly noisy or low-resolution medical images, the frequency decomposition process can amplify artifacts or lose critical structural information, potentially degrading performance. In future studies, we aim to address these limitations by optimizing the knowledge distillation framework and leveraging large-model-driven frequency-domain learning.

6. Conclusion and further work

This study introduces AFDSeg, a novel heterogeneous network-driven model designed to effectively extract and integrate multi-level frequency domain features in medical images, thereby enhancing segmentation performance. This improvement is realized through a dual-branch encoding structure, that facilitates feature learning. First, the FAHS module is designed to adaptively extract high-frequency texture and detailed features, while the PLFA module enhances the most informative low-frequency features to strengthen structural consistency. Second, an additional segmentation branch is incorporated, employing the LHSD module to suppress noise in the high-frequency feature extraction, thereby ensuring the effective integration of discriminative features for enhanced segmentation accuracy. Finally, a heterogeneous network architecture is employed to enable effective feature interaction and seamless fusion across diverse components.

To rigorously assess the effectiveness of our method, comprehensive experiments on five widely used public datasets were performed: Kvasir-SEG, BUSI, ISIC 2017, ACDC, and Synapse. Both qualitative and quantitative comparisons against state-of-the-art methods were conducted to comprehensively evaluate the segmentation performance of our approach. The experimental results demonstrate that our method consistently outperforms all baseline models in key performance metrics, including mIoU, ACC, and Dice scores. Additionally, extensive ablation studies were performed to systematically analyze the contribution of each module. To improve the interpretability of model predictions, we leveraged visualization techniques including Grad-CAM [54], entropy maps [52], and t-SNE maps [53]. These tools provide strong visual evidence that our AFDSeg effectively attends to and accurately delineates target segmentation regions in medical images. Furthermore, to assess the generalization capability of our method, we conducted cross-domain experiments on three external datasets: Kvasir Capsule-SEG, BUS42, and M&Ms. The results indicate that our method consistently captures key features efficiently and achieves superior generalization performance across diverse medical imaging domains, outperforming all competing approaches.

Beyond medical image segmentation, our heterogeneous AFDSeg exhibits strong adaptability to other fine-grained visual recognition tasks, such as remote sensing image segmentation and industrial defect detection. Specifically, the FAHS module enhances high-frequency texture feature extraction, allowing for precise delineation of fine details such as building edges and road networks, while the PFLA module strengthens the representation of low-frequency structural patterns, including land cover categories, to improve the semantic segmentation of satellite/aerial imagery. The denoising capability of LHSD module and multi-scale alignment through MPFR enhance anomaly detection in manufacturing, where subtle surface defects often result in localized high-frequency perturbations.

Future studies will focus on: (1) developing a more computationally efficient heterogeneous network by optimizing the knowledge distillation framework, allowing AFDSeg to transfer its learned representations to a compact single-branch model and reduce computational costs while maintaining accuracy. (2) dynamically optimizing filter window sizes using learnable neural operators and reparameterization techniques to adaptively determine optimal window sizes. (3) modeling temporal variations and frequency-domain features in 3D medical image segmentation through a dual-path 3D segmentation network (the spatiotemporal path integrates a lightweight 3D Swin

