# Decoding with Structured Awareness: Integrating Directional, Frequency-Spatial, and Structural Attention for Medical Image Segmentation

**Fan Zhang**[1], **Zhiwei Gu**[1], **Hua Wang**[2*]

[1]School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China
[2]School of Computer and Artificial Intelligence, Ludong University, Yantai 264025, China
zhangfan@sdtbu.edu.cn, guzhiwei409@gmail.com, hwa229@163.com

## Abstract

To address the limitations of Transformer decoders in capturing edge details, recognizing local textures and modeling spatial continuity, this paper proposes a novel decoder framework specifically designed for medical image segmentation, comprising three core modules. First, the Adaptive Cross-Fusion Attention (ACFA) module integrates channel feature enhancement with spatial attention mechanisms and introduces learnable guidance in three directions (planar, horizontal, and vertical) to enhance responsiveness to key regions and structural orientations. Second, the Triple Feature Fusion Attention (TFFA) module fuses features from Spatial, Fourier and Wavelet domains, achieving joint frequency-spatial representation that strengthens global dependency and structural modeling while preserving local information such as edges and textures, making it particularly effective in complex and blurred boundary scenarios. Finally, the Structural-aware Multi-scale Masking Module (SMMM) optimizes the skip connections between encoder and decoder by leveraging multi-scale context and structural saliency filtering, effectively reducing feature redundancy and improving semantic interaction quality. Working synergistically, these modules not only address the shortcomings of traditional decoders but also significantly enhance performance in high-precision tasks such as tumor segmentation and organ boundary extraction, improving both segmentation accuracy and model generalization. Experimental results demonstrate that this framework provides an efficient and practical solution for medical image segmentation.

## Introduction

Medical image segmentation plays a pivotal role in intelligent healthcare and clinical applications, aiming to accurately delineate organs, tumors, or lesions from complex medical images, thereby providing clinicians with structured and intuitive reference information. This not only significantly improves the accuracy of diagnosis and lesion assessment but also offers crucial support for surgical planning, radiotherapy dose design, and treatment monitoring. With the rapid growth of medical imaging data, automated segmentation methods have become indispensable for reducing clinicians' workload, enhancing diagnostic efficiency, and ensuring result consistency. Deep learning, as a driving force of artificial intelligence, has greatly advanced the application of Convolutional neural networks (CNNs) in image segmentation. Early models such as LeNet (LeCun et al. 2002) and AlexNet (Krizhevsky, Sutskever, and Hinton 2012) demonstrated the advantages of deep convolutional architectures in feature extraction. Subsequent models like VGG (Simonyan and Zisserman 2014) and ResNet (He et al. 2016) enhanced feature representation and training stability through residual and multi-path designs. Building on this foundation, U-Net (Ronneberger, Fischer, and Brox 2015) proposed an encoder-decoder architecture that captures global semantic information through downsampling and recovers spatial details via skip connections that fuse multi-level features. However, traditional skip connections often rely on simple addition operations, which may lead to the loss of spatial details and the inclusion of redundant information, making it difficult to balance global and local features. In recent years, Vision Transformers have shown great potential in medical image segmentation due to their ability to capture long-range dependencies through self-attention mechanisms. Representative methods such as Swin-UNet (Cao et al. 2022), PVT (Wang et al. 2021), MaxViT (Tu et al. 2022), MERIT (Rahman and Marculescu 2024), and ConvFormer (Lin et al. 2023) have improved overall segmentation performance but still struggle with modeling fine-grained textures and edge details, and the development of large models has also provided insights for our research (Ma et al. 2024). Furthermore, recent developments in composed image retrieval, including ENCODER (Li et al. 2025a), FineCIR (Li et al. 2025b), OFFSET (Chen et al. 2025a), HUD (Chen et al. 2025b), PAIR (Fu et al. 2025), and MEDIAN (Huang et al. 2025), reveal that integrating structured reasoning with uncertainty modeling and hierarchical feature aggregation can significantly improve representation robustness and semantic interpretability, inspiring advances in medical image segmentation. Therefore, inspired by the aforementioned advances, we propose a novel decoder framework to address these challenges by preserving global perception while enhancing the representation of edges and structural details. This decoder consists of three key modules:

- ACFA: A direction-aware module that strengthens structural orientation and spatial consistency.

*Corresponding author.

- TFFA: A tri-branch fusion module integrating spatial, Fourier, and wavelet representations to balance global and local features.
- SMMM: A multi-scale skip-fusion module that suppresses redundancy, refines boundaries, and improves feature alignment for accurate and detailed segmentation.

