FDM-Net: Frequency-domain modulation network for medical image segmentation

Original Paper | Published: 24 April 2026

Volume 20, article number 284 (2026) Cite this article

Signal, Image and Video Processing

Aims and scope →

Submit manuscript →

ShuMin Ma & Cao Shi

Abstract

Due to the inherent noise and artifacts of 3D medical imaging, issues such as low contrast and signal interference make segmentation still highly challenging. Existing methods mainly focus on modeling spatial information, while the intrinsic texture and directional cues embedded in frequency-domain representations remain underexplored. To address this, we propose a Frequency-Domain Modulation Network (FDM-Net), which incorporates frequency-domain cues into the segmentation process. To differentiate between the global morphological components of primary features and the fine-grained texture components, we develop a Frequency Decomposition (FD) module that separates images into heterogeneous frequency components, thereby highlighting overall contours and microscopic edges. To capture the characteristics of both high-frequency and low-frequency features, we introduce a Frequency-domain Adaptive Convolution (FAConv), which dynamically models convolutional kernel weights to enhance the dual-branch

FDM-Net: Frequency-Domain Modulation Network for Medical Image Segmentation

ShuMin Ma[1] and Cao Shi[1*]

[1*]College of Information Science and Technology, Qingdao University of Science and Technology,Qingdao, 266061, Shandong, China.

*Corresponding author(s). E-mail(s): caoshi@yeah.net;
Contributing authors: shuminma418@qq.com;

Abstract

Due to the inherent noise and artifacts of 3D medical imaging, issues such as low contrast and signal interference make segmentation still highly challenging. Existing methods mainly focus on modeling spatial information, while the intrinsic texture and directional cues embedded in frequency-domain representations remain underexplored. To address this, we propose a Frequency-Domain Modulation Network (FDM-Net), which incorporates frequency-domain cues into the segmentation process. To differentiate between the global morphological components of primary features and the fine-grained texture components, we develop a 3D Frequency-Domain Decomposition (FD) module that separates images into heterogeneous frequency components, thereby highlighting overall contours and microscopic edges. To capture the characteristics of both high-frequency and low-frequency features, we introduce a Frequency-domain Adaptive Convolution (FAConv), which dynamically models convolutional kernel weights to enhance the dual-branch responsiveness to different frequencies. Furthermore, to effectively integrate multi-scale and multi-branch representations, we design a Frequency-Guided Hybrid Attention (FGHA) module that aggregates dimension-specific information. Through visualization and experimental analysis, extensive results on three different imaging modality datasets demonstrate that the proposed FDM-Net achieves superior performance and robustness compared to state-of-the-art methods, while maintaining low computational cost.

Keywords: Frequency Domain,Medical Image Segmentation,Adaptive Convolution,Frequency-domain Decomposition,Feature Fusion

1 Introduction

Accurate medical image segmentation is central to computer-aided diagnosis, supporting tasks such as lesion delineation, surgical planning, and disease grading [1–3]. Manual annotation, however, is labor-intensive, inefficient, and subject to inter-observer variability [4, 5]. In 3D imaging, precise lesion segmentation is particularly critical for treatment planning, making the development of efficient and reliable automatic methods an essential research focus.

Over the past decade, deep learning techniques have been widely applied in the field of computer vision. Convolutional neural networks (CNNs) [6], with their powerful feature representation and capability of extracting deep semantic information, have driven significant progress in image segmentation. In particular, the U-shaped encoder–decoder architecture introduced by U-Net [7] has become a fundamental approach for medical image segmentation. ConvNeXt [8], inspired by the design principles of Transformers [9], achieves competitive segmentation performance while maintaining a purely convolutional architecture. In addition, other convolution-based networks such as [10–12] have also demonstrated remarkable segmentation accuracy. However, due to the inherently local receptive field of convolutions, CNNs face limitations in capturing global interactions and structured information. To mitigate this issue, Yu et al. [13] proposed dilated convolutions, which expand the receptive field by increasing the sampling rate. On the other hand, several Transformer-based architectures have been proposed. For example, Swin Transformer [14] introduces a shifted window attention mechanism that alleviates the heavy computational burden of conventional Transformers. I2U-Net [15] incorporates a hierarchical feature interaction mechanism between the encoder and decoder, enabling richer information flow and enhancing multi-scale semantic modeling and segmentation accuracy. Although these methods perform well in integrating spatial-domain features,

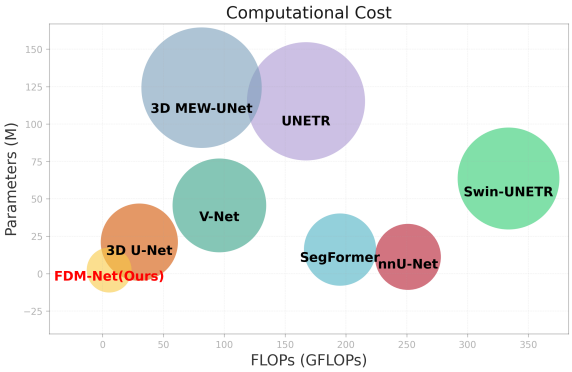

Fig. 1: Visualization of the comparison between methods in terms of computational complexity. The computational cost is represented using parameter size (bubble size) and GFLOPs. Our proposed method (FDM-Net) achieves better or comparable segmentation performance while having the smallest number of parameters and GFLOPs.

they pay limited attention to texture and directional information, which are crucial for highlighting lesion heterogeneity in medical imaging [16].

In recent years, research on medical image segmentation has increasingly recognized the role of frequency-domain representations in enhancing fine-grained texture capture and boundary delineation [17–19]. Li et al. [20] proposed a segmentation network based on global frequency-domain information, but it exhibits low sensitivity to local details. SFFNet [21] focuses on frequency decomposition with particular emphasis on regions near edges. However, most studies tend to process frequency-domain features in a unified manner, without explicitly separating low-frequency components, which primarily reflect overall structural patterns, from high-frequency components, which characterize subtle textures and directional variations.

To address these challenges, we propose FDM-Net, a frequency-domain–aware framework. It employs a Frequency Decomposition (FD) module to extract complementary high-frequency and low-frequency components, thereby enabling stable capture of both texture and structural features [22, 23]. FAConv enhances the network's sensitivity to different frequency characteristics by learning frequency-specific representations, which further improves contour detection and fine-grained texture modeling. FGHA strengthens global information correlation and local information interaction. By modulating features in the frequency domain, the framework integrates frequency- and spatial-domain information in a synergistic manner, aligning segmentation more closely with medical structural characteristics.

