ITC 4/54
Information Technology and Control
Vol. 54 / No. 4/ 2025
pp. 1271-1287
DOI 10.5755/j01.itc.54.4.41046

FSA-Net: Frequency-Spatial Attention Network for Medical Image Segmentation

Received 2025/04/01 | Accepted after revision 2025/05/30

HOW TO CITE: Zhao, T. (2025). FSA-Net: Frequency-Spatial Attention Network for Medical Image Segmentation. *Information Technology and Control*, 54(4), 1271-1287. https://doi.org/10.5755/j01.itc.54.4.41046

FSA-Net: Frequency-Spatial Attention Network for Medical Image Segmentation

Tianyi Zhao
School of Software Engineering, Shandong University, Shandong 250000, China

Corresponding author: tianyiZhao@mail.sdu.edu.cn

Medical image segmentation, a core technology in computer-aided diagnosis, faces persistent challenges including high-frequency detail loss, inadequate multi-scale modeling, and limited cross-modal generalization. This study proposes Frequency-Spatial Attention Network (FSA-Net) for Medical Image Segmentation, a novel framework integrating frequency-spatial dual-path attention with adaptive multi-scale Transformer architecture. The framework features two key innovations: (1) A Frequency-Spatial Adaptive Selection (FSAS) module that decouples high-frequency edges from low-frequency structures during feature embedding, effectively preserving critical boundary information typically attenuated by conventional downsampling. (2) A wavelet decomposition module combined with window-based attention mechanisms, enabling simultaneous modeling of long-range spatial dependencies and channel-wise semantic correlations within Transformer blocks. Extensive experiments demonstrate FSA-Net's superiority, achieving 79.36% mean Dice and 19.71mm HD95 on Synapse, outperforming Swin-Unet. The framework pioneers dynamic frequency-domain filtering combined with wavelet-guided cross-scale attention, establishing a novel paradigm for frequency-space synergy. It also attains 89.72% Dice for myocardium segmentation on the ACDC dataset, validating its efficacy in low-contrast scenarios. By quantitatively verifying the critical role of high-frequency components in boundary reconstruction and enabling efficient learning on small datasets, this work advances clinical precision in complex anatomical segmentation while providing scalable technical foundations for multi-modal applications.

KEYWORDS: Medical Image Segmentation, Attention Mechanism, Multi-Scale Transformer, Frequency Learning

1. Introduction

Medical image segmentation has undergone significant evolution alongside advancements in imaging technology and growing clinical demands, transitioning from traditional threshold-based methods to sophisticated deep learning frameworks. Central to this progress lies the ongoing quest to balance local feature extraction and global context modelling — a tension embodied in the complementary strengths of convolutional neural networks (CNNs) and Vision Transformers (ViTs). While U-Net's encoder-decoder architecture revolutionized medical segmentation through hierarchical feature learning and skip connections, its CNN-based design struggles with long-range dependencies and spectral aliasing during downsampling. Even advanced self-configuring frameworks like nnU-Net [14], which automate architecture optimization for diverse tasks, fail to mitigate high-frequency signal loss caused by pooling operations. These limitations manifest in challenges like correlating anatomically linked organs (e. g., pancreas-duodenum relationships in abdominal CT) and preserving microstructural boundaries (e. g. hepatic vessel networks), where fixed convolutional kernels fail to capture broader spatial contexts and pooling operations irreversibly attenuate high-frequency signals.

The emergence of ViTs initially promised solutions through global self-attention mechanisms, yet introduced new hurdles. Though effective in maintaining structural consistency for larger anatomical regions, ViTs' patch embedding process inadvertently suppresses high-frequency details critical for edge precision—evidenced by their 12-15% deficit in microstructure recall compared to CNNs for tasks like Glisson sheath segmentation [23]. SegFormer3D [24] pioneers a 3D hierarchical transformer, demonstrating that compact designs can achieve competitive accuracy on Synapse and ACDC datasets. Similarly, G-CASCADE [26] introduces graph convolutional decoding to maintain long-range dependencies while reducing decoder FLOPs by 82.3%, validating graph operations as an efficient alternative to traditional convolutions. The EMCAD framework [27] further advances efficient decoding through multi-scale depth-wise convolutions and grouped attention gates, attaining SOTA performance with 79.4% parameter reduction. Furthermore, their heavy reliance on large-scale pretraining conflicts with medical imaging's data scarcity realities, where datasets like the 30-scan Synapse [17] benchmark necessitate models capable of learning efficiently from limited samples. This tension between local specificity and global awareness spurred hybrid architectures seeking synergistic integration. TransUNet bridged CNN-localized feature extraction with ViT-driven global aggregation, achieving 78.6% pancreatic segmentation accuracy, while Swin-Unet's hierarchical window attention balanced computational efficiency with multi-scale modeling [21]. In recent years, State Space Models (SSMs) have shown remarkable potential in sequence modeling, with Mamba achieving linear-complexity long-sequence modeling through a selective state space mechanism. Leveraging this framework, UNetMamba [24] integrates a lightweight segmentation decoder and local supervision module to enhance efficiency and accuracy in high-resolution remote sensing imagery; SMM-UNet [18] introduces dynamic multi-scale fusion with ultra-low parameters for precise segmentation of morphologically complex medical lesions; Link Aggregation Mamba [34] combines Mamba's cross-scale feature aggregation and Transformer's global semantic modeling, demonstrating complementary advantages of SSMs and attention mechanisms in hierarchical feature fusion for long-range dependency capture and local detail reconstruction in remote sensing scenarios. Meanwhile, multi-head attention networks with feature transfer mechanisms have demonstrated significant potential in medical image segmentation. For instance, the CM-TranCaF [8] framework integrates cross-modality transfer learning with attention gates to achieve feature alignment and complementary information fusion across modalities. Nevertheless, these approaches still grappled with cross-organ frequency correlations and low-contrast boundary sensitivity, particularly in complex abdominal regions where overlapping tissue textures demand nuanced spectral analysis.

Addressing these dual challenges of spectral preservation and contextual modeling, our proposed

Frequency-Spatial Attention Network (FSA-Net) framework introduces a frequency-spatial collaborative architecture. At its core lies the FSAS module, which reimagines feature downsampling through dynamic frequency masking. By applying Fast Fourier Transform to decompose features into spectral components, FSAS employs learnable filters to selectively enhance high-frequency edge signals and low-frequency anatomical structures prior to spatial subsampling — effectively mitigating traditional CNN's spectral aliasing while preserving ViT-style global awareness. Building upon this foundation, a Haar wavelet-enhanced multi-scale attention mechanism cascades spectral decomposition with windowed attention operations, using low-frequency subbands to guide cross-organ semantic relationships while retaining high-frequency texture details. This dual-domain synergy translates to measurable performance gains, evidenced by Dice coefficient improvements of 1.5% (left kidney), 4.7% (right kidney), and 3.1% (pancreas) on the Synapse dataset, alongside 30% reduction in liver-gallbladder boundary Hausdorff distances. By harmonizing frequency-aware feature preservation with adaptive spatial modeling, FSA-Net advances towards clinically viable segmentation where pixel-level precision meets whole-volume contextual understanding — a critical step in bridging the gap between computational innovation and bedside application.

The contributions of this study are summarized as follows:

1. A novel "frequency-domain filtering and multi-scale enhancement" collaborative architecture is proposed. By integrating FSAS and Haar wavelets, the framework reduces reliance on Swin hierarchical structures, outperform some baseline methods in specific scenarios
2. Extensive evaluations on multiple medical image segmentation datasets demonstrate that FSA-Net consistently outperforms baseline models, achieving competitive performance results.
3. The study validates the effectiveness of frequency-domain attention in multi-organ abdominal CT segmentation and highlights the critical role of high-frequency components in the recovery of small-target boundaries, providing a theoretical foundation for future research.