## Related Works

### Vision Encoders

CNNs are widely used as encoders in medical image segmentation for their efficiency in extracting multi-level spatial and semantic features through local convolutions, weight sharing, and mature architectures. VGGNet (Simonyan and Zisserman 2014) showed that stacking small kernels with increased depth improves performance, while GoogLeNet (Szegedy et al. 2015) employs Inception modules and global average pooling to achieve efficient feature extraction with fewer parameters. ResNet (He et al. 2016) introduced residual connections to overcome degradation in deep networks, enabling deeper models with high efficiency. However, CNNs' fixed receptive fields limit their ability to capture long-range dependencies and global context, which is critical for complex shapes or blurred boundaries. Vision Transformers (ViTs) address this via self-attention, with variants like Swin Transformer (Liu et al. 2021) and PVT (Wang et al. 2021) enhancing global modeling. PVTv2 (Wang et al. 2022) further combines convolutions for local features, overlapping patch embedding, and linear attention with mean pooling. Despite improvements in global representation, these methods still struggle with modeling short-range dependencies effectively.

### Medical Image Segmentation

Medical image segmentation, a key task in medical image analysis, benefits significantly from deep learning's capability for powerful feature extraction and end-to-end modeling. Classic CNN-based frameworks such as U-Net (Ronneberger, Fischer, and Brox 2015), V-Net (Milletari, Navab, and Ahmadi 2016), SegNet (Badrinarayanan, Kendall, and Cipolla 2017), and FCN (Long, Shelhamer, and Darrell 2015) have laid the foundation for automated delineation of organs and lesions. U-Net's encoder–decoder architecture with skip connections enables multi-scale fusion, while V-Net extends this idea to 3D data, and SegNet and FCN improve spatial preservation and end-to-end prediction, respectively. Recent advances further enhance representational capacity: AD-LA Former (Wang, Wang, and Zhang 2025) combines dynamic convolutions and positional attention for complex structures, whereas EMCAD (Rahman, Munir, and Marculescu 2024) introduces a multi-scale attention decoder that strengthens salient region modeling through channel, spatial, and grouped gating attention. Transformer-based models such as TransUNet (Chen et al. 2021), Swin-UNet (Cao et al. 2022), and CSWin-UNet (Liu et al. 2025) extend long-range contextual modeling via self-attention mechanisms, while MISSFormer (Huang et al. 2022b) and LeViT-UNet (Xu et al. 2023) improve efficiency through multi-scale and lightweight design. Hybrid architectures that combine CNNs and Transformers—such as Hiformer (Heidari et al. 2023) and TBConvL-Net (Iqbal et al. 2025)—leverage the strengths of both paradigms to balance locality, global dependency, and computational cost. Beyond conventional segmentation models, progress in efficient learning (Yao, Li, and Xiao 2024), federated harmonization (Xiao et al. 2024), cross-modal contrastive learning (Wang et al. 2025), and diffusion-based self-supervision (Xiao et al. 2025) has inspired new directions for robust feature representation. Moreover, advancements in normalized 3D scene modeling (Yao et al. 2023) and multimodal reasoning (Zhang et al. 2025) provide insights into efficient information fusion and reasoning that could further benefit medical image segmentation.

## Methods

Our goal is to develop a novel medical image segmentation decoder that can effectively preserve the integrity of both local and global information while placing greater emphasis on key regional features. The following sections provide a detailed description of the decoder's structural design and functionality, as shown in the Figure 1, which consists of three core modules:

### Adaptive Cross-Fusion Attention

To enhance the model's responsiveness to key regions and its ability to model structural directions, we design an ACFA module with directional awareness. This module integrates channel enhancement and spatial modeling mechanisms while introducing learnable guidance along three directions (planar, vertical, and horizontal), enabling fine-grained feature enhancement across different scales and orientations. Specifically, for an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, channel and spatial gating are first applied to extract features:

$$\widehat{X}_{l-1}^{CG} = X \odot \delta\left(CG_{avg}\left(X\right) + CG_{\max}\left(X\right)\right) \tag{1}$$

$$\widehat{X}_{l-1}^{SG} = X \odot \delta\left(f_{7\times7}^{Conv}\left(SG\left(X\right)\right)\right) \tag{2}$$