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

- To extract both low- and high-frequency features, we design a 3D frequency decomposition module. A Gaussian soft-gating mechanism is used to smoothly filter high-frequency components while separating low- and high-frequency information. This dual-branch learning strategy enhances the model's ability to perceive morphological boundaries and micro-textures.
- We introduce a 3D adaptive convolution in the frequency domain that constructs kernel weights based on frequency information. By separately learning low- and high-frequency representations, FAConv improves the network's responsiveness to different frequency ranges, thereby adaptively capturing diverse frequency information.
- To handle the microstructural patterns and heterogeneity of 3D medical data, FGHA leverages cross-attention in the frequency domain to capture multi-dimensional and multi-scale information, enabling more comprehensive representation integration and feature extraction.
- On multiple datasets, including OAI-ZIB [24], Pancreas [25], and Hippocampus [25], our method outperforms the comparative approaches in segmentation performance, demonstrating strong robustness and low computational cost (Fig.1).

2 Related Work

2.1 Image Semantic Segmentation

Semantic segmentation methods rooted in deep learning frameworks have developed rapidly [26–29]. Recent approaches based on a single backbone primarily leverage CNNs and their variants. The pioneering U-Net framework, along with its adaptive successor nnU-Net and Dense UNet [30], relies on an encoder–decoder architecture with skip connections to enhance feature representation. Attention mechanisms further improve the model's ability to focus on salient features. Attention U-Net [31] incorporates attention gates into the standard U-Net framework, enabling dynamic localization of important regions while suppressing background. ResNet [32] introduces residual connections to alleviate gradient vanishing issues, thereby pushing accuracy boundaries. Transformer-only architectures such as Swin U-Net [33] adopt a shifted window attention mechanism, effectively reducing computational complexity while maintaining strong robustness. However, these single-framework approaches [8, 34] remain limited in contextual understanding or require large-scale datasets for training.

Hybrid frameworks have been widely applied in image segmentation. UNETR [35] incorporates Transformers into a U-shaped architecture to capture multi-scale information. TransUNet [36] enhances the understanding of complex structures by combining local feature extraction of CNNs with global context modeling of Transformers. Hulin et al. [37] proposed a dual-branch network that improves segmentation accuracy through iterative feature interaction between CNN and Transformer branches, along with bilateral difference learning.

Although these architectures effectively integrate spatial-domain information of images, they primarily focus on spatial representations while overlooking the potential of frequency-domain modeling. In contrast, high-frequency information in the frequency domain plays a crucial role in delineating boundaries, capturing fine details, and improving overall segmentation accuracy [38].

2.2 Frequency-domain Network

Frequency-domain transformations have been widely applied in computer vision tasks, including representation learning [39, 40], image restoration [41, 42], and image generation [43]. These techniques provide the ability to decompose information across different spatial scales, thereby enhancing a model's capability to capture either global or local information.

In the field of image segmentation, frequency-domain transformation techniques have been widely applied due to their effectiveness in denoising and multi-resolution feature extraction. A straightforward approach is to embed frequency-domain operations into specific network components, such as convolutions or Transformers. Lu et al. [44] proposed a dual-tree complex wavelet transform, which, by leveraging approximate shift invariance, multi-directional selectivity, frequency-domain aliasing suppression, and phase preservation, improves CNNs' boundary precision and noise suppression capabilities in images. The Frequency-Guided Feature Network [45] incorporates frequency enhancement into attention modules to achieve multi-scale feature fusion and spatial context modeling, thereby enhancing network generalization. Recent work [46] further employs contourlet decomposition to transfer high-order texture features and spatial structural information, demonstrating superior performance.

An increasing number of methods are now exploring more sophisticated strategies for frequency-space fusion. MIFNet [47] integrates local, global, and frequency features within a single module, building upon the use of non-local filters and multi-scale attention to process spatial and frequency features. The Spatial-Frequency Network [48] incorporates contextual feature dependencies in both the frequency and spatial domains to improve network accuracy. However, these approaches still suffer from fragmented utilization of frequency-domain information or treat frequency representations as a unified feature space. Therefore, a more comprehensive exploration of the integration and exploitation of frequency-domain information remains necessary.

3 Methodology

3.1 Overall Architecture

The overview of the proposed FDM-Net framework is shown in Fig.2(a). FDM-Net is a dual-branch parallel network following an encoder–decoder architecture. It integrates four key modules: Frequency Decomposition Module(FD), 3D Frequency-Domain Adaptive Convolution(FAConv), Frequency-Guided Hybrid Attention(FGHA), and the fusion module. First, FD decomposes the input image into high-frequency and low-frequency subbands, yielding a low-frequency component X^L and high-frequency components X^H grouped radially and directionally. To better learn spectral coefficients, a 3D frequency-adaptive convolution kernel is employed in the shallow encoder to enhance response discrimination. Residual connections are applied separately to process high-frequency and low-frequency features. In the deep encoder, the FGHA module attends to both high-frequency and low-frequency representations, enhancing cross-frequency consistency along spatial, channel, and specific dimensions. The fusion module integrates the outputs of the dual-branch encoders, while deep decoder outputs are skip-connected to shallow layers to facilitate supervised learning.

3.2 Frequency Decomposition Module

As shown in Fig.3, the frequency decomposition module initially decomposes CT or MRI scans into a low-frequency component and multiple direction-sensitive high-frequency subbands. These subbands serve as inputs to the decoder, where structural and textural cues are extracted separately. The core idea is to map the input image to the frequency domain via a 3D Fourier transform and construct differentiable soft-gated masks along diagonal and radial dimensions in the frequency domain. Specifically, after transforming the input to the frequency domain, features are first separated into low- and high-frequency components based on the frequency radius. The high-frequency portion is then divided into two directional subbands along the angular dimension and further decomposed into several radial bands to obtain masks. Each mask is multiplied element-wise with the input spectrum and transformed back to the spatial domain via the inverse fast Fourier transform, achieving multi-directional and multi-scale decomposition. This process can be formulated as:

$$X_f = \mathcal{S}\big(\mathcal{F}\{x\}\big) \tag{1}$$

Here, x denotes the input tensor, F represents the 3D fast Fourier transform, S is the frequency spectrum center shift, and X_f is the transformed power spectrum. Subsequently, a polar coordinate framework is constructed for radial banding and directional decomposition.