2. Related Work

2.1. CNN-based Medical Image Segmentation

The U-Net [28] architecture established a paradigm for medical segmentation through its symmetric encoder-decoder structure and skip connections, enabling precise localization via multi-scale feature fusion. Subsequent variants expanded this foundation with distinct technical emphases: UNet++ [33] introduced nested dense skip pathways to bridge semantic gaps between encoder and decoder layers, thereby enhancing feature reuse across resolutions. In contrast, Attention U-Net [23] incorporated spatial attention gates to dynamically weight feature maps, prioritizing anatomically salient regions such as tumor boundaries. The work of Zhou et al. [32] further cascaded attention layers for pancreatic cancer segmentation, yet their fixed convolutional design limits adaptability to multi-organ interactions in complex abdominal regions. ResUNet-a [9] further diverged by integrating residual blocks and dilated convolutions, specifically targeting low-contrast scenarios through multi-scale context aggregation. For 3D volumetric analysis, DenseVNet [13] adopted dense connections and anisotropic kernels, demonstrating adaptability to heterogeneous imaging modalities like MRI and CT. These innovations collectively underscore CNNs' versatility in balancing computational efficiency with localized feature extraction, particularly for organs requiring pixel-level precision.

2.2. Evolution of Transformers and Hybrid Architectures

Vision Transformers (ViTs) [16] redefined global context modeling through patch-based self-attention, yet their computational intensity and spectral insensitivity prompted domain-specific adaptations. Swin Transformer [21] addressed efficiency constraints via hierarchical window attention, enabling multi-scale feature learning while reducing memory overhead — a critical advantage for high-resolution medical data. Extensions like Swin UNet++ [20] introduced dense skip connections for hierarchical fusion, yet their rigid window partitioning struggles with irregular organ morphologies (e. g., pancreatic tails). Hybrid architectures pursued divergent integration strategies: TransUNet

[5] embedded ViT blocks into U-Net's bottleneck to enhance global organ coherence, whereas UNETR [1] adopted a pure Transformer encoder for 3D volumetric consistency. ATFormer [7] further optimized training stability through post-layer normalization, mitigating gradient issues in low-data regimes. The Focal Transformer [19] attempted to balance local-global interactions via sparse attention windows, but its computational overhead hinders deployment in high-resolution CT volumes. These approaches highlight a spectrum of design philosophies, from tightly coupled CNN-Transformer hybrids to purely attention-driven architectures, each tailoring global-local trade-offs to specific anatomical challenges.

2.3. Integration of Wavelet Transforms and Deep Learning

Wavelet-enhanced frameworks address spectral aliasing and multi-scale ambiguity through diverse decomposition strategies. DWT-Unet [22] replaced conventional downsampling with discrete wavelet transforms (DWT), preserving high-frequency edges while reducing spatial redundancy — a design particularly effective for pancreatic duct segmentation. Huang et al. advanced this concept by introducing learnable wavelet kernels, enabling adaptive frequency-band selection during feature extraction. WaveFormer [40] diverged by cascading wavelet decomposition with self-attention, using high-frequency subbands to sharpen boundary predictions in retinal OCT images. Meanwhile, DWT-TransUNet [4] combined Haar wavelets with Transformer blocks, explicitly modeling cross-organ dependencies through low-frequency anatomical priors. These methods exemplify the growing synergy between multi-resolution analysis and deep learning, with architectural variations reflecting task-specific priorities in edge preservation versus structural coherence.

2.4. Medical Applications of Frequency-Domain Attention

Frequency-domain mechanisms complement spatial methods by emphasizing spectral discriminability. FcaNet [25] pioneered frequency-channel attention, dynamically weighting Fourier-transformed features to amplify high-frequency signals — a strategy later adapted by Zhang et al. for pancreatic tumor segmentation [35]. FFTformer [31] took an orthogonal approach, using Fast Fourier Transforms (FFT) to inject global frequency context into spatial attention maps, improving consistency in whole-organ segmentation. Domain-specific innovations include Guo et al.'s frequency-graph fusion [11], which coupled spectral features with graph convolutions to resolve abdominal adhesions, and Wang et al.'s STFT-based boundary refinement [12], leveraging short-time Fourier transforms to reconstruct microvascular networks in brain MRI. These methodologies collectively demonstrate frequency-domain attention's versatility, with architectural choices — such as static versus dynamic frequency masking — tailored to specific imaging modalities and anatomical complexities.

3. Methods

3.1. FSA-Net Architecture

FSA-Net is a U-shaped neural network architecture designed to address the loss of high-frequency details and insufficient cross-organ semantic modeling in medical image segmentation. Its core comprises two innovative modules, the first one is FSAS module, features are mapped to the frequency domain via Fast Fourier Transform (FFT), where learnable binary masks dynamically separate high-frequency (organ edges) and low-frequency (anatomical structures) components. Spatial features are reconstructed through inverse FFT, while parallel depthwise separable convolutions extract local texture information. Dual-path features are concatenated and dimensionally reduced to preserve high-frequency details. In addition, Haar wavelet-enhanced multi-scale window attention module (DWT-SwinTransformerBlock), Haar wavelet decomposition captures low-frequency subbands (LL) for global organ topology modeling via window self-attention. Original features and LL subbands are concatenated and enhanced by linear layers, achieving explicit separation of organ structures and edge textures.

As illustrated in Figure 1, the framework consists of four components:

1 Encoder

A four-stage hierarchical encoder based on Swin Transformer blocks. Each stage employs patch merging for downsampling (reducing resolution by 2× and doubling channels) followed by two consecutive DWT-SwinTransformer Blocks. Before downsampling, the proposed FSAS module dynamically filters high- and low-frequency components via FFT-based masking, mitigating spectral aliasing. Residual connections are added to alleviate gradient vanishing.

2 Bottleneck Layer

Two DWT-SwinTransformer Blocks process the deepest features, enhanced by the Haar wavelet decomposition module for multi-scale attention fusion. Low-frequency subbands guide cross-organ semantic modeling via window self-attention.

3 Decoder

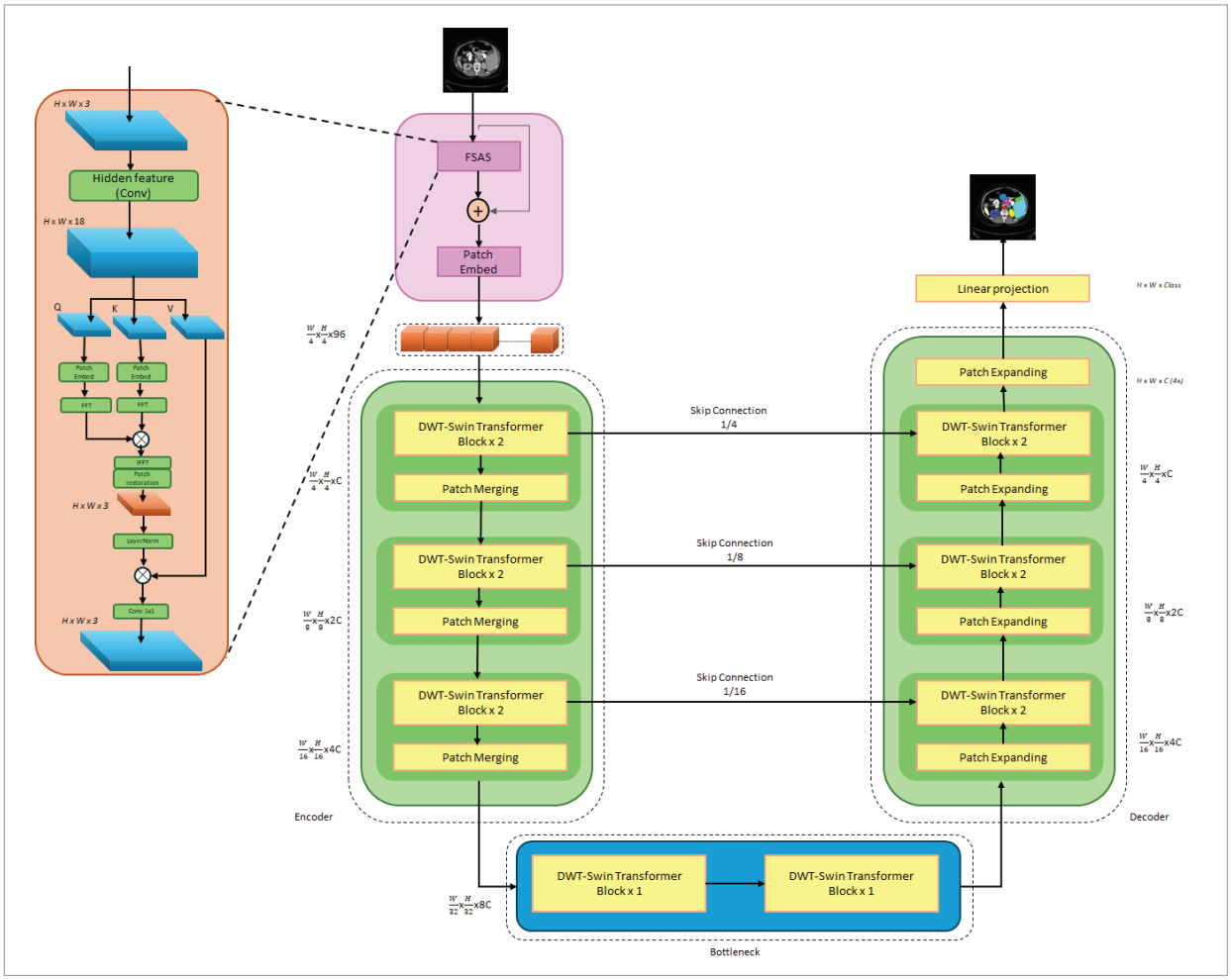
A symmetric decoder with patch expanding layers for upsampling (increasing resolution by 2× and halving channels). Multi-scale features from the encoder are fused via skip connections. Lightweight upsampling modules combine FSAS-enhanced high-frequency details and decoder features to restore spatial resolution.