Here, $CG_{avg}\left(\right)$ and $CG_{\max}\left(\right)$ represent average channel gating and maximum channel gating, respectively. $SG\left(\right)$ denotes spatial feature gating, $f_{7\times7}^{Conv}\left(\right)$ refers to a convolution operation with a $7 \times 7$ kernel, $\delta\left(\right)$ is the Sigmoid activation function, and $\odot$ indicates matrix multiplication. Subsequently, to obtain deeper-level features, the feature map is divided along the channel dimension into four subsets $\widehat{X}_{l-1}^{SG_1}, \widehat{X}_{l-1}^{SG_2}, \widehat{X}_{l-1}^{SG_3}, \widehat{X}_{l-1}^{SG_4}$, each of which is combined with learnable parameters corresponding to specific directions for fine-grained modulation. Depthwise separable convolutions are then applied to extract critical responses for each direction. Specifically, three branches introduce learnable weight parameters with directional guidance:

$$Tensor^{HW} = f_{[0,1]}^{Uniform}\left(Param\left[1, \frac{C}{4}, H, W\right]\right) \tag{3}$$

$$Tensor^{H} = f_{[0,1]}^{Uniform}\left(Param\left[1, \frac{C}{4}, H, 1\right]\right) \tag{4}$$

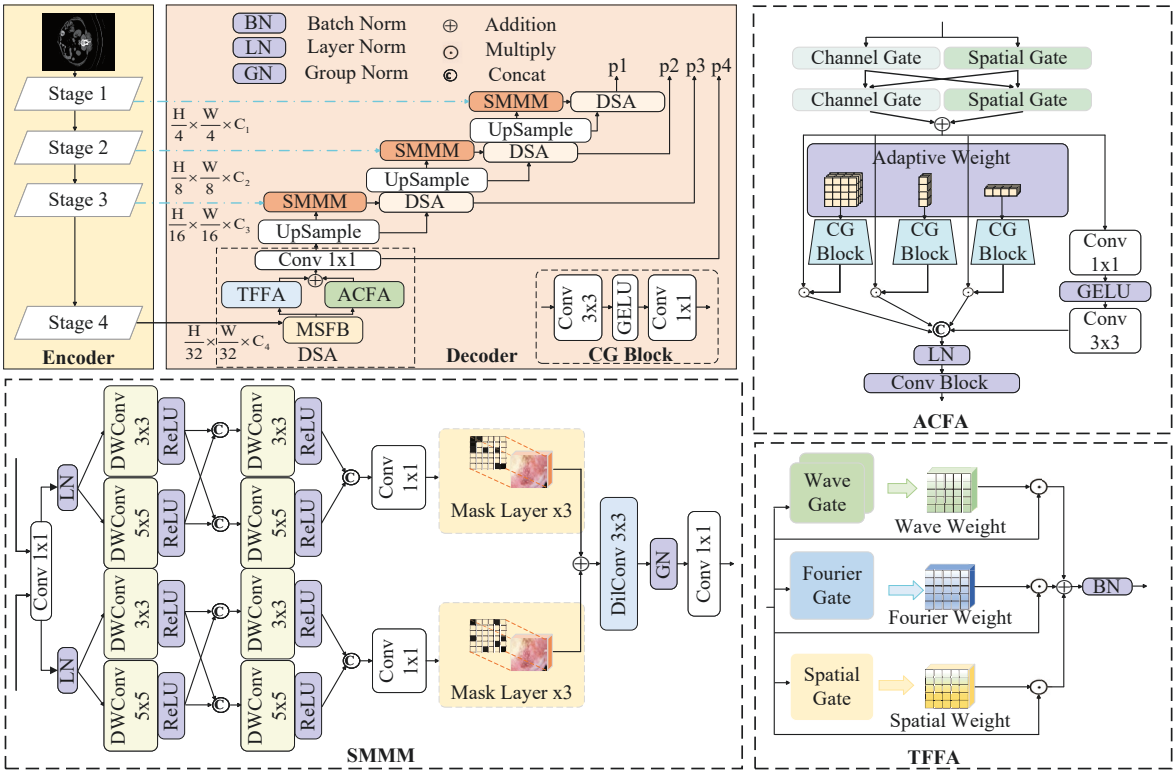


Figure 1: Proposed network architecture of the model.

$$Tensor^{W} = f_{[0,1]}^{Uniform}\left(Param\left[1, \frac{C}{4}, 1, W\right]\right) \quad (5)$$