$$\mathrm{MLF}(R) = \mathbf{1}[R \le r_{\mathrm{ratio}}] \tag{2}$$

$$\mathrm{MHF}(R) = 1 - \mathrm{MLF}(R) \tag{3}$$

$$X_{\mathrm{LF}} = X_f \cdot \mathrm{MLF}(R), \tag{4}$$

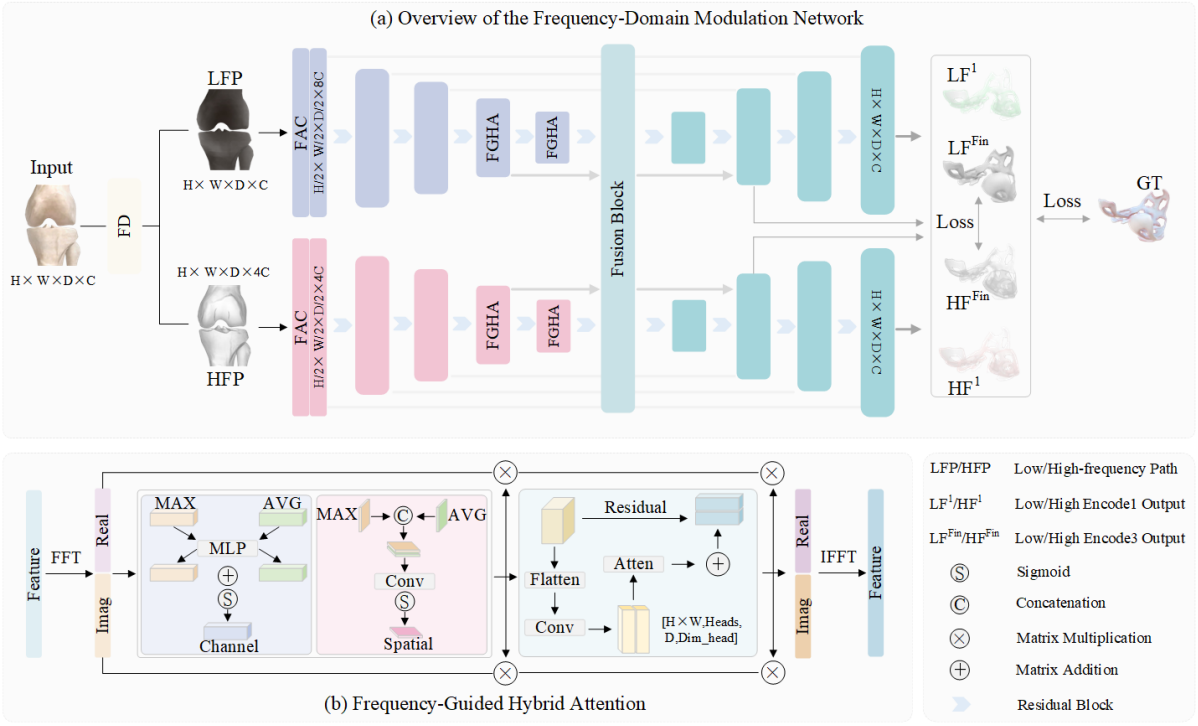

Fig. 2: The overall framework of FDM-Net is shown in (a). It is a dual-branch network for high-frequency and low-frequency features, which decomposes and integrates features through FD, FAConv, and FGHA. Two types of losses are used to align predictions, labels, and the dual branches. (b) FGHA integrates multi-scale information via a cross-attention mechanism, enhancing the processing of features.

$$LF = \mathcal{F}^{-1}\left\{\mathcal{S}^{-1}(X_{\mathrm{LF}})\right\} \tag{5}$$

Here, R denotes the radial radius, r_{ratio} is the radial radius ratio, $\mathrm{MLF}(R)$ is the low-pass mask, X_{LF} is the low-frequency component in the frequency domain, F^{-1} is the 3D inverse fast Fourier transform, and LF is the low-frequency feature transformed back to the spatial domain. $\mathrm{MHF}(R)$ represents the complementary high-frequency set.

$$M_{\mathrm{radial}}(R) = \exp\left(-\frac{(R-r_c)^2}{2\sigma_r^2}\right) \tag{6}$$

$$r_c^{(b)} = \frac{1}{2}\left(r_{b-1} + r_b\right) \tag{7}$$

$$M_r^{(b)}(R) = M_{\mathrm{radial}}^{(b)}(R) \cdot \mathrm{MHF}(R) \tag{8}$$

The above process achieves radial banding. The high-frequency component is divided into b concentric radial bands along the radial direction (in this work, b=2). These bands are evenly spaced within the interval $[r_{\mathrm{ratio}}, 1]$, and smooth transitions are realized using a Gaussian-shaped soft mask M_{radial}. Here, r_c denotes the center radius of a band. $M_r^{(b)}(R)$ represents the radial band mask.

$$d\theta = \min\left|\theta - \theta_i\right| \tag{9}$$

$$M_{\mathrm{orient}}(\theta) = \exp\left(-\frac{d\theta^2}{2\sigma_t^2}\right) \tag{10}$$

$$M_{b,i}(R,\theta) = M_r^{(b)}(R) \cdot M_{\mathrm{orient}}^{(i)}(\theta) \tag{11}$$

The above process achieves directional decomposition. Within each radial band, features are further grouped according to the angular coordinate θ, representing the planar angle. First, i directional centers are generated (in this work, i=2), with $d\theta$ denoting the angular interval. A Gaussian weighting $M_{\mathrm{orient}}(\theta)$ is applied. $M_{b,i}(R,\theta)$ represents the radial–directional decomposition mask, with $i = 1, 2$.

$$X_{(b,i)}^c = X_f^c \cdot M_{b,i}(R,\theta) \tag{12}$$

$$HF^{(c,b,g)} = \mathcal{F}^{-1}\left\{\mathcal{S}^{-1}\left(X_{(b,i)}^c\right)\right\} \tag{13}$$

$$HF = \mathrm{Concat}_{c=1,\,b=1,2,\,i=1,2}\; HF^{(c,b,i)} \tag{14}$$

The mask for each input channel is multiplied element-wise with the frequency spectrum and concatenated along the channel dimension. An inverse FFT and center shift are then applied to obtain the high-frequency feature map HF.