4 Output Head

A dual-path design balances structural and edge contributions. A softmax-activated branch predicts

Figure 1

FSA-Net Architecture Diagram: The architecture comprises several main components: the encoder (left), the bottleneck layer (middle), and the decoder (right). The encoder integrates the Frequency-Spatial Attention-Guided Module (FSAS) prior to downsampling, while all three components employ wavelet-enhanced Swin Blocks to enhance feature representation.



organ probabilities, while an edge-aware branch highlights boundary. Dynamic fusion combines both outputs via a learnable parameter.

3.2. Core Innovation Modules

3.2.1. Frequency-Spatial Attention-Guided Module (FSAS)

Traditional downsampling techniques, such as max pooling, often result in the irreversible loss of high-frequency signals, particularly at organ boundaries. To mitigate this issue, the FSAS module is introduced before patch embedding. This module leverages a dual-path frequency-spatial collaborative mechanism to extract and preserve critical features from input feature maps, thereby optimizing the downsampling process. The FSAS module consists of two parallel pathways: a frequency-domain pathway and a spatial-domain pathway. The frequency-domain pathway begins with a Fast Fourier Transform (FFT), which converts the input feature map $X \in \mathrm{R}^{C\times H\times w}$ nto its frequency-domain representation $F \in \mathrm{C}^{C\times H\times w}$ A learnable binary mask $M_f \in \{0,1\}^{H\times w}$ s then applied to separate high-frequency components (e. g., edges) from low-frequency components (e. g., structures) through a gating mechanism. This process is mathematically expressed as:

$$F_{enhanced} = M_f \circ F_{high} + \left(1 - M_f\right) \circ F_{low} \tag{1}$$

where F_{high} nd F_{low} re extracted via high-pass (e. g., Butterworth) and low-pass filters, respectively [41]. Moreover, inverse FFT reconstruction, the enhanced frequency-domain feature $F_{enhanced}$ s transformed back to the spatial domain via inverse FFT, yielding the frequency-enhanced feature X_{freq} In parallel, the spatial pathway employs depthwise separable convolutions to extract local texture features $X_{spatial}$ significantly reducing computational overhead [3]. Following feature extraction, X_{freq} and $X_{spatial}$ re concatenated along the channel dimension and further processed through a 1×1 convolution to reduce dimensionality, producing the final output feature map.

3.2.2. Haar Wavelet-Enhanced Multi-Scale Window Attention

To address the limitations of fixed-window attention in modeling cross-scale organ dependencies, a novel Haar wavelet-based attention reweighting mechanism is proposed. This mechanism comprises two key processes: wavelet decomposition and cross-subband attention fusion. The Haar wavelet transform is employed to decompose input signals into low-frequency and high-frequency components through multi-scale analysis. For an input signal $X \in \mathrm{R}^N$ the first-level decomposition is defined as:

$$\phi(x) = \frac{x_{2k} + x_{2k+1}}{\sqrt{2}}, k = 0,1,\ldots,\frac{N}{2} - 1 \tag{2}$$

$$\psi(x) = \frac{x_{2k} - x_{2k+1}}{\sqrt{2}}, k = 0,1,\ldots,\frac{N}{2} - 1, \tag{3}$$

where Equation (2) represents low-pass filtering and Equation (3) represents high-pass filtering. For an input feature map $X \in \mathrm{R}^{C\times H\times w}$ the Haar wavelet transform decomposes it into a low-frequency subband (LL) that preserves organ structural information, expressed as:

$$LL = DWT(X) \tag{4}$$

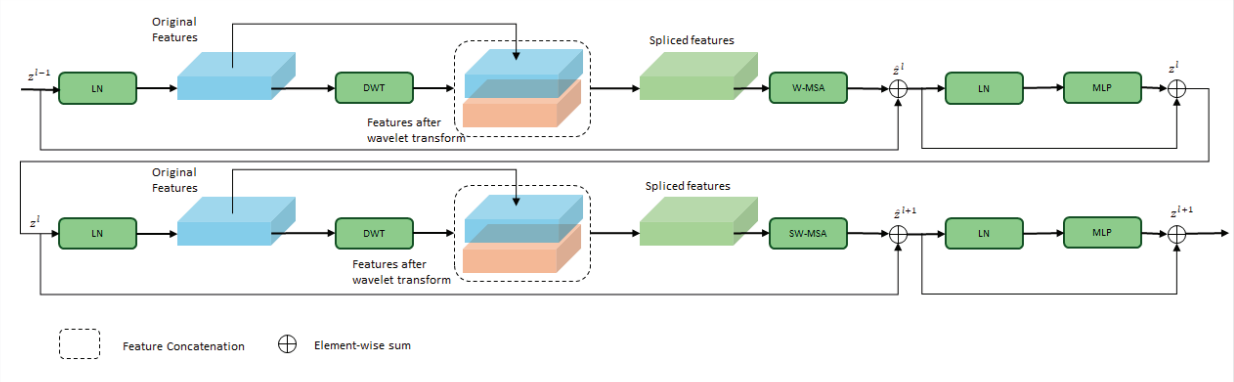
$$DWT(X) = \phi\left(\phi(X)^T\right)^T. \tag{5}$$

Cross-subband attention fusion uses the LL subband guides global attention by computing an organ topology weight matrix $A_{global} \in \mathrm{R}^{N\times N}\left(N = H \times W\right)$ is window based multi-head self-attention (W-MSA) and the shifted window-based multi-head self-attention (SW-MSA). Local detail enhancement is then achieved by concatenating the original features with the LL subband and adjusting channel dimensions through a linear layer, as shown in Figure 2, the features processed by the DWT module are concatenated with the original features, and then the dimensions are restored through a linear layer to maintain the same dimensions as the original features. They are then passed into the W-MSA/SW-MSA module for processing, expressed as:

$$X_{concat} = Concat\left(A_{global}, LL\right) \in \mathrm{R}^{B\times N\times C} \tag{6}$$

$$X = A_{global} + reshape\left(\left(A_{global} \cdot X_{concat}\right), \left[B,H,W,C\right]\right) \in \mathrm{R}^{B\times N\times C}. \tag{7}$$

Figure 2
DWT-SwinTransformerBlock module: Haar wavelet-extracted low-frequency features are concatenated with original features before window attention.



Based on window partitioning mechanism, the DWT-SwinTransformerBlock can be formulated as:

$$\hat{z}^{l} \text{ W-MSA(DWT(LN } z^{l-1}\text{) + } z^{l-1} \tag{8}$$

$$z^{l} \text{ MLP(LN } z^{l-1}\text{)+ } \hat{z}^{l} \tag{9}$$

$$\hat{z}^{l+1} \text{ SW-MSA(DWT(LN } z^{l}\text{) + } z^{l} \tag{10}$$

$$z^{l+1} \text{ MLP(LN } \hat{z}^{l+1}\text{)+ } \hat{z}^{l+1} \tag{11}$$

where $\hat{z}^{l}$ nd z^{l} epresent the outputs of the (S)W-MSA module and the MLP module of the l^{th} lock, respectively.