Here, $f_{[0,1]}^{Uniform}()$ denotes initialization using a Uniform function, which sets the tensor values to random numbers within the range $[0, 1]$. These tensors are then processed through depthwise separable convolutions to further model the structural distribution patterns in different directions. The parameters are optimized in an end-to-end manner, enabling the model to automatically learn the most suitable directional attention patterns for the data distribution during training. Specifically:

$$\widehat{X}_{l-1}^{HW} = f_{3\times3}^{DWConv}\left(\vartheta\left(f_{1\times1}^{Conv}\left(Tensor^{HW}\right)\right)\right) \quad (6)$$

$$\widehat{X}_{l-1}^{H} = f_{3}^{DWConv1d}\left(\vartheta\left(f_{1}^{Conv1d}\left(Tensor^{H}\right)\right)\right) \quad (7)$$

$$\widehat{X}_{l-1}^{W} = f_{3}^{DWConv1d}\left(\vartheta\left(f_{1}^{Conv1d}\left(Tensor^{W}\right)\right)\right) \quad (8)$$

Here, $f_{3}^{DWConv1d}()$ and $f_{1}^{Conv1d}()$ represent 1D depthwise separable convolution with a kernel size of 3 and 1D convolution with a kernel size of 1, respectively, while $\vartheta()$ denotes the GELU activation function. To complement any detail information that might be missed in other directions, the fourth branch employs a set of standard convolution operations to capture more generalized contextual information, specifically:

$$\widehat{X}_{l-1}^{4} = f_{1\times1}^{Conv}\left(\vartheta\left(f_{3\times3}^{DWConv}\left(\widehat{X}_{l-1}^{SG_4}\right)\right)\right) \quad (9)$$

Finally, the features from the three directional branches and the fourth branch are concatenated, followed by LayerNorm and convolutional fusion to obtain the final direction-aware enhanced output.

## Triple Feature Fusion Attention

Traditional spatial convolution structures rely on local convolution kernels with fixed receptive fields, which can capture local texture information but are limited in modeling long-range dependencies, global structural relationships, and cross-scale semantic interactions. In particular, during feature fusion, simple spatial stacking or averaging operations, although preserving spatial distribution information, tend to cause semantic smoothing and redundancy, which can obscure key edges or subtle regions and weaken the model's sensitivity to details and boundaries. To address these limitations, we propose the TFFA module, which consists of three branches: a wavelet branch, a Fourier branch, and a spatial feature branch. In the wavelet branch, the module employs DoG and Mexican Hat wavelet functions for local spatio-frequency analysis. Both of these wavelet functions are classic filtering or edge-detection operators with strong spatial localization and frequency analysis capabilities. Compared to traditional wavelets such as Haar or Daubechies, they provide superior edge and texture representation. For the input tensor $X \in \mathbb{R}^{B\times C\times H\times W}$, the

| Methods | Average | | Spl | RKid | LKid | Gal | Liv | Sto | Aor | Pan |
|---|---|---|---|---|---|---|---|---|---|---|
| | DSC ↑ | HD95 ↓ | | | | | | | | |
| TransUNet | 77.49 | 31.69 | 85.08 | 77.02 | 81.87 | 63.16 | 94.08 | 75.62 | 87.23 | 55.86 |
| Swin-UNet | 79.13 | 21.55 | 90.66 | 79.61 | 83.28 | 66.53 | 94.29 | 76.60 | 85.47 | 56.58 |
| LeViT-UNet | 78.53 | 78.53 | 88.86 | 80.25 | 84.61 | 62.23 | 93.11 | 72.76 | 87.33 | 59.07 |
| MISSFormer | 81.96 | 18.20 | 91.92 | 82.00 | 85.21 | 68.65 | 94.41 | 80.81 | 86.99 | 65.67 |
| ScaleFormer | 82.86 | 16.81 | 89.40 | 83.31 | 86.36 | 74.97 | 95.12 | 80.14 | 88.73 | 64.85 |
| HiFormer-B | 80.39 | **14.70** | 90.99 | 79.77 | 85.23 | 65.23 | 94.61 | 81.08 | 86.21 | 59.52 |
| DAEFormer | 82.63 | 16.39 | 91.82 | 82.39 | 87.66 | 71.65 | 95.08 | 80.77 | 87.84 | 63.93 |
| PVT-CASCADE | 81.06 | 20.23 | 90.10 | 80.37 | 82.23 | 70.59 | 94.08 | 83.69 | 83.01 | 64.43 |
| LKA | 82.77 | 17.42 | 91.45 | 81.93 | 84.93 | 71.05 | 94.87 | 83.71 | 87.48 | 66.76 |
| EMCAD | 83.63 | 15.68 | 92.17 | 84.10 | 88.08 | 68.87 | **95.26** | **83.92** | 88.14 | 68.51 |
| AD-LA Former | 83.48 | 21.31 | 88.72 | 70.82 | 86.50 | **83.30** | 95.17 | 67.28 | **91.34** | **84.69** |
| CSWin-UNet | 81.12 | 18.86 | 89.05 | 78.53 | 83.51 | 67.85 | 95.23 | 81.74 | 87.13 | 65.94 |
| **Ours** | **83.92** | 18.91 | **92.46** | **86.47** | **89.26** | 67.51 | 94.72 | 83.33 | 87.63 | 69.95 |