3.3 3D Frequency-Domain Adaptive Convolution

The 3D frequency-domain adaptive convolutional structure is illustrated in Fig.4. FAConv constructs kernel weights by learning spectral coefficients in the frequency domain. These coefficients are divided into non-overlapping groups to separately capture different frequency components . Subsequently, the kernel weight distribution is dynamically adjusted according to the spatial information of the input features, enabling frequency-response adaptation across the entire kernel. Additionally, the responses of the convolutional kernel across different frequency bands are weighted, allowing features at each frequency band to receive varying levels of attention, thereby achieving spatially dependent frequency modulation. The detailed process is shown in Algorithm1.

Algorithm 1 Illustration of the FAConv process

Require: Input feature map X, static kernel set $\{W_1, \ldots, W_K\}$,
Ensure: Output feature map Y_{final}

Step 1: Global feature compression
$X_{\text{gap}} \Leftarrow \text{GAP}(\text{iDFT}(X)) \in \mathbb{R}^{C_{\text{in}} \times 1 \times 1 \times 1}$
$X_1 \Leftarrow \text{Sigmoid}(\text{FC}(X_{\text{gap}})) \in \mathbb{R}^{C_{\text{att}} \times 1 \times 1 \times 1}$
Step 2: Generate dynamic attention coefficients
$A_c, A_s, A_k, A_f \Leftarrow \text{Cross-Attention}(X_1)$
if use Local-Attention **then**
    $A_{\text{local}} \Leftarrow \text{Local-Attention}(X_1)$
else
    $A_{\text{local}} \Leftarrow 1$
end if
$\pi \Leftarrow A_c \odot A_s \odot A_k \odot A_f \odot A_{\text{local}}$
Step 3: Generate dynamic kernel
$W^{'} \Leftarrow \{W_1 \oplus W_2 \oplus \ldots \oplus W_K\} \odot \{\pi_1 \oplus \pi_2 \oplus \ldots \oplus \pi_K\} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k \times k \times k}$
Step 4: Convolution
$Y_{\text{final}} \Leftarrow \text{Sigmoid}(\text{Conv1}(\text{GAP}(\text{DFT}(X)))) * W^{'} \in \mathbb{R}^{C_{\text{out}} \times D \times H \times W}$
Return Y_{final}

Specifically, a set of mutually independent static kernel collections $\{W_1, W_2, \ldots, W_K\}$ is first initialized, where K denotes the number of kernels. The input X is processed through $iDFT$ and global average pooling (GAP) to obtain the compressed feature X_1. X_1 is then fused through channel attention, spatial attention, kernel attention, filter attention, as well as parallel local kernel attention, to generate the final attention coefficients $\{\pi_1, \pi_2, \ldots, \pi_K\}$. The four attention dimensions and the local attention dimension are eventually aligned through broadcasting. The attention coefficients are multiplied with the static kernels to produce the dynamic weights $W^{'}$. X is further processed through DFT, GAP, and convolution, and then multiplied with $W^{'}$ to supplement the features and obtain the final output Y_{final}.

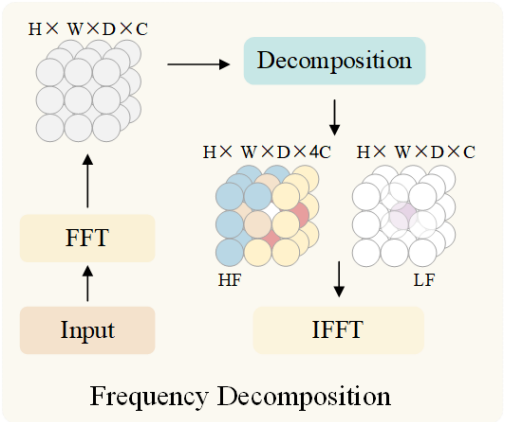

Fig. 3: Structure of the Frequency Decomposition Module. It is used to decompose the input into high-frequency and low-frequency components.

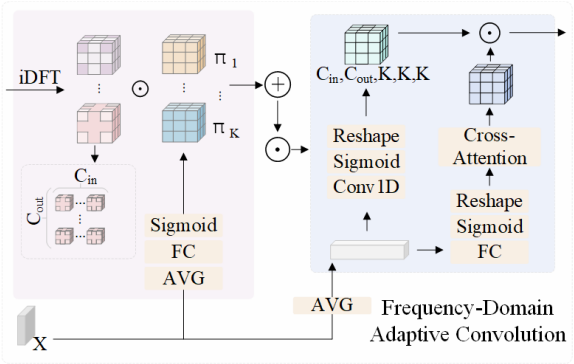

Fig. 4: Structure of the 3D Frequency-Domain Adaptive Convolution. It uses non-overlapping grouped weights to learn high-frequency and low-frequency feature components, consisting of non-overlapping weight parts and a modulation part.

3.4 Frequency-Guided Hybrid Attention and Feature Fusion

As shown in Fig.2(b), the Frequency-Guided Hybrid Attention (FGHA) module integrates and refines deep features extracted by the dual-branch shallow encoder, enhancing the network's ability to model discriminative features. FGHA employs channel, spatial, and axial cross-attention to model features in the frequency domain.

First, a 3D FFT is applied to map the features to the frequency domain, obtaining the real part F_r and imaginary part F_i, facilitating attention extraction from deep features. Cross-attention operations are then performed separately on the real and imaginary parts, producing attention weights F_c, F_s, and F_a. These attention maps are aligned using broadcasting and multiplied with the real part to obtain the adjusted feature map. Finally, the features are transformed back to the spatial domain.

This module integrates low-frequency and high-frequency branch information, complementing specific details. The entire process can be expressed as follows:

$$\hat{F_n} = F(F_n) \tag{15}$$

$$F_{r,i} = \Im(\text{SA}(CA(\hat{F_n}))) \tag{16}$$

$$F_{\text{out}}^{'} = \text{AxialAttnD}(F_r) + i\,\text{AxialAttnD}(F_i) \tag{17}$$

$$F_{\text{out}} = F^{-1}(F_{\text{out}}^{'}) \tag{18}$$

Here, F_n denotes the input feature map, and $\hat{F_n}$ represents the frequency-domain features. F_r and

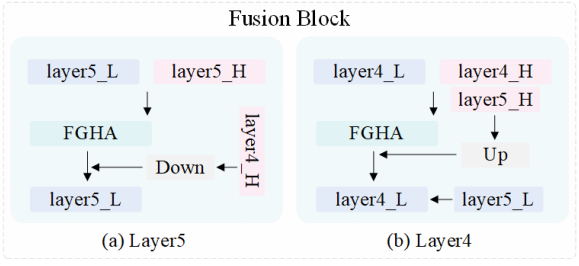

Fig. 5: Illustration of the fusion module. It is used to integrate dual-branch features and multi-scale information.