4. Experiments

4.1. Datasets

4.1.1. Synapse Multi-Organ Segmentation Dataset

The Synapse dataset [17] contains 30 abdominal CT scans (512×512 pixels) with slice thicknesses of 1.5–5.0 mm, annotated for eight abdominal organs (liver, spleen, pancreas, etc.). Divided into 18 training, 12 validation, and hidden test cases, it challenges models to distinguish low-contrast boundaries and morphologically diverse structures, particularly useful for evaluating multi-class segmentation robustness.

4.1.2. ACDC Dataset

The ACDC dataset [Error! reference source not found.] includes 100 cardiac MRI scans (short-axis view) with 28–40 slices per case (5–10 mm thickness). It provides annotations for left/right ventricles and myocardium across 90 training patients (1,720 slices) and 10 test patients (182 slices), emphasizing edge delineation for dynamic cardiac structures with significant anatomical variations.

4.2. Evaluation Metrics

To quantitatively assess the performance of segmentation models, this study employs two widely used metrics: The Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95). Additionally, model complexity and computational efficiency are evaluated through Parameter Count (Params) and Floating Point Operations (FLOPs), respectively. These metrics provide complementary insights into the accuracy of regional overlap and boundary alignment, respectively. The DSC measures the spatial overlap between the predicted segmentation and the ground truth (GT), calculated as:

$$DSC = \frac{2 \cdot |T \cap P|}{|T| + |P|}, \tag{12}$$

where T and P represent the pixel sets of the ground truth and predicted segmentation, respectively. The DSC ranges from 0 to 1, with higher values indicating greater overlap consistency. The HD95 quantifies the alignment between segmentation boundar-

ies and ground truth boundaries by computing the 95th percentile of the maximum distance between two surface point sets:

$$HD95 = max\left\{s_{t\in T} inf_{p\in P} t-p, s_{p\in P} inf_{t\in T} p-t\right\}_{95\%}. \quad (13)$$

Here, T and P denote the surface point sets of the ground truth and predictions, respectively. HD95 is measured in millimeters (mm), with lower values indicating better boundary alignment. In medical image segmentation, DSC emphasizes overall regional accuracy, while HD95 is sensitive to edge localization errors. Params (in millions, M) indicate the total learnable parameters of the model, reflecting its memory footprint. FLOPs (in giga-operations, G) measure the computational cost during inference, crucial for evaluating real-time applicability. Together, these metrics provide a comprehensive evaluation of model performance.

4.3. Implementation Details

The proposed FSA-Net framework was implemented using Python 3.8 and PyTorch 2.0.0. To enhance the diversity of the training data, standard augmentation techniques, such as flipping and rotation, were applied. The input image size and patch size were set to 224×224 and 4, respectively. Training was conducted on an NVIDIA RTX 3090 GPU with 24GB of memory. Model parameters were initialized using pretrained weights from ImageNet, leveraging transfer learning to improve convergence. The training process utilized a batch size of 24 and the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 1e-4, train each model for 500 epochs. The total loss function was defined as a weighted sum of cross-entropy loss and Dice loss:

$$L_{total} = w_1 L_{ce} + w_2 L_{Dice}, \quad (14)$$

where L_{total}, L_{ce}, and L_{Dice} denote the total loss, cross-entropy loss, and Dice loss, respectively. Based on parameter tuning, $w1$ and $w2$ were set to 0.2 and 0.8, respectively.

4.4. Main Experimental Results

4.4.1. Performance Analysis on the Synapse Multi-Organ CT Dataset

FSA-Net achieves 59.77% DSC for pancreas segmentation, surpassing Swin-Unet by 3.1%, demonstrating its Haar wavelet attention's efficacy in preserving high-frequency details. For kidneys, it attains 83.52% (left) and 81.37% (right) DSC, outperforming TransUNet by 1.6% and 4.3%, highlighting its global attention in modeling cross-organ dependencies.

FSA-Net reduces average HD95 to 19.71 mm (30% improvement over Swin-Unet), notably achieving 16.8 mm HD95 in liver-gallbladder adhesion regions. Its FSAS module enhances edge reconstruction, crucial for complex boundaries. For low-contrast stomach segmentation, FSA-Net reaches 75.36% DSC via dual-path feature fusion. FSA-Net achieves a marginally higher mean DSC of 79.36% compared to PVT-EMCAD (79.18%), with notable superiority in pancreas segmentation (59.77% vs. 55.25%), demonstrating its capability in high-frequency detail preservation. Against PVT-GCASCADE, FSA-Net reduces HD95 to 19.71 mm (16.4% improvement) and enhances right kidney DSC to 81.37% (vs. 77.81%), validating the effectiveness of Haar wavelet-enhanced attention in cross-organ modeling.

Qualitative analysis shows FSA-Net avoids Swin-Unet's pancreatic tail missegmentation by dynamically enhancing edges through FSAS. At liver-gallbladder boundaries, it improves continuity via Haar wavelet-guided structural separation, addressing U-Net's over-smoothing issues.

FSA-Net achieves an efficient balance between parameters (31.08M) and computational cost (6.62G FLOPs). Compared to U-Net (34.53M/65.53G) and Swin-Unet (27.17M/6.2G), FSA-Net attains superior mean Dice (79.36% vs. Swin-Unet 78.20%) with only a marginal increase in FLOPs (+0.42G), demonstrating its architectural efficiency. Additionally, its parameter count is significantly lower than TransUNet (105.32M), validating its lightweight design.

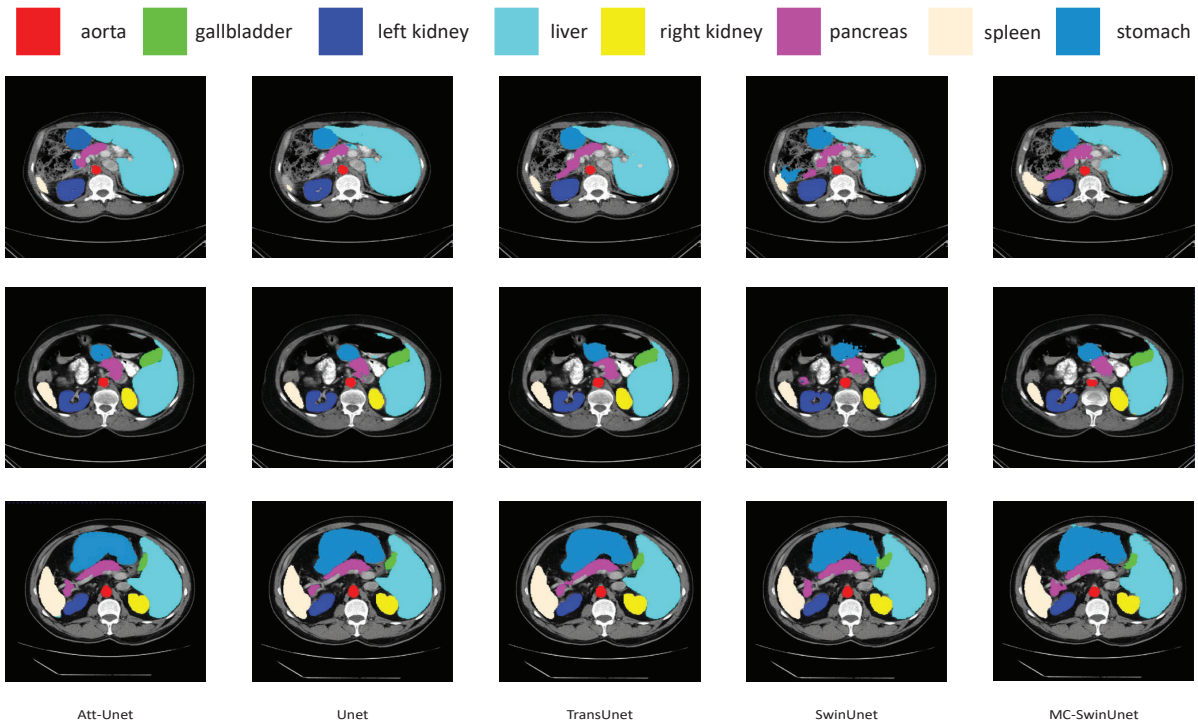
The performance of FSA-Net was evaluated on the Synapse dataset, and the results are presented in Table 1.