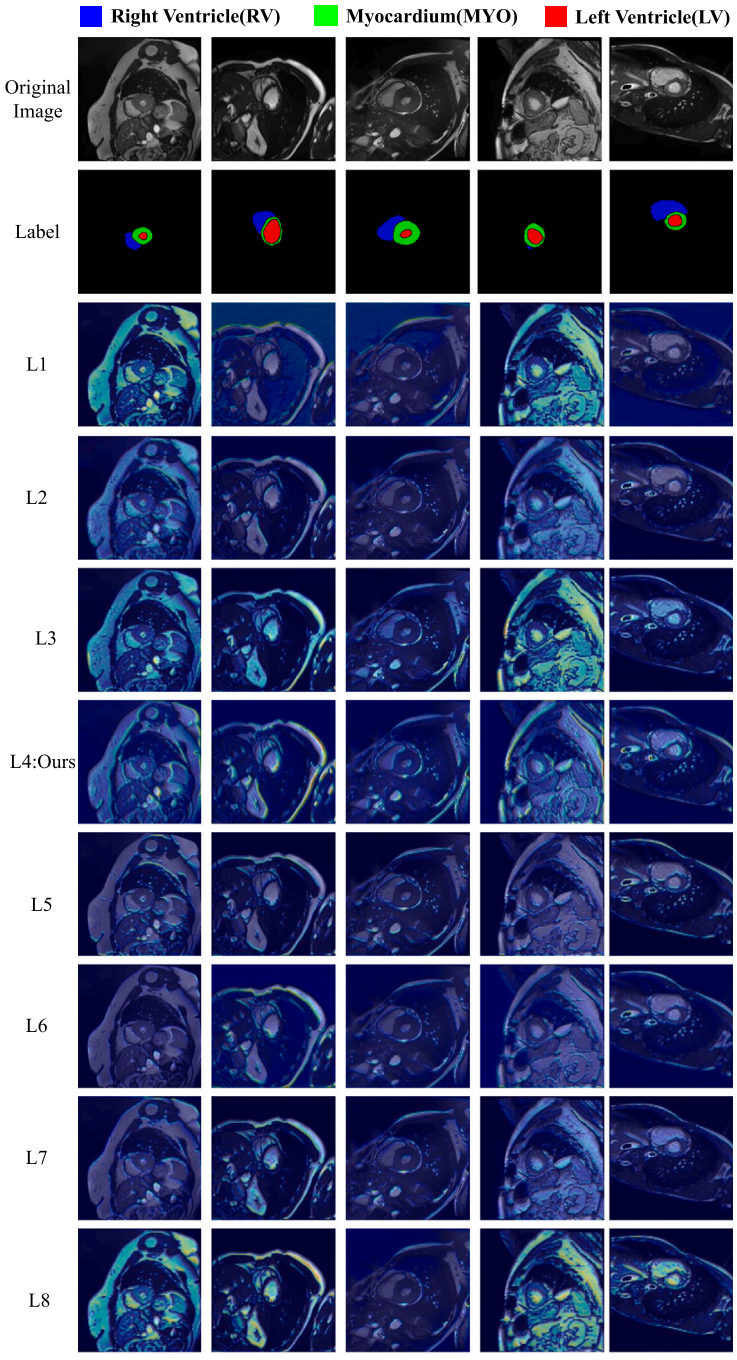


Fig. 14. Visualize the feature response of different filter window sizes after FAHS using Grad-CAM [54].