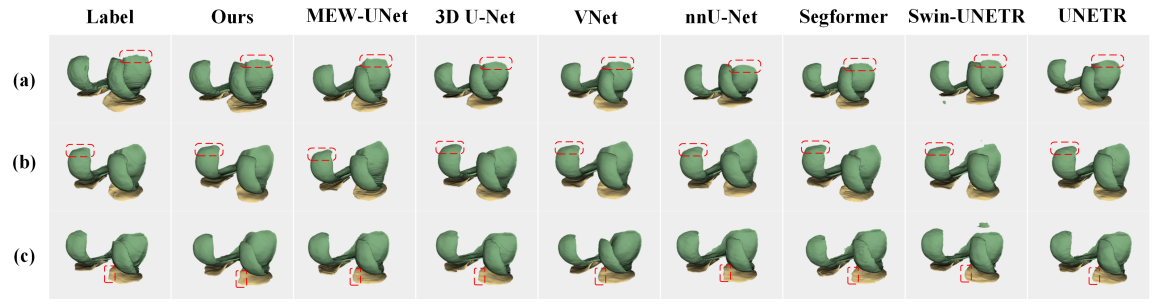

Fig. 6: Visualization of segmentation results from eight methods on representative OAI-ZIB samples. The green and yellow regions represent femoral cartilage (FC) and tibial cartilage (TC), respectively.

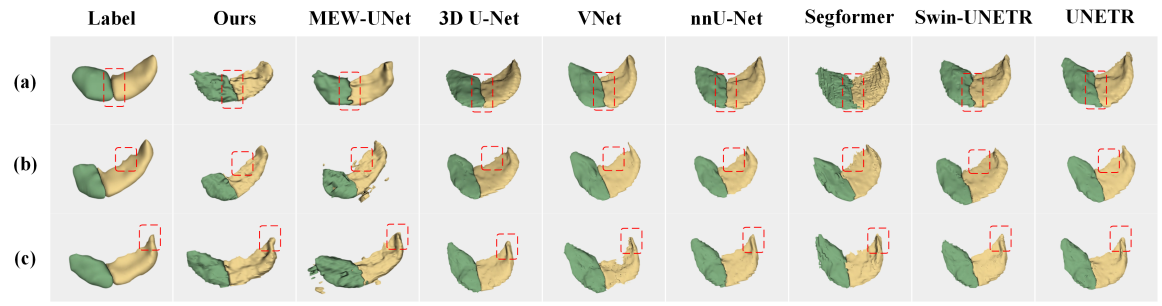

Fig. 7: Visualization of segmentation results on three representative hippocampus samples. The green and yellow regions represent the anterior and posterior parts, respectively.

F_i correspond to the real and imaginary parts in the frequency domain, respectively. CA and SA denote channel attention and spatial attention. $\Im$ represents broadcasting. $AxialAttnD$ refers to axial attention along the D dimension. F'_{out} is the reconstructed complex spectrum from the processed real part $\text{AxialAttnD}(F_r)$ and imaginary part $i\,\text{AxialAttnD}(F_i)$. The 3D inverse Fourier transform (IFFT) restores the features back to the spatial domain, yielding the enhanced feature map F_{out}.

Fig.5 illustrates the operations performed by the feature fusion module, which is designed to integrate information from the dual branches to ensure robustness and mutual enhancement between them. Taking the low-frequency branch as an example, the high-frequency feature $H5$ and the low-frequency feature $L5$ are the outputs of the fourth layer of the decoder. After passing through the FGHA module, they are concatenated and further processed by subsequent convolutional blocks to extract features. Meanwhile, the fourth-layer output of the high-frequency branch is downsampled and added in to supplement fine details, producing the final fifth-layer output of the low-frequency branch. The feature fusion process for the fourth layer of the low-frequency branch is similar, but it also incorporates the final output of the fifth layer to enrich the features.

3.5 Loss Function

The overall loss function (Eq.19) used by the model can be summarized as:

$$L_{\text{Total}} = L_{\text{Cons}} + \lambda L_{\text{BA}} \tag{19}$$

Here, L_{Cons} represents the consistency constraint loss, which is used to align the outputs of the low-frequency and high-frequency branches. L_{BA} denotes the branch alignment loss, ensuring the consistency of predictions between the two branches, and λ represents the weighting factor.

$$L_{\text{Cons}} = \sum_{f\in\{L,H\}} \sum_{t\in\{\text{fin,side}\}} L^{f}_{\text{Cons}}\left(\hat{y}^{f_t}, y\right) \tag{20}$$

$$L_{Dice} = 1 - \frac{2}{N}\sum_{i=1}^{N} \frac{\sum_{i=1}^{I} g_i p_i + \epsilon}{\sum_{i=1}^{I} g_i + \sum_{i=1}^{I} p_i + \epsilon} \tag{21}$$

The L_{Cons} loss includes both low- and high-frequency losses, computed over the dual branches as well as the auxiliary branch, and uses Dice loss (Eq.21). Here, y represents the ground truth labels, y_{ft} represents the predicted outputs, L and H denote the low-frequency and high-frequency branches, and fin and mid denote the intermediate layer outputs and the final outputs of the dual branches. The L_{BA} loss employs 3D DFL, as shown in Eq.22, and uses the final outputs of the dual branches as pseudo-labels for each other, enabling mutual supervision. Here, $\hat{y}^L$ and $\hat{y}^H$ denote the main outputs of the LF and HF branches of the

model, respectively.

$$L_{\mathrm{BA}} = L_{\mathrm{BA}}^{\mathrm{FF}}\left(\hat{y}^{L}, \hat{y}^{H}\right) \tag{22}$$

$$W(u) = \mathrm{Norm}\left(\|F(\hat{y}^{(1)})(u) - F(\hat{y}^{(2)})(u)\|_2^{\alpha}\right) \tag{23}$$

$$L_{\mathrm{BA}} = \lambda_u \cdot \frac{1}{|\Omega|} \sum_{u \in \Omega} W(u)\, \|F(\hat{y}^{(1)})(u) - F(\hat{y}^{(2)})(u)\|_2^2 \tag{24}$$

The final formulations of 3D DFL and L_{BA} are presented in Eq.3.5 and Eq.24, where $F(\cdot)$ represents the 3D Fourier transform, $u \in \Omega$ denotes the frequency coordinates, and Ω represents the set of all positions in the frequency domain. $\|F(\hat{y}^{(1)})(u) - F(\hat{y}^{(2)})(u)\|_2^{\alpha}$ is the difference between two predicted points at frequency u, defined as the squared Euclidean distance. $W(u)$ is the frequency weighting matrix, where $\alpha > 0$ is the focusing factor and $\mathrm{Norm}(\cdot)$ denotes normalization by the maximum value. $|\Omega|$ is the total number of positions in the frequency domain, serving as the denominator for summation normalization.