Table 1: The comparison results of the model with previous methods on the Synapse dataset. Bold indicates the best result, dsc is presented for abdominal organs spleen (Spl), right kidney (Rkid), left kidney (Lkid), gallbladder (Gal), liver (Liv), stomach (Sto), aorta (Aor), and pancreas (Pan). the same below.

| Methods | DSC | SE | SP | ACC |
|---|---|---|---|---|
| TransUNet | 81.23 | 82.63 | 95.77 | 92.07 |
| Swin-Unet | 88.15 | 83.64 | **98.69** | 95.82 |
| LKA | 90.99 | 90.55 | 98.49 | 96.98 |
| Hiformer-B | 90.93 | 88.67 | 98.57 | 96.69 |
| EGE-Unet | 89.28 | 88.39 | 98.11 | 96.48 |
| UltraLightVM-UNet | 90.91 | 90.53 | 97.90 | 96.46 |
| EMCAD | 90.06 | **93.70** | 96.81 | 96.55 |
| AD-LA Former | 87.68 | 93.18 | 98.65 | 97.03 |
| TBConvL-Net | 90.89 | 91.19 | 97.61 | 96.07 |
| **Ours** | **91.40** | 92.75 | 97.78 | **97.26** |

Table 2: Comparison experiments on the ISIC 2017 dataset.

| Methods | DSC | SE | SP | ACC |
|---|---|---|---|---|
| TransUNet | 84.99 | 85.78 | 96.53 | 94.52 |
| Swin-Unet | 86.03 | 79.98 | 98.53 | 94.45 |
| LKA | 88.88 | 84.75 | 98.34 | 95.34 |
| Hiformer-B | 88.10 | 84.61 | 97.74 | 94.85 |
| EGE-Unet | 89.04 | 90.44 | 95.91 | 94.58 |
| VM-UNetV2 | 89.73 | 88.64 | 97.13 | 95.06 |
| UltraLightVM-UNet | 89.40 | 86.80 | 97.81 | 95.58 |
| EMCAD | 89.73 | 92.43 | 96.75 | 96.17 |
| DMSA-Unet | 90.63 | **97.14** | 97.14 | 96.16 |
| AD-LA Former | 85.91 | 91.24 | **99.04** | 96.46 |
| **Ours** | **90.71** | 93.34 | 96.18 | **96.62** |

Table 3: Comparison experiments on the ISIC 2018 dataset.

wavelet transform is applied as follows:

$$W(a,b) = \int_{-\infty}^{\infty} x(t)\,\psi_{a,b}(t)dt \tag{10}$$

where $\psi_{a,b}()$ denotes the mother wavelet function, instantiated with different wavelet types. DoG (Difference of Gaussians) is defined as the difference between two Gaussian-blurred images of different scales:

$$Dog(x,y) = G_{\sigma_1}(x,y) - G_{\sigma_2}(x,y), \sigma_1 < \sigma_2 \tag{11}$$

It emphasizes image details within specific frequency bands, serving as a band-pass filter approximating LoG. By adjusting the Gaussian kernel parameter $\sigma$, specific texture scales can be enhanced. Medical images often have low contrast and blurred textures; DoG can highlight areas with significant grayscale changes, enhancing edge and contour perception. Its formula is:

$$\psi_{a,b}^{Dog}(x) = -\frac{1}{\sqrt{a}}\left(\frac{x-b}{a}\right)e^{-\frac{\left(\frac{x-b}{a}\right)^2}{2}} \tag{12}$$