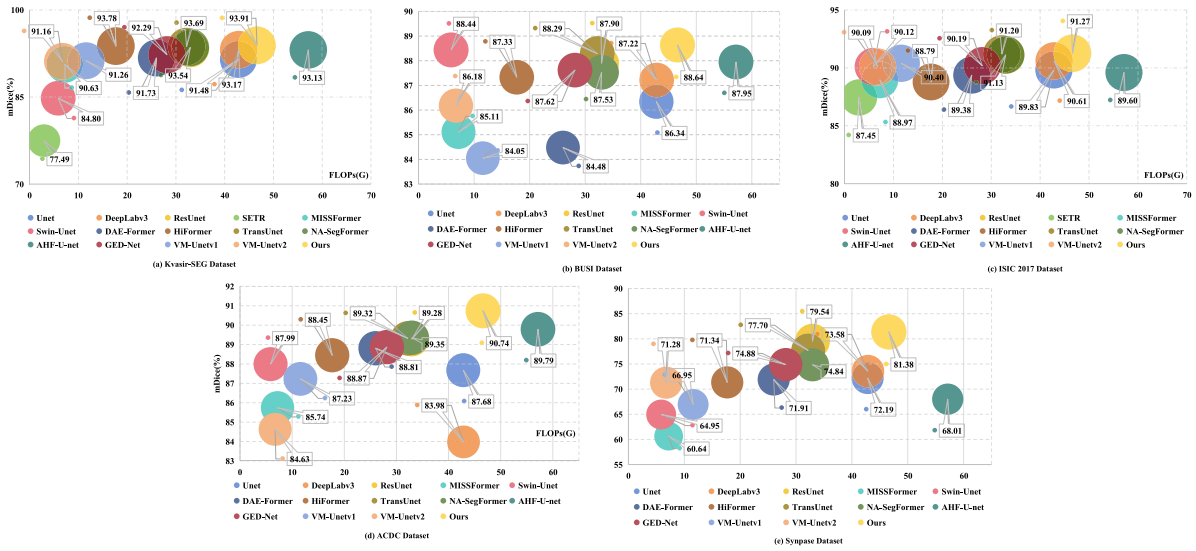


Fig. 15. Bubble plots of FLOPs and mDice composite metrics for each dataset.

Transformer to capture localized anatomical changes, whereas the frequency-domain path leverages a learnable wavelet transform to extract multi-scale frequency features, allowing for a more robust feature representation and segmentation accuracy for complex anatomical structures). (4) exploring large-model-driven frequency-domain learning by leveraging the cross-domain transferability of large models to reduce dependency on extensive annotated datasets and improve the segmentation performance on low-quality medical images. (5) performing clinical integration studies in collaboration with radiologists by deploying AFDSeg as a PACS plugin to evaluate its impact on diagnostic workflows and reduce inter-observer variability.

CRediT authorship contribution statement

Dong Liu: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Jin Kuang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, 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 research was funded by Natural Science Foundation of Hunan Province (No. 2023JJ50392, No. 2023JJ50393), Scientific Research Fund of Hunan Provincial Education Department (No. 23A0588), and Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province.

Data availability

The datasets used in this paper are all publicly available: Kvasir-SEG dataset can be found at https://datasets.simula.no/kvasir-seg/; BUSI dataset can be found at https://scholar.cu.edu.eg/?q=afahmy/pages/dataset; ISIC 2017 dataset can be found at https://challenge.isic-archive.com/data/#2017; ACDC dataset can be found at https://acdc.creatis.insa-lyon.fr/description/databases.html; Synapse dataset can be found at https://www.synapse.org/Synapse:syn3193805/wiki/217789; Kvasir Capsule-SEG dataset can be found at https://datasets.simula.no/kvasir-capsule-seg/; BUI42 dataset can be found at https://github.com/xbhlk/STU-Hospital.git; M&Ms dataset can be found at https://www.ub.edu/mnms/. In addition, the code for the proposed model in this paper is available on https://github.com/promisedong/AFDSeg.

References

[1] C. Shen, W. Li, H. Chen, X. Wang, F. Zhu, Y. Li, X. Wang, B. Jin, Complementary information mutual learning for multimodality medical image segmentation, Neural Netw. 180 (2024) 106670.

[2] S. Zhong, W. Wang, Q. Feng, Y. Zhang, Z. Ning, Cross-view discrepancy-dependency network for volumetric medical image segmentation, Med. Image Anal. 99 (2025) 103329.

[3] C. Suo, T. Zhou, K. Hu, Y. Zhang, X. Gao, Cross-level collaborative context-aware framework for medical image segmentation, Expert Syst. Appl. 236 (2024) 121319.

[4] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv. arXiv preprint arXiv:1706.05587 5.

[5] R. Azad, R. Arimond, E.K. Aghdam, A. Kazerouni, D. Merhof, Dae-former: Dual attention-guided efficient transformer for medical image segmentation, in: International Workshop on PRedictive Intelligence in Medicine, Springer, 2023, pp. 83–95.