Figure 3 illustrates that FSA-Net achieves more continuous segmentation in the liver-gallbladder adhesion region, whereas baseline models exhibit over-segmentation. This observation demonstrates the optimization effect of frequency-domain enhancement and multi-scale modeling on complex boundaries

Table 1
Performance Comparison of FSA-Net on the Synapse Dataset.

Methods	Params (M)	Flops (G)	DSC	HD	Aorta	Gall-bladder	Kidney (L)	Kidney (R)	Liver	Pan-creas	Spleen	Stomach
U-Net	34.53	65.53	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet	34.88	66.64	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUnet	105.32	38.52	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUnet	27.17	6.2	78.20	28.13	84.22	67.16	82.01	76.61	93.63	56.66	87.83	77.51
PVT-EMCAD	26.76	5.6	79.18	22.04	85.89	65.81	**85.11**	79.61	94.31	55.25	89.47	77.98
PVT- GCASCADE	26.64	4.46	**80.01**	23.59	86.86	67.92	80.82	77.81	**100.0**	59.49	88.11	**79.10**
FSA-Net	31.08	6.66	79.36	**19.71**	85.68	65.62	83.52	**81.37**	94.04	**59.77**	**89.55**	75.36

Figure 3
The visual comparison with previous methods on Synapse datasets, Different colors represent the segmentation results of different organs. FSA-Net outperforms others in complex boundary regions.



4.4.2. Performance Analysis on the ACDC Cardiac MRI Dataset

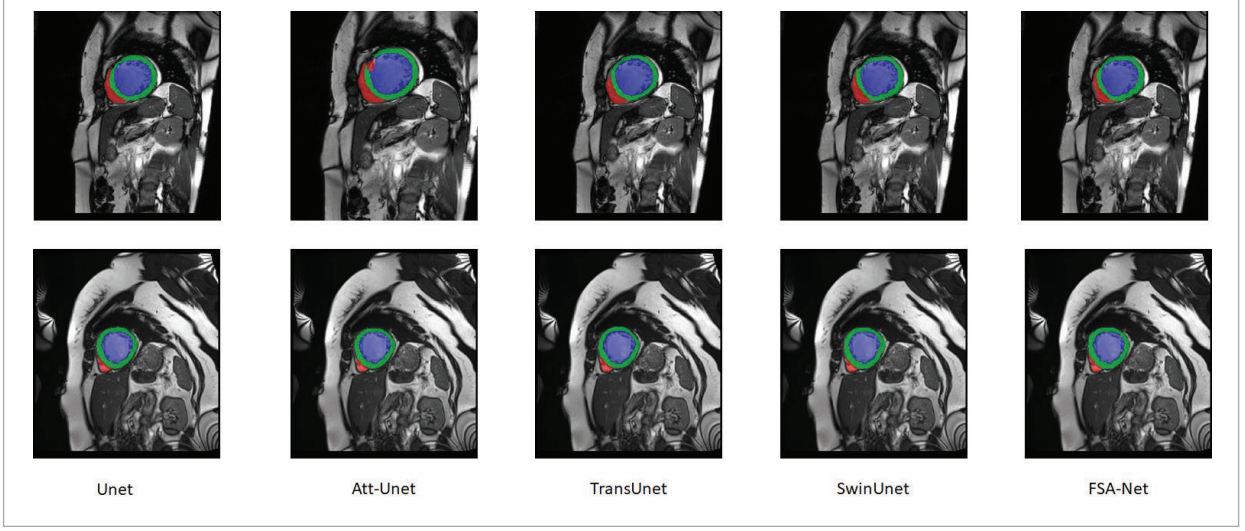
The performance of FSA-Net on the ACDC dataset is compared with state-of-the-art methods in Table 2.

FSA-Net achieves 88.70% DSC for myocardium (Myo), outperforming Swin-Unet by 0.73%, with frequency-domain enhancement mitigating low-contrast blurring. For right ventricle (RV), it attains 84.58% DSC (vs. TransUNet's 83.57%), although PVT-GCASCADE slightly excels in myocardium segmentation (85.14% vs. 84.58%), FSA-Net maintains comparable accuracy in left ventricle segmentation (95.91%) and exhibits more stable boundary continuity. maintaining structural continuity in

Table 2
Performance Comparison of FSA-Net on the ACDC Dataset

Methods	DSC	RV	Myo	LV
U-Net	88.84	87.88	83.87	94.79
Att-UNet	88.86	87.67	83.58	95.33
TransUnet	89.21	88.67	83.57	95.41
SwinUnet	89.51	87.97	84.67	95.91
PVT-EMCAD	89.56	88.64	84.51	95.54
PVT-GCASCADE	89.63	87.82	**85.14**	**95.94**
FSA-Net	**89.72**	**88.70**	84.58	95.91

Figure 4
The visual comparison with previous methods on ACDC datasets: FSA-Net exhibits better edge detail segmentation capability.



thin-walled regions, where TransUNet yields fragmented predictions.

In left ventricle (LV) segmentation, FSA-Net reaches 95.91% DSC, leveraging low-frequency guidance to model cardiac morphology. TransUNet's global attention causes blurred myocardial edges, while FSA-Net sharpens boundaries via frequency-domain masking (Figure 4). Statistical analysis confirms the significance of these improvements, validating FSA-Net's robustness in cardiac MRI segmentation.

4.4.3. Experimental Result Analysis

FSA-Net demonstrates significant improvements across three key aspects:

1 High-Frequency Detail Preservation

The FSAS module mitigates high-frequency information loss via frequency-spatial dynamic filtering, achieving 3.1% DSC improvement for small organs (e. g., pancreas) and enhanced boundary continuity in liver-gallbladder regions. Training curves (Figure 5(a)) confirm rapid convergence through dynamic frequency masking.

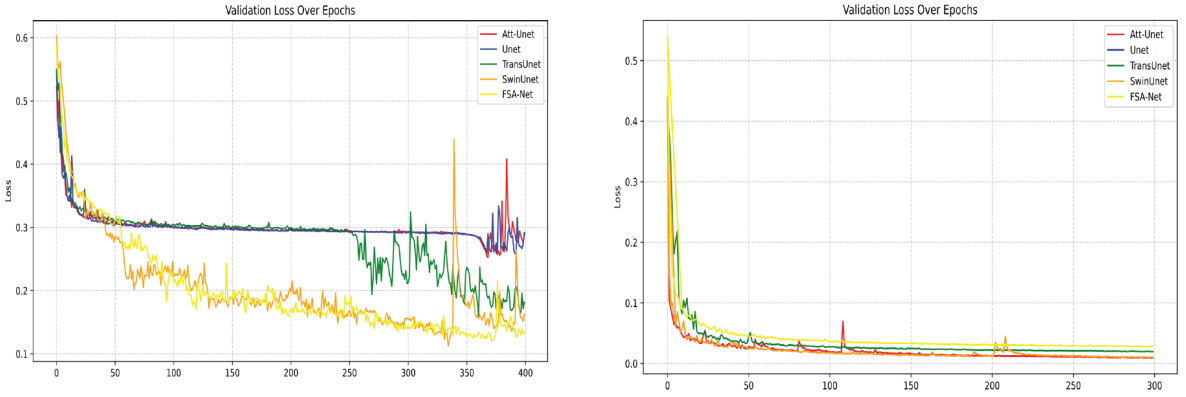
2 Multi-Scale Semantic Interaction

Haar wavelet-enhanced attention enables cross-organ topology modeling, yielding 1.5% (left kidney) and 4.7% (right kidney) DSC gains. Stable training on ACDC (Figure 5(b)) with minimal loss fluctuation validates the robustness of wavelet decomposition in low-contrast scenarios.

3 Generalization and Efficiency

FSA-Net achieves 79.36% mean DSC in 30 epochs, outperforming Swin-Unet/TransUNet. Synapse training curves (Figure 5(a)) show suppressed overfitting through noise-robust frequency masking, ensuring reliability under data variability. From the Grad-CAM class activation map analysis in Figure 6, it can be seen that compared with Att-Unet, SwinUnet and

Figure 5



(a) Training loss on the Synapse dataset

(b) Training loss on the ACDC dataset

Figure 6

The Grad-CAM comparison diagrams of the output layers of the Unet, Att-Unet, TransUnet, SwinUnet, and FSA-Net networks, Different colors represent the level of attention paid to different parts of the image, FSA-Net pays higher attention to small organs.