Here, a is the scale parameter and b is the shift parameter, both of which are learnable, enabling dynamic fusion of multi-wavelet features. Mexican Hat, in contrast, detects edge zero-crossings by taking the second derivative after smoothing, while effectively suppressing noise interference:

$$\psi_{a,b}^{MH}(x) = \frac{2}{\sqrt{3a}\pi^{\frac{1}{4}}}\left(1-\left(\frac{x-b}{a}\right)^2\right)e^{-\frac{\left(\frac{x-b}{a}\right)^2}{2}} \tag{13}$$

The wavelet transform provides localized multi-scale information, whereas the Fourier transform is inherently global, capturing overall structures and modeling long-range dependencies, thus compensating for the limited perception of large-scale structures in convolutional models. It transforms the image from the spatial domain to the frequency domain, representing periodic structures with sine and cosine waves. The frequency-domain features are modulated by learnable weight matrices, where high-frequency components correspond to edges and textures, and low-frequency components represent contours and backgrounds. Its formula is:

$$F(u,v) = \iint f(x,y)\,e^{-j2\pi(ux+vy)}dxdy \tag{14}$$

In the spatial branch, pointwise convolution is applied to extract spatial features. Unlike simple channel stacking or

averaging, TFFA introduces an attention gating mechanism after the three branches, assigning dynamic weights to spatial, frequency, and wavelet outputs to achieve adaptive fusion. This mechanism avoids over-smoothing caused by traditional fusion strategies and enhances the model's response to salient regions in complex scenarios. The final fused result is processed with batch normalization and activation functions to improve feature stability and nonlinear modeling capability, thereby achieving superior performance in tasks involving complex boundaries and fine-grained segmentation.

### Structural-Aware Multi-Scale Masking Module

In medical image segmentation, skip connections are essential for linking multi-scale features between the encoder and decoder and mitigating spatial detail loss during upsampling. However, traditional skip connections often use simple feature addition, lacking selective fusion and leading to redundant or irrelevant information. To overcome this limitation, we propose SMMM, a structure-aware multi-scale fusion strategy that strengthens feature representation and discrimination. Specifically, for encoder features $X \in \mathbb{R}^{B\times C\times H\times W}$ and decoder features $Y \in \mathbb{R}^{B\times C\times H\times W}$ of the same shape, both are first processed by parallel pointwise convolutions to activate spatial cues, followed by multi-scale perception modules. Each module integrates depthwise separable convolutions with kernel sizes ($3\times3, 5\times5$), combined with a two-stage channel split and ReLU activation, effectively enlarging the receptive field and improving the ability to capture complex structural boundaries. The detailed formulation is as follows:

$$\widehat{X}_M = f_{1\times1}^{Conv}\left(X\right) \tag{15}$$

$$\widehat{X}_{M-1}^{S_1} = \gamma\left(f_{3\times3}^{DWConv}\left(\widehat{X}_M\right)\right) \tag{16}$$

$$\widehat{X}_{M-1}^{S_2} = \gamma\left(f_{5\times5}^{DWConv}\left(\widehat{X}_M\right)\right) \tag{17}$$

$$\widehat{X}_M^{S_1} = \gamma\left(f_{3\times3}^{DWConv}\left(Cat\left(\widehat{X}_{M-1}^{S_1}, \widehat{X}_{M-1}^{S_2}\right)\right)\right) \tag{18}$$

$$\widehat{X}_M^{S_2} = \gamma\left(f_{5\times5}^{DWConv}\left(Cat\left(\widehat{X}_{M-1}^{S_1}, \widehat{X}_{M-1}^{S_2}\right)\right)\right) \tag{19}$$

$$\widehat{X} = f_{1\times1}^{Conv}\left(Cat\left(\widehat{X}_M^{S_1}, \widehat{X}_M^{S_2}\right)\right) \tag{20}$$

where $\gamma()$ denotes the ReLU activation function. After the multi-scale feature fusion module, the encoder and decoder features are fed into a masking module for spatial saliency modeling. This module applies three different channel-gating filters to identify the most discriminative regions in the spatial domain, and uses a Softmax activation function to implement a weighted strategy that emphasizes high-response areas. This saliency design effectively handles challenges such as blurred lesions and indistinct contours in medical images. Next, the filtered features are added together and further fused using dilated convolution with a dilation rate of 2. This expands the receptive field, improves the capture of lesion shapes and structural boundaries, and introduces richer contextual information without compressing the feature map resolution. Finally, the fused features are processed through a normalization layer and a pointwise convolution for channel alignment and stabilization of feature distributions, preventing gradient vanishing and improving training stability.