[6] J. Ruan, S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, 2024, arXiv preprint arXiv:2402.02491.

[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[8] Z. Zhu, Z. Wang, G. Qi, N. Mazur, P. Yang, Y. Liu, Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction, Pattern Recognit. 153 (2024) 110553.

[9] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

[10] 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: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[11] S. Bansal, S. Madisetty, M.Z.U. Rehman, C.S. Raghaw, G. Duggal, N. Kumar, et al., A comprehensive survey of mamba architectures for medical image analysis: Classification, segmentation, restoration and beyond, 2024, arXiv preprint arXiv:2410.02362.

[12] M. Zhang, Y. Yu, S. Jin, L. Gu, T. Ling, X. Tao, VM-UNET-V2: rethinking vision mamba unet for medical image segmentation, in: International Symposium on Bioinformatics Research and Applications, Springer, 2024, pp. 335–346.

[13] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, et al., TransUNet: Rethinking the U-net architecture design for medical image segmentation through the lens of transformers, Med. Image Anal. 97 (2024) 103280.
[14] Y. Wang, C. Jiang, S. Luo, Y. Dai, J. Zhang, Graph neural network enhanced dual-branch network for lesion segmentation in ultrasound images, Expert Syst. Appl. 256 (2024) 124835.
[15] L. Chen, Y. Fu, L. Gu, C. Yan, T. Harada, G. Huang, Frequency-aware feature fusion for dense image prediction, IEEE Trans. Pattern Anal. Mach. Intell. (2024).
[16] Y. Rao, W. Zhao, Z. Zhu, J. Lu, J. Zhou, Global filter networks for image classification, Adv. Neural Inf. Process. Syst. 34 (2021) 980–993.
[17] L. Chi, B. Jiang, Y. Mu, Fast fourier convolution, Adv. Neural Inf. Process. Syst. 33 (2020) 4479–4488.
[18] A. Tragakis, Q. Liu, C. Kaul, S.K. Roy, H. Dai, F. Deligianni, R. Murray-Smith, D. Faccio, GLFNET: Global-local (frequency) filter networks for efficient medical image segmentation, 2024, arXiv preprint arXiv:2403.00396.
[19] P. Li, R. Zhou, J. He, S. Zhao, Y. Tian, A global-frequency-domain network for medical image segmentation, Comput. Biol. Med. 164 (2023) 107290.
[20] Z. Zou, H. Yu, J. Huang, F. Zhao, Freqmamba: Viewing mamba from a frequency perspective for image deraining, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 1905–1914.
[21] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, A. Klein, Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge, vol. 5, 2015, p. 12.
[22] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 263–273.
[23] Z. Zhu, Z. Zhang, G. Qi, Y. Li, Y. Li, L. Mu, A dual-branch network for ultrasound image segmentation, Biomed. Signal Process. Control. 103 (2025) 107368.
[24] W. Meng, S. Liu, H. Wang, AFC-unet: Attention-fused full-scale CNN-transformer unet for medical image segmentation, Biomed. Signal Process. Control. 99 (2025) 106839.
[25] W. Dong, B. Du, Y. Xu, Shape-intensity-guided U-net for medical image segmentation, Neurocomputing 610 (2024) 128534.
[26] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H.D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia, ISM, IEEE, 2019, pp. 225–2255.
[27] A.A. Munia, M. Abdar, M. Hasan, M.S. Jalali, B. Banerjee, A. Khosravi, I. Hossain, H. Fu, A.F. Frangi, Attention-guided hierarchical fusion U-net for uncertainty-driven medical image segmentation, Inf. Fusion 115 (2025) 102719.
[28] Y. Zhang, J. Yin, Y. Gu, Y. Chen, Multi-level feature attention network for medical image segmentation, Expert Syst. Appl. 263 (2025) 125785.
[29] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).
[30] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.