4 Experimental

Datasets. We quantitatively evaluated our method using publicly available datasets acquired with different imaging modalities: OAI-ZIB, Hippocampus, and Pancreas.

OAI-ZIB is a knee cartilage dataset with expert manual annotations of femoral and tibial cartilage. The dataset consists of MRI scans of 507 subjects, with 253 used for training and 254 for testing.

The Pancreas dataset contains 420 three-dimensional CT scans (281 for training and 139 for testing). A major challenge lies in the imbalanced label distribution, which involves large-, medium-, and small-scale structures.

The hippocampus dataset aims to accurately segment two adjacent small structures from MRI scans. It contains 394 MRI scans, of which 263 are used for training and 131 for testing.

Evaluation metrics. To objectively assess the performance of different models, we used the Dice similarity coefficient (DSC) and the average symmetric surface distance (ASSD) as evaluation metrics.

Implementation details. Our proposed FDM-Net was implemented in PyTorch 1.10.1 with CUDA 11.8 and trained and tested on an NVIDIA GeForce RTX 4090 GPU . The network was trained for 450 epochs using the SGD optimizer with a momentum of 0.9, an initial learning rate of 0.3, a weight decay of 5×10^{-5}, and a batch size of 1. To improve generalization and prevent overfitting, several data augmentation strategies were applied during training, including random rotations, random flipping, and random cropping.

4.1 Experimental Results.

4.1.1 Quantitative Results

To evaluate the segmentation performance of FDM-Net on medical images, we compared it with seven 3D segmentation networks. The segmentation results of different models on the OAI-ZIB dataset are shown in Table.1. Quantitative results indicate that our method, FDM-Net, outperforms the comparisons. On OAI-ZIB, FDM-Net achieves Dice scores of 90.12% and 89.73% for Femoral Cartilage and Tibial Cartilage, respectively, with corresponding ASSD values of 0.5023 mm and 0.4753 mm. Compared with 3D MEW-UNet, the Dice score for Tibial Cartilage increased by 3.85%, and ASSD improved by 0.025 mm. This demonstrates more accurate predictions of the cartilage regions and improved boundary delineation. This is crucial for the clinical assessment of knee osteoarthritis, as precise measurement of cartilage thickness and volume is a key parameter for monitoring disease progression.

To evaluate the generalization ability of FDM-Net across different organs, we conducted experiments on the hippocampus dataset. The comparison results are shown in Table.2. For the Posterior region, our method has a slightly lower DSC than nnU-Net, possibly due to the small sample size, but achieves the best ASSD. Compared with other methods, the Dice score also exceeds that of 3D MEW-UNet, which also uses frequency-domain information, by 5.84%.

On the pancreas dataset, the performance of FDM-Net is shown in Table.3. The tumor region achieved a Dice score of 70.69% and an ASSD of 17.15 mm, representing a significant improvement over the comparison networks. The notable increase in DSC may be attributed to the low-frequency branch focusing on global information. However, segmentation performance for the Pancreas region remains moderate. It is worth emphasizing that FDM-Net achieves these results with extremely low computational cost, requiring only 2.4M parameters and 0.68 GFLOPs, which is several times less than other methods.

4.1.2 Qualitative Results

Fig.6 provides a qualitative visualization of the segmentation performance of the models on selected samples from the OAI-ZIB dataset. Each row corresponds to one sample, with red dashed boxes highlighting regions of segmentation discrepancy. Consistent with the quantitative metrics, FDM-Net shows improvements in capturing contour details and achieving precise segmentation. As shown in Fig.6(c), our method is more sensitive to irregular edges. In Fig.6(b), CNN-based UNet and VNet exhibit over-smoothing at the boundaries, performing worse than the Transformer-based methods and the well-optimized nnU-Net. In contrast, the frequency-domain-based methods (Ours and MEW-UNet) produce segmentation results that

	Femoral Cartilage		Tibial Cartilage		Average	
Methods	DSC(%)	ASSD(mm)	DSC(%)	ASSD(mm)	DSC(%)	ASSD(mm)
V-Net	87.71	1.1520	84.11	1.7328	85.91	1.4424
3D U-Net	89.03	0.7766	84.77	1.0323	86.90	0.9045
nnU-Net	89.05	0.6825	85.53	0.8340	87.29	0.7583
SegFormer	85.55	1.8193	83.79	3.2474	84.67	2.5334
UNETR	88.34	0.6962	85.27	0.9987	86.80	0.8475
Swin-UNETR	89.27	0.6274	85.30	0.8436	87.28	0.7355
3D MEW-UNet	89.11	0.5492	85.88	0.5003	87.50	0.5248
Ours	**90.12**	**0.5023**	**89.73**	**0.4753**	**89.93**	**0.4888**

Table 1: Results of OAI-ZIB datasets. The best results are highlighted in bold.

	Anterior		Posterior		Average	
Methods	DSC(%)	ASSD(mm)	DSC(%)	ASSD(mm)	DSC(%)	ASSD(mm)
V-Net	80.02	6.5452	76.72	3.6717	78.37	5.1084
3D U-Net	88.75	1.4018	87.33	1.3113	88.04	1.3566
nnU-Net	90.00	1.4973	**89.00**	1.5273	**89.50**	1.5123
SegFormer	85.97	2.9538	84.98	2.0609	85.48	2.5074
UNETR	88.79	1.5799	87.10	1.4908	87.95	1.5354
Swin-UNETR	89.31	1.4324	87.83	1.4866	88.57	1.4595
3D MEW-UNet	83.48	1.7545	82.45	1.5914	82.97	1.6730
Ours	**90.23**	**1.2332**	88.29	**1.3073**	89.26	**1.2703**

Table 2: Results of Hippocampus datasets. The best results are highlighted in bold.

more closely match the label shapes, although MEW-UNet's excessive emphasis on fine features leads to slight distortions in certain regions. Overall, our model better preserves contours and aligns more closely with the label shapes. Fig.7(b) shows segmentation visualizations on a specific sample from the Hippocampus dataset. Notably, in Fig.7(B) and (C), MEW-UNet produces false positives in some regions, while our method achieves accurate segmentation. Considering the overall results, FDM-Net provides a more complete delineation.