| Methods | DSC | RV | Myo | LV |
|---|---|---|---|---|
| Swin-UNet | 88.07 | 85.77 | 84.42 | 94.03 |
| TransUnet | 89.71 | 86.67 | 87.27 | 95.18 |
| Cascaded MERIT | 91.85 | 90.23 | 89.53 | 95.80 |
| PVT-GCASCADE | 91.95 | 90.31 | 89.63 | 95.91 |
| DMSA-UNet | 92.28 | 90.32 | 90.49 | 96.02 |
| EMCAD | 92.12 | 90.65 | 89.68 | 96.02 |
| AD-LA Former | 90.09 | 88.68 | 88.94 | 95.30 |
| CSWin-UNet | 91.46 | 89.68 | 88.94 | 95.76 |
| **Ours** | **92.75** | **91.18** | **90.40** | **96.67** |

Table 4: Comparison experiments on the ACDC dataset.

## Experiments

### Experimental Setup

We implemented our model using PyTorch 1.11.0 and conducted all experiments on a single NVIDIA A100 GPU with 40GB memory. Following the experimental settings of EMCAD, we adopted PVTv2-b2 pre-trained on ImageNet as the encoder. The learning rate and weight decay were set to 1e-4, the AdamW optimizer was used during training, normalized masks [0,1], and no augmentation. The batch size was fixed at 12. We trained the model for 200 epochs on the ISIC 2017 and ISIC 2018 datasets, and for 300 and 400 epochs on the Synapse and ACDC datasets, respectively.

### Datasets

**Skin Lesion Segmentation** We used the ISIC 2017 and ISIC 2018 datasets, released by the International Skin Imaging Collaboration. ISIC 2017 contains about 2,000 high-resolution dermoscopic images with precise lesion annotations, mainly for boundary segmentation evaluation. ISIC 2018 expands to 2,594 images with more lesion types, providing a benchmark for multi-class segmentation and generalization assessment.

**Synapse Multi-Organ Segmentation** The Synapse dataset, from the MICCAI 2015 Multi-Atlas Labeling Beyond the Cranial Vault challenge, includes 30 abdominal CT scans with pixel-level annotations of eight organs, such as the liver, pancreas, kidneys, and spleen. Its standardized imaging and high-quality masks make it a widely used benchmark for multi-organ segmentation, cross-structure recognition, and generalization studies.

**ACDC Dataset** The ACDC dataset, part of the MICCAI 2017 cardiac segmentation challenge, comprises cardiac MRI cine sequences from 100 subjects, including healthy and pathological cases (e.g., dilated and hypertrophic cardiomyopathy). It provides precise annotations for the left and right ventricles and myocardium, serving as a benchmark for cardiac structure segmentation and functional analysis.

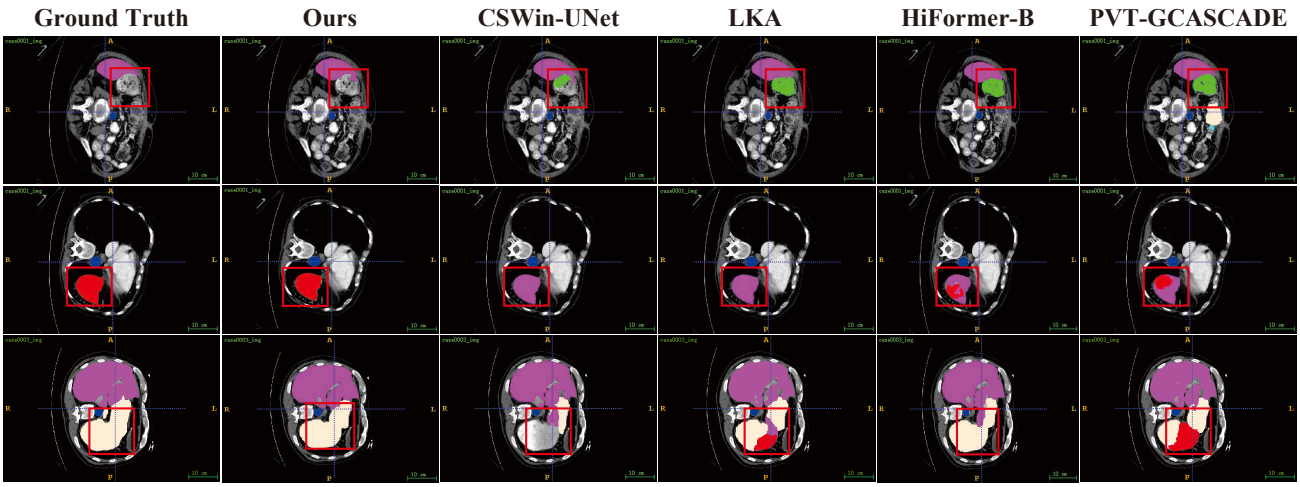


Figure 2: Comparison of the performance of our model with the visualization results of other models on the Synapse dataset.