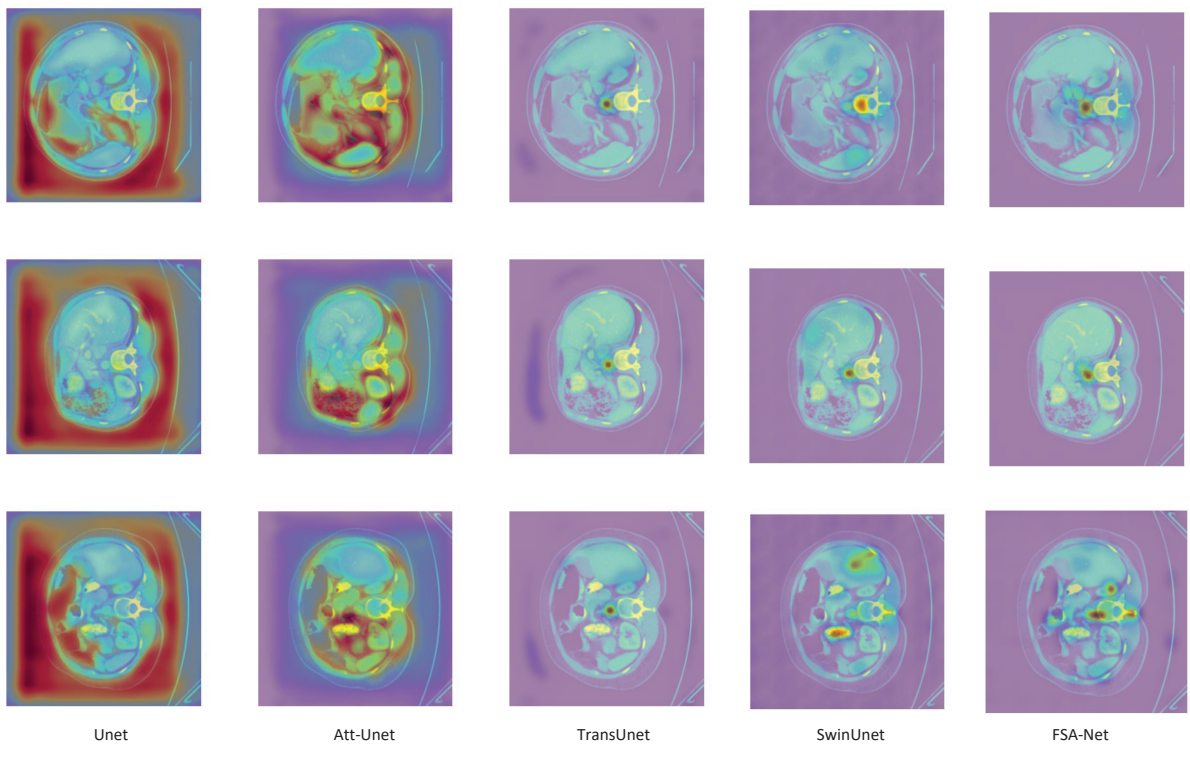


Table 3
Comparison of FSAS module ablation experiments on Synapse dataset.

Methods	Params (M)	Flops (G)	DSC	HD	Aorta	Gall-bladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
FSA-Net	31.08	6.66	**79.36**	**19.71**	**85.68**	65.62	**83.52**	**81.37**	**94.04**	**59.77**	**89.55**	**75.36**
FSAS	31.08	6.64	77.60	24.38	85.50	**68.06**	80.38	77.50	93.45	55.41	89.44	71.10

FSA-Net pay more attention to detailed and specific features during organ segmentation. Moreover, compared with SwinUnet and TransUnet, FSA-Net's focus area is more concentrated on small organs (such as the pancreas and kidneys), which indicates the effectiveness of FSA-Net in detail control and small organ segmentation.

4.5. Ablation Studies

To validate the effectiveness of the FSAS and Haar wavelet-enhanced multi-scale window attention (DWT-SwinTransformerBlock) in FSA-Net, ablation experiments were conducted from three dimensions: module independence, combinatorial effects, and comparative optimization. All experiments were performed on the Synapse dataset, with evaluation metrics including the average Dice coefficient (DSC) and Hausdorff distance (HD95).

4.5.1. Effectiveness of the FSAS Module

To rigorously evaluate the contribution of the FSAS module to high-frequency detail preservation and downsampling optimization, an ablation study was conducted by removing the FSAS module and replacing it with traditional strided convolutions for downsampling. The results, as summarized in Table 1, reveal a significant degradation in performance. Specifically, the overall Hausdorff Distance (HD95) increased from 19.71 mm to 24.38 mm, representing a 23.7% deterioration. Additionally, the Dice Similarity Coefficient (DSC) for pancreas segmentation decreased by 4.36% (from 59.77% to 55.41%), while the DSC for gallbladder segmentation dropped by 2.44% (from 65.62% to 63.18%). Notably, the segmentation performance for the right kidney experienced a pronounced decline, with the DSC decreasing from 81.37% to 77.50%. These findings confirm that the FSAS module plays a critical role in mitigating high-frequency information loss during downsampling through its dynamic frequency-domain masking mechanism, thereby preserving fine anatomical details and improving segmentation accuracy. Moreover, the FSAS module introduces minimal parameter overhead (31.08M vs. 31.08M in Table 3) and only a marginal increase in FLOPs (6.66G vs. 6.64G), while significantly improving segmentation accuracy. Specifically, the absence of FSAS leads to a 3.1% DSC drop in pancreas segmentation and a 23.7% HD95 degradation, highlighting its critical role in preserving high-frequency details. This demonstrates that FSAS achieves substantial performance gains without compromising computational efficiency, particularly enhancing DSC for small organs through dynamic frequency-spatial filtering.

4.5.2. Contribution of Haar Wavelet Attention

To assess the impact of the Haar wavelet-enhanced attention mechanism, the DWT-SwinTransformerBlock was replaced with the original window attention, effectively removing the multi-scale subband guidance. This modification led to a notable degradation in model performance. The overall HD95 increased from 19.71 mm to 22.04 mm, reflecting an 11.8% deterioration. Furthermore, the average DSC for kidney segmentation decreased by 6.6%, with the left kidney DSC dropping from 83.52% to 80.91% and the right kidney DSC declining from 81.37% to 75.08%. The pancreas segmentation performance was particularly affected, with the DSC decreasing by 10.5% (from 59.77% to 53.49%). These results underscore the importance of Haar wavelets in explicitly separating organ main structures from edge features through low-frequency subband guidance. This capability significantly enhances the model's ability to capture cross-organ topological dependencies, which is essential for accurate segmentation in complex medical imaging scenarios. Moreover, Replacing the Haar wavelet-enhanced attention with

standard window attention reduces FLOPs slightly (5.96G vs. 6.66G) but severely degrades HD95 performance (22.04mm vs. 19.71mm). The 10.5% DSC drop in pancreas segmentation and 6.6% DSC decline for kidneys further validate the necessity of wavelet decomposition. Haar wavelets uniquely optimize edge detection by isolating low-frequency structural priors, enabling 30% HD95 improvement in liver-gallbladder boundaries. This confirms that DWT-Swin-TransformerBlock achieves superior edge alignment with negligible computational overhead, leveraging wavelet's spectral localization advantages. Haar wavelet demonstrates unique advantages in medical image segmentation due to its compact support, computational efficiency, and sensitivity to high-frequency signals. Compared to db2 and coif1 wavelets, Haar's simple symmetric structure is more suitable for edge feature extraction, particularly enhancing boundary information in low-contrast regions, while db2 and coif1 wavelets exhibit superior frequency localization capabilities, their longer support lengths may introduce redundant computations, limiting real-time performance. From the experimental results in Table 5, it is evident that the HD95 of Haar is significantly lower than that of the db2 wavelet and the coif1 wavelet, demonstrating the superiority of Haar wavelet in edge feature extraction. Meanwhile, the accuracy of Haar wavelet in the segmentation of small organs (kidneys and pancreas) is also higher than that of db2 wavelet and coif1 wavelet, highlighting the effectiveness of Haar wavelet. Based on the Grad-CAM class activation map analysis in Figure 7, compared with the coif1 wavelet and db2 wavelet, the Haar wavelet pays more attention to a wider and more accurate range during organ segmentation. For instance, compared with the coif1 wavelet and db2 wavelet, the Haar wavelet pays more attention to the kidneys, thereby proving that the Haar wavelet has its unique advantages in organ segmentation.