To further investigate the reasons behind the improved segmentation performance of our method, we conducted a comparative analysis of Grad-CAM attention maps of FDM-Net and 3D MEW-UNet across multiple samples. As shown in Fig.8. FDM-Net exhibits more concentrated high-response regions on the target structures, demonstrating stronger localization of key structures. The third row of Fig.8 partly illustrates the significant improvement of FDM-Net in tumor region segmentation. The region is shown in the deepest red, indicating that the model focuses most on the target. Qualitative analysis shows that FDM-Net excels in capturing boundaries and fine details of key structures.

4.2 Ablation experiment

To evaluate the contribution of several modules in the network, we conducted ablation experiments on the Hippocampus dataset. Table.4 presents the results of the ablation experiments. In these experiments, we used Dice and HD95 as evaluation metrics.

Effectiveness of the FD module. As shown in Table.4, introducing the FD module achieved excellent baseline performance. In the Anterior and Posterior regions, the Dice scores were 88.66% and 86.69%, while the HD95 values were 1.3661 mm and 1.5311 mm, respectively, highlighting the effectiveness of the dual high-frequency and low-frequency branches.

Effectiveness of the FD and FAConv modules. After adding the FAConv module, segmentation performance was further improved thanks to adaptive learning of high-frequency and low-frequency features. The HD95 score in the Posterior region reached the optimal 1.5116 mm. Dice scores also improved by 0.58% and 0.9%. These results indicate that the FAConv module not only effectively reduces uncertainty in boundary prediction but also enhances the model's discriminative ability in complex regions, making the segmentation results more precise and stable.

Effectiveness of the FD and FGHA modules. When combined with FGHA alone, the performance improvement was limited, although Dice scores still outperformed some comparative methods. This may be because, without FAConv's dynamic feature modeling, FGHA struggles to obtain reasonable weight distributions, and its collaborative modeling advantage is not fully realized. Notably, the introduction of the FGHA module reduced HD95 in the Anterior region, indicating its positive role in improving segmentation boundary accuracy.

Effectiveness of the FD, FAConv, and FGHA modules. The full FDM-Net achieved the highest Dice scores, with 90.23% in the Anterior

	Pancreas		Tumor		Cost	
Methods	DSC(%)	ASSD(mm)	DSC(%)	ASSD(mm)	Params(M)	FLOPs(G)
V-Net	74.81	**2.1520**	50.11	18.7328	45.60	96.00
3D U-Net	75.04	2.7766	51.77	18.2523	21.03	30.20
nnU-Net	**76.75**	2.6825	52.86	18.0740	11.20	250.95
Swin-UNETR	76.54	2.4401	52.65	18.59	63.58	333.64
3D MEW-UNet	75.14	2.2949	51.65	18.6174	124.28	81.34
Ours	74.61	2.9132	**53.96**	**17.15**	**2.40**	**0.68**

Table 3: Results of Pancreas datasets. The best results are highlighted in bold.

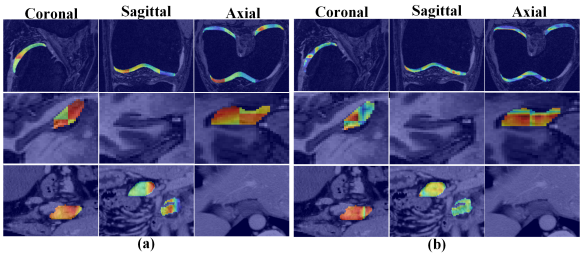

Fig. 8: Grad-CAM visualizations of FDM-Net and 3D MEW-UNet on samples from three datasets. The first row corresponds to OAI-ZIB, the second to hippocampus, and the third to pancreas. (a) FDM-Net, (b) MEW-UNet. Red indicates high attention regions, while blue indicates low attention regions.

	Anterior		Posterior	
	DSC(%)	HD95(mm)	DSC(%)	HD95(mm)
FD	88.66	1.3661	86.69	1.5311
FD + FAConv	89.24	1.3214	87.59	**1.5116**
FD + FGHA	88.56	1.3292	86.68	1.5776
FDM-Net	**90.23**	**1.2258**	**88.29**	1.5305

Table 4: Ablation study on the Pancreas datasets.

region and 88.29% in the Posterior region. Experimental results demonstrate that each component contributes to improving segmentation accuracy and exhibits complementary effects.

To further investigate the contribution of each module to the feature representation of the target regions, we visualized the features using t-SNE (Fig.9). When only the FD module was introduced, the intra-class points were relatively scattered, indicating weak feature cohesion. After adding the FGHA module, the intra-class distribution did not change significantly, consistent with the quantitative metrics. When combining FD and FAConv, the intra-class points became more compact, showing higher cohesion. In the complete FDM-Net (FD, FAConv, and FGHA), the two clusters within each class almost completely overlapped, forming more regular and dense clusters. This indicates that the global structural encoding was enhanced.

5 Discussion

Conclusion. In this work, we propose FDM-Net, a frequency-domain-based dual-branch segmentation framework. The method achieves a good balance between computational cost and segmentation performance. It also explores introducing frequency-domain modeling into medical image segmentation, providing a feasible alternative to traditional spatial-domain approaches. By employing frequency-domain decomposition and multi-scale feature integration, the model's ability to capture both fine details and global features is enhanced.

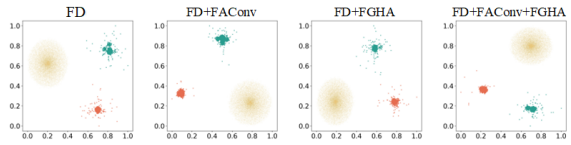

Fig. 9: t-SNE visualization of feature representations for the same data under different module combinations. The background is shown in yellow, anterior in green, and posterior in orange.

Limitation and Future Work. Our proposed method has certain limitations. Although it achieves improvements on several segmentation tasks by decomposing and integrating high-frequency and low-frequency features, the robustness of segmentation for some targets remains challenging. Future research could further explore improvement strategies, such as adjusting the frequency decomposition approach or enhancing the multi-scale feature fusion module, to improve the model's robustness and applicability.