[31] Z. Zhu, M. Sun, G. Qi, Y. Li, X. Gao, Y. Liu, Sparse dynamic volume TransUNet with multi-level edge fusion for brain tumor segmentation, Comput. Biol. Med. (2024) 108284.
[32] X. Huang, Z. Deng, D. Li, X. Yuan, Y. Fu, Missformer: An effective transformer for 2d medical image segmentation, IEEE Trans. Med. Imaging 42 (5) (2022) 1484–1494.
[33] D. Liu, C. Lu, H. Sun, S. Gao, NA-segformer: A multi-level transformer model based on neighborhood attention for colonoscopic polyp segmentation, Sci. Rep. 14 (1) (2024) 22527.
[34] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024, arXiv preprint arXiv:2401.09417.
[35] Z. Zhu, K. Yu, G. Qi, B. Cong, Y. Li, Z. Li, X. Gao, Lightweight medical image segmentation network with multi-scale feature-guided fusion, Comput. Biol. Med. 182 (2024) 109204.
[36] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, L. Li, Mamba-unet: Unet-like pure visual mamba for medical image segmentation, 2024, arXiv preprint arXiv:2402.05079.
[37] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Inf. Fusion 91 (2023) 376–387.
[38] Y.-S. Ye, M.-R. Chen, H.-L. Zou, B.-B. Yang, G.-Q. Zeng, GID: Global information distillation for medical semantic segmentation, Neurocomputing 503 (2022) 248–258.
[39] D. An, P. Liu, Y. Feng, P. Ding, W. Zhou, B. Yu, Dynamic weighted knowledge distillation for brain tumor segmentation, Pattern Recognit. 155 (2024) 110731.
[40] E.B. Loussaief, H.A. Rashwan, M. Ayad, A. Khalid, D. Puig, Adaptive weighted multi-teacher distillation for efficient medical imaging segmentation with limited data, Knowl.-Based Syst. (2025) 113196.
[41] Y. Huang, C. Zhou, L. Chen, J. Chen, S. Lan, Medical frequency domain learning: Consider inter-class and intra-class frequency for medical image segmentation and classification, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2021, pp. 897–904.
[42] Y. Chen, X. Zhang, L. Peng, Y. He, F. Sun, H. Sun, Medical image segmentation network based on multi-scale frequency domain filter, Neural Netw. 175 (2024) 106280.
[43] W. Xu, R. Xu, C. Wang, X. Li, S. Xu, L. Guo, PSTNet: Enhanced polyp segmentation with multi-scale alignment and frequency domain integration, IEEE J. Biomed. Heal. Inform. (2024).
[44] D. Jha, P. Smedsrud, M. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. Johansen, Kvasir-seg: A segmented polyp dataset, international conference on multimedia modeling,
[45] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2020) 104863.
[46] D. Gutman, N.C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC), 2016, arXiv preprint arXiv:1605.01397.
[47] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M.A.G. Ballester, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11) (2018) 2514–2525.
[48] 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: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6202–6212.
[49] P.H. Smedsrud, V. Thambawita, S.A. Hicks, H. Gjestang, O.O. Nedrejord, E. Næss, H. Borgli, D. Jha, T.J.D. Berstad, S.L. Eskeland, et al., Kvasir-capsule, a video capsule endoscopy dataset, Sci. Data 8 (1) (2021) 142.
[50] Z. Zhuang, N. Li, A.N. Joseph Raj, V.G. Mahesh, S. Qiu, An RDAU-NET model for lesion segmentation in breast ultrasound images, PLoS One 14 (8) (2019) e0221535.
[51] V.M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P.M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, et al., Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge, IEEE Trans. Med. Imaging 40 (12) (2021) 3543–3554.
[52] A. Saporta, T.-H. Vu, M. Cord, P. Pérez, ESL: Entropy-guided self-supervised learning for domain adaptation in semantic segmentation, 2020, arXiv preprint arXiv:2006.08658.
[53] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).
[54] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.