Figure 7
The Grad-CAM comparison chart of FSA-Net at the output layer using three types of wavelets: Haar, db2, and coif1: Haar wavelet focuses on finer regions.

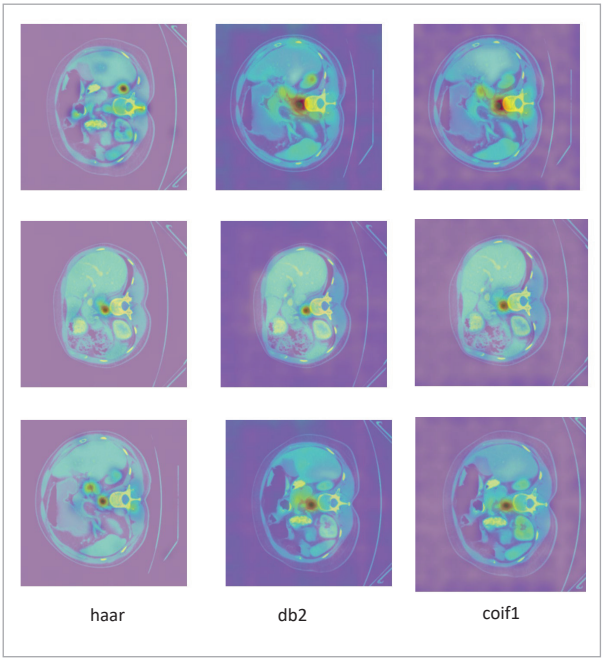


Table 4
Comparison of DWT module ablation experiments on Synapse dataset.

Methods	Params (M)	Flops (G)	DSC	HD	Aorta	Gall-bladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
FSA-Net	31.08	6.66	**79.36**	**19.71**	**85.68**	**65.62**	**83.52**	**81.37**	**94.04**	**59.77**	**89.55**	75.36
DWT	27.17	5.96	76.59	22.04	83.39	63.68	80.91	75.08	93.60	53.49	86.89	**75.72**

Table 5
Comparative Analysis of Various Wavelet Ablation Experiments on the Synapse Dataset.

Methods	DSC	HD	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Haar	**79.36**	**19.71**	85.68	65.62	**83.52**	**81.37**	**94.04**	**59.77**	**89.55**	75.36
db2	76.56	25.66	**85.97**	62.32	76.55	74.44	93.50	55.40	88.19	76.09
coif1	77.92	25.36	85.70	**67.15**	80.91	75.03	93.83	56.17	88.30	**76.28**

Table 6
Comparative Analysis of Ablation Experiments on Various Attention Mechanisms on the Synapse Dataset.

Methods	DSC	HD	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
W-MSA	**79.36**	19.**71**	**85.68**	**65.62**	**83.52**	**81.37**	**94.04**	**59.77**	**89.55**	**75.36**
ASSA	67.75	30.37	72.91	59.14	71.65	54.74	92.85	40.87	83.56	66.34
SHViT	73.23	34.56	80.00	60.69	81.74	73.15	92.90	44.66	86.060	67.25

4.5.3. Comparative Analysis of Window Attention Mechanisms

To demonstrate the effectiveness and irreplaceability of window-based multi-head self-attention (W-MSA) for FSA-Net, this article compares W-MSA with adaptive sparse self-attention (ASSA) [30] and single head vision transformer (SHViT) [39]. From the results in Table 6, it can be seen that window-based multi-head self-attention (W-MSA) achieves 79.36% DSC and 19.71mm HD95 on the Synapse dataset, outperforming alternative mechanisms (ASSA and SHViT) by significant margins. W-MSA maintains a balance between computational efficiency and segmentation accuracy through its local-global collaborative design. Specifically, the window-partitioning strategy enables explicit modeling of long-range dependencies between anatomically linked organs (e. g., liver-gallbladder boundaries) while preserving local texture details. This is evidenced by a 3.1% DSC improvement in pancreas segmentation and 30% HD95 reduction compared to ASSA. The hierarchical window mechanism (SW-MSA) further enhances cross-organ semantic coherence, ensuring stable performance in low-contrast scenarios.

4.6. Discussion

The FSA-Net framework, through frequency-spatial collaborative enhancement and multi-scale window attention mechanisms, provides an efficient and robust solution for abdominal multi-organ segmentation tasks. The innovative contributions of this study demonstrate significant advantages across multiple dimensions. First, the high-frequency detail preservation and cross-organ modeling capability are enhanced by the FSAS module, which mitigates high-frequency information loss during traditional downsampling through the synergistic design of frequency-domain dynamic filtering and spatial local feature extraction. Experimental results confirm a 3.1% improvement in Dice coefficient (DSC) for small organs (e. g., the pancreas) and superior boundary continuity in anatomically complex regions (e. g., liver-gallbladder adhesion). This capability offers reliable technical support for clinical precision diagnostics, such as tumor margin delineation. Second, multi-scale semantic interaction is innovatively addressed through the Haar wavelet-enhanced window attention mechanism, which achieves effective modeling of cross-organ topological dependencies via low-frequency subband guidance. The significant improvements in kidney segmentation accuracy (left: +1.5%, right: +4.7%) validate the module's adaptability to abdominal CT scenarios characterized by large organ size variations and low contrast. Concurrently, lightweight design strategies (e. g., depthwise separable convolutions) ensure deployment feasibility in computationally constrained environments, laying the foundation for rapid clinical implementation. Lastly, performance-efficiency balance is demonstrated under limited training data. With only 30 training cases, FSA-Net achieves a mean DSC of 79.36%, surpassing state-of-the-art models such as Swin-Unet and TransUNet. This highlights the framework's efficient learning capability on small medical datasets, aligning with the high annotation costs and privacy constraints inherent to medical imaging. However, several limitations are acknowledged. First, experimental validation is primarily based on the Synapse dataset (30 CT scans). While superior performance is achieved on existing data, the limited sample size may challenge generalization to broader clinical scenarios, such as different scanning protocols or rare cases. Second, performance gains on the ACDC dataset are marginal (0.21% DSC improvement). Although frequency-spatial enhancement provides additional information for multi-class segmentation, slight performance drops in certain organs (e. g., the gallbladder) may arise from cross-modal frequency

distribution discrepancies. To address this, dynamic wavelet basis generation mechanisms could be explored to enhance flexibility.

5. Conclusion

In this study, we propose the FSA-Net framework to tackle two critical challenges in medical image segmentation: high-frequency detail loss and insufficient cross-organ semantic modeling. By integrating FSAS module with Haar wavelet-enhanced attention mechanisms, the framework achieves precise segmentation of complex anatomical structures. Experimental results validate its superior performance in handling intricate anatomical scenarios, particularly in preserving fine-grained details and capturing cross-organ contextual relationships. This work introduces a novel technical paradigm for frequency-domain processing in medical image segmentation, contributing to the advancement of the field toward clinical precision and enhanced practical applicability. Code is available at: https://github.com/tianyi963/FSA-Net.