References

[1] Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[2] Chen, Z., Xu, Q., Liu, X., Yuan, Y.: Un-sam: Domain-adaptive self-prompt segmentation for universal nuclei images. Medical Image Analysis, 103607 (2025)

[3] Alirr, O.I., Abd Rahni, A.A.: An automated liver vasculature segmentation from ct scans for hepatic surgical planning. International Journal of Integrated Engineering **13**(1), 188–200 (2021)

[4] Chen, Z., Guo, X., Woo, P.Y., Yuan, Y.: Super-resolution enhanced medical image diagnosis with sample affinity interaction. IEEE Transactions on Medical Imaging **40**(5), 1377–1389 (2021)

[5] Chen, Z., Yang, C., Zhu, M., Peng, Z., Yuan, Y.: Personalized retrogress-resilient federated learning toward imbalanced medical data. IEEE Transactions on Medical Imaging **41**(12), 3663–3674 (2022)

[6] Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K.: Medical image analysis using convolutional neural networks: a review. Journal of medical systems **42**(11), 226 (2018)

[7] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing

and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[8] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[10] Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016). Ieee

[11] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

[12] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[13] Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[15] Dai, D., Dong, C., Yan, Q., Sun, Y., Zhang, C., Li, Z., Xu, S.: I2u-net: A dual-path u-net with rich information interaction for medical image segmentation. Medical Image Analysis **97**, 103241 (2024)

[16] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[17] Liu, J., Zhang, D., He, L., Yu, X., Han, W.: Mfagnet: Multi-scale frequency attention gating network for land cover classification. International Journal of Remote Sensing **44**(21), 6670–6697 (2023)

[18] Yang, C., Zhang, Z.: Pfd-net: Pyramid fourier deformable network for medical image segmentation. Computers in Biology and Medicine **172**, 108302 (2024)

[19] Li, X., Xu, F., Yong, X., Chen, D., Xia, R., Ye, B., Gao, H., Chen, Z., Lyu, X.: Sscnet: A spectrum-space collaborative network for semantic segmentation of remote sensing images. Remote Sensing **15**(23), 5610 (2023)

[20] Li, P., Zhou, R., He, J., Zhao, S., Tian, Y.: A global-frequency-domain network for medical image segmentation. Computers in Biology and Medicine **164**, 107290 (2023)

[21] Yang, Y., Yuan, G., Li, J.: Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. IEEE Transactions on Geoscience and Remote Sensing (2024)

[22] Lu, H., Wang, H., Zhang, Q., Won, D., Yoon, S.W.: A dual-tree complex wavelet transform based convolutional neural network for human thyroid medical image segmentation. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 191–198 (2018). IEEE

[23] Wang, Z., Li, X., Duan, H., Su, Y., Zhang, X., Guan, X.: Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform. Expert Systems with Applications **171**, 114574 (2021)

[24] Ambellan, F., Tack, A., Ehlke, M., Zachow, S.: Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. Medical Image Analysis **52**, 109–118 (2019)

[25] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., *et al.*: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)

[26] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

[27] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)

[28] Jing, J., Wang, Z., Rätsch, M., Zhang, H.: Mobile-unet: An efficient convolutional neural network for fabric defect detection. Textile Research Journal **92**(1-2), 30–42 (2022)

[29] Koonce, B.: Efficientnet. In: Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization, pp. 109–123. Springer, ??? (2021)

[30] Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y., Chen, G.: Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quantitative imaging in medicine and surgery **10**(6), 1275 (2020)

[31] Zhu, Z., Yan, Y., Xu, R., Zi, Y., Wang, J.: Attention-unet: A deep learning approach for fast and accurate segmentation in medical imaging. Journal of Computer Science and Software Applications **2**(4), 24–31 (2022)

[32] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[33] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). Springer

[34] Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 36–46 (2021). Springer

[35] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)

[36] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[37] Kuang, H., Wang, Y., Liu, J., Wang, J., Cao, Q., Hu, B., Qiu, W., Wang, J.: Hybrid cnn-transformer network with circular feature interaction for acute ischemic stroke lesion segmentation on non-contrast ct scans. IEEE Transactions on Medical Imaging **43**(6), 2303–2316 (2024)

[38] Gao, F., Wang, X., Gao, Y., Dong, J., Wang, S.: Sea ice change detection in sar images based on convolutional-wavelet neural networks. IEEE Geoscience and Remote Sensing Letters **16**(8), 1240–1244 (2019)

[39] Fu, D., Liu, J., Zhong, H., Zhang, X., Zhang, F.: A novel self-supervised representation learning framework based on time-frequency alignment and interaction for mechanical fault diagnosis. Knowledge-Based Systems **295**, 111846 (2024)

[40] Zhao, W., Fan, L.: Time-series representation learning via time-frequency fusion contrasting. Frontiers in Artificial Intelligence **7**, 1414352 (2024)

[41] Ding, H., Huang, Y., Chen, N., Lu, J., Li, S.: Image inpainting for periodic discrete density defects via frequency analysis and an adaptive transformer-gan network. Applied Soft Computing **167**, 112410 (2024)

[42] Cao, Y., Ma, R., Zhao, K., An, P.: Wfil-net: image inpainting based on wavelet downsampling and frequency integrated learning module. Multimedia Systems **31**(1), 21 (2025)

[43] Qian, Y., Cai, Q., Pan, Y., Li, Y., Yao, T., Sun, Q., Mei, T.: Boosting diffusion models with moving average sampling in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8911–8920 (2024)

[44] Lu, H., Wang, H., Zhang, Q., Won, D., Yoon, S.W.: A dual-tree complex wavelet transform based convolutional neural network for human thyroid medical image segmentation. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 191–198 (2018). IEEE

[45] Li, X., Xu, F., Gao, H., Liu, F., Lyu, X.: A frequency domain feature-guided network for semantic segmentation of remote sensing images. IEEE Signal Processing Letters **31**, 1369–1373 (2024)

[46] Ji, D., Zhao, F., Lu, H., Wu, F., Ye, J.: Structural and statistical texture knowledge distillation and learning for segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(5), 3639–3656 (2025)

[47] Fan, J., Li, J., Liu, Y., Zhang, F.: Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation. Engineering Applications of Artificial Intelligence **129**, 107638 (2024)

[48] Zhang, T., Dick, R.P.: Spatial-frequency network for segmentation of remote sensing images. In: 2023 IEEE International Conference on Image Processing (ICIP), pp. 3553–3557 (2023)