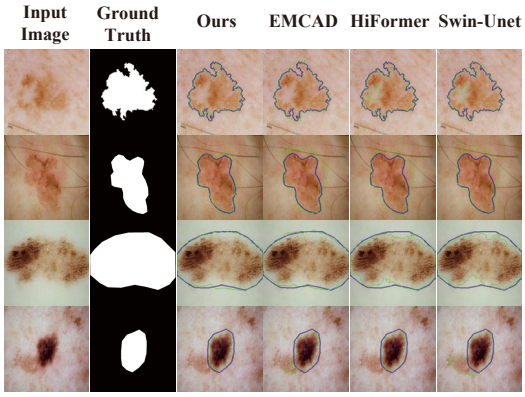


Figure 3: Comparison of segmentation results on the ISIC 2017 dataset with other previous methods.

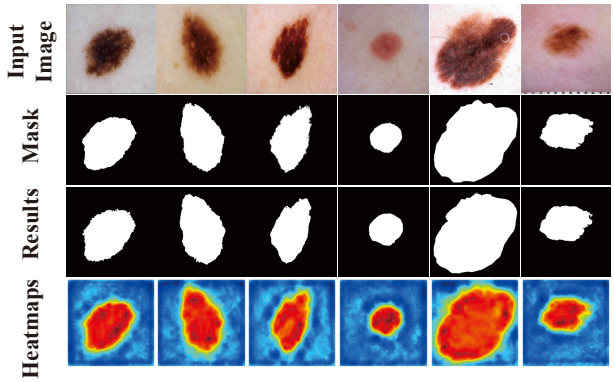


Figure 4: Attention heatmap visualization for the proposed model on ISIC 2018.

## Quantitative and Qualitative Results

On the Synapse dataset (Table 1), our model achieved an average DSC of 83.92%, surpassing state-of-the-art methods such as EMCAD (83.63%) and AD-LA Former (83.48%), demonstrating a clear advantage in multi-organ segmentation. It achieved the best performance on key organs including the spleen (92.46%), right kidney (86.47%), and left kidney (89.26%), highlighting its ability to capture complex organ boundaries and fine-grained structures. Although AD-LA Former obtained the highest score on the gallbladder (83.30%), its overall average remains lower than ours. We additionally compared our model with other recent methods, including ScaleFormer (Huang et al. 2022a), DAEFormer (Azad et al. 2023), PVT-CASCADE (Rahman and Marculescu 2023a), LKA (Guo et al. 2023), EGE-Unet (Ruan et al. 2023), UltraLightVM-UNet (Wu et al. 2024), Cascaded MERIT (Rahman and Marculescu 2023b), VM-UNetV2 (Zhang et al. 2024), and DMSA-UNet (Li et al. 2024) across Synapse, ISIC 2017, ISIC 2018, and ACDC datasets. On ISIC 2017 (Table 2), our model achieved the highest DSC (91.40%) and ACC (97.26%), outperforming EMCAD (90.06%) and LKA (90.99%), while maintaining a balanced sensitivity and specificity. On ACDC (Table 4), it achieved an average DSC of 92.75%, with RV, Myo, and LV scores of 91.18%, 90.40%, and 96.67%, exceeding DMSA-UNet (92.28%) and Cascaded MERIT (91.85%). On ISIC 2018 (Table 3), it reached the best DSC (90.71%) and ACC (96.62%) and a competitive SE (93.34%), second only to DMSA-UNet (97.14%). These comprehensive results, further supported by detailed visualization comparisons (Figures 2 and 3), confirm the robustness, structural awareness, and strong cross-task generalization of our model across multi-organ, cardiac, and skin lesion segmentation tasks, underscoring its effectiveness in handling diverse and challenging medical imaging scenarios.

| Methods | Average | | Aor | Gal | LKid | RKid | Liv | Pan | Spl | Sto |
|---|---|---|---|---|---|---|---|---|---|---|
| | DSC ↑ | HD95 ↓ | | | | | | | | |
| Baseline | 81.