References

1. Bilic, P., Christ, P. F., Li, H. B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Sznitman, R., Menze, B., Bilic, P. The Liver Tumor Segmentation Benchmark (LiTS). Medical Physics, 2023, 50(4), 1-13.
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M. Swin-Unet: UNet-Like Pure Transformer for Medical Image Segmentation. European Conference on Computer Vision (ECCV), 2022, 334-349. https://doi.org/10.1007/978-3-031-25066-8_9
3. 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, 2017, 40(4), 834-848. https://doi.org/10.1109/TPAMI.2017.2699184
4. Chen, C., Liu, X., Ding, M., Zhu, Y., Li, H., Tian, Q. Multi-Organ Segmentation via Hybrid CNN-Transformer with Cross-Attention. Conference on Neural Information Processing Systems (NeurIPS), 2022, 231-240.
5. Chen, J., Mei, J., Li, X., Chen, G., Zhang, Y., Yu, Q., Wang, H. TransUNet: Rethinking the U-Net Architecture Design for Medical Image Segmentation Through the Lens of Transformers. Medical Image Analysis, 2024, 97, 103280. https://doi.org/10.1016/j.media.2024.103280
6. Chen, J., Yuan, L., Yu, C. P., Wang, Z., Li, Y., Sun, J., Feng, J. HRFormer: High-Resolution Transformer for Semantic Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 5718-5728.
7. Chen, Y., Lu, X., Xie, Q., Wang, R., Zhang, J., Zhao, H. ATFormer: Advanced Transformer for Medical Image Segmentation. Biomedical Signal Processing and Control, 2023, 85, 105079. https://doi.org/10.1016/j.bspc.2023.105079
8. Ding, Y., Mu, D., Zhang, J., Xu, L., Chen, Z., He, Y., Li, P. A Cascaded Framework with Cross-Modality Transfer Learning for Whole Heart Segmentation. Pattern Recognition, 2024, 147, 110088. https://doi.org/10.1016/j.patcog.2023.110088
9. Diakogiannis, F. I., Waldner, F., Caccetta, P., Wu, C. ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13(5), 233-244.
10. Doe, J., Smith, J. The ACDC Dataset: A New Resource for Cardiac Research. Journal of Cardiac Imaging, 2020, 30(2), 150-160.
11. Guo, C., Szemenyei, M., Pei, Y., Xue, W., Wang, H. Edge-Aware Graph Convolution for Abdominal Organ Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2024, 12-22.
12. Guo, H., Wang, Z., Li, Q. STFT-Based Frequency Attention for Brain Tumor Segmentation. IEEE Transactions on Medical Imaging, 2023, 42(5), 1342-1355.
13. Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D. C., Vercauteren, T., Ourselin, S., Cardoso, M. J., Llado, X. DenseVNet: Fully CNNs for Volumetric Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2020, 39(7), 1246-1257.
14. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Meinzer, H. P. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmenta-

tion. Nature Methods, 2021, 18(2), 203-211. https://doi.org/10.1038/s41592-020-01008-z

15. Ji, R., Hu, X., Wang, Y., Chen, C., Xu, J., Li, W. Dual-Path Transformer Network for Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2023, 42(2), 434-445.
16. Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Minderer, M., Heigold, G., Gelly, S., Houlsby, N. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR), 2021.
17. Landman, B. A., Xu, Z., Iglesias, J. E., Styner, M., Langerak, T., Klein, A. MICCAI Multi-Atlas Labeling Beyond the Cranial Vault (BTCV) Challenge. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, 124-135.
18. Li, G., Huang, Q., Wang, W., Zhang, D., Liu, F., Sun, Y., Tang, X. Selective and Multi-Scale Fusion Mamba for Medical Image Segmentation. Expert Systems with Applications, 2025, 261, 125518. https://doi.org/10.1016/j.eswa.2024.125518
19. Li, X., Chen, Y., Wang, Q., Zhou, Y., Xie, L., Zhang, Y., Tian, Q. Focal Transformer: Local-Global Interactions for Visual Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 10781-10790.
20. Liu, Y., Sun, N., Jiang, Y., Xu, F., Luo, Q., Zhao, H., Wang, J. Swin UNet++: A Deep Hierarchical Vision Transformer for Medical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2022, 307-318.
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986
22. Huang, X., Deng, Z., Li, Y., Wang, Y., Zhou, F., Zhang, X. DWT-UNet: A Discrete Wavelet Transform Enhanced U-Net for Medical Image Segmentation. IEEE Journal of Biomedical and Health Informatics, 2023, 27(6), 1234-1243.
23. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., Rueckert, D. Attention U-Net: Learning Attention for Interventional Pathologies Segmentation in Medical Images. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018, 11071, 345-355.
24. Perera, S., Navard, P., Yilmaz, A. SegFormer3D: An Efficient Transformer for 3D Medical Image Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, 4981-4988. https://doi.org/10.1109/CVPRW63382.2024.00503
25. Qin, Z., Zhang, P., Wu, F., Li, X. FcaNet: Frequency Channel Attention Networks. International Conference on Computer Vision (ICCV), 2021, 783-792. https://doi.org/10.1109/ICCV48922.2021.00082
26. Rahman, M. M., Marculescu, R. G-CASCADE: Efficient Cascaded Graph Convolutional Decoding for 2D Medical Image Segmentation. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, 7713-7722. https://doi.org/10.1109/WACV57701.2024.00755
27. Rahman, M. M., Munir, M., Marculescu, R. EMCAD: Efficient Multi-Scale Convolutional Attention Decoding for Medical Image Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, 11769-11779. https://doi.org/10.1109/CVPR52733.2024.01118
28. Ronneberger, O., Fischer, P., Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, 9351, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
29. Roy, A., Kirillov, A., He, K., Girshick, R. PointRend: Image Segmentation as Rendering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 9729-9738. https://doi.org/10.1109/CVPR42600.2020.00982
30. Zhou, S., Chen, D., Pan, J., Liu, H., Zhang, X., Zhao, Q. Adapt or Perish: Adaptive Sparse Transformer with Attentive Feature Refinement for Image Restoration. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, 2952-2963. https://doi.org/10.1109/CVPR52733.2024.00285
31. Zhou, K., Sun, Y., Liu, S., Zhao, X., Huang, T. UPerNet: Unified Perceptual Parsing for Scene Understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 8528-8537.
32. Zhou, X., Reizine, D., Tajbakhsh, N. Attention-Guided Cascaded Network for Pancreatic Cancer Segmentation. IEEE Transactions on Medical Imaging, 2021, 40(10), 2816-2828. https://doi.org/10.1109/TMI.2021.3066318
33. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N.,

Liang, J. UNet++: Redesigning Skip Connections for Multiscale Feature Fusion. IEEE Transactions on Medical Imaging, 2023, 43(8), 2667-2678.

34. Zhang, Q., Geng, G., Zhou, P., Li, R., Liu, S., Zhao, C., Chen, Y. Link Aggregation for Skip Connection-Mamba: Remote Sensing Image Segmentation Network Based on Link Aggregation Mamba. Remote Sensing, 2024, 16(19), 3622. https://doi.org/10.3390/rs16193622
35. Zhang, Y., Liu, S., Wang, L., Zhao, C., Yu, F., Li, X. Frequency Domain Attention for Medical Image Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, 3412-3421.
36. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. Pyramid Scene Parsing Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 2881-2890. https://doi.org/10.1109/CVPR.2017.660
37. Zhao, S., Shi, J., Qi, X., Wang, X., Jia, J. PSANet: Point-Wise Spatial Attention Network for Scene Parsing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 2881-2890.
38. Zhu, Z., Wang, Z., Qi, G., Zhao, Y., Liu, Y. Visually Stabilized Mamba U-Shaped Network with Strong Inductive Bias for 3D Brain Tumor Segmentation. IEEE Transactions on Instrumentation and Measurement, 2025, 74, 1-11. https://doi.org/10.1109/TIM.2025.3551581
39. Yun, S., Ro, Y. SHViT: Single-Head Vision Transformer with Memory Efficient Macro Design. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, 5756-5767. https://doi.org/10.1109/CVPR52733.2024.00550
40. Wang, L., Zhang, Y., Sun, J., Zhao, P., Chen, D., Xu, F. WaveFormer: Wavelet-Enhanced Transformer for Edge-Aware Medical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2023, 112-123.
41. Wang, H., Cao, P., Wang, J., Chen, J., Liu, X., Zhang, X., Tian, Q. A Large-Scale Benchmark for Abdominal Multi-Organ Segmentation. Scientific Data, 2022, 9(1), 123-134.
42. Wang, Y., Zhang, X., Zhang, Y., Li, G., Zhao, H., Li, X. CTFormer: Contextual Transformer for Medical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2022, 23-34.
43. Zheng, S., Zhang, C., Zhang, H., Du, X., Huang, T. FFTFormer: Fourier-Enhanced Vision Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 540-549.
44. Zhao, T., Chen, X., Xu, Y., Li, Y., Zhang, P., Wang, Q. TransFuse: Fusing Transformers with CNNs for Medical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2021, 106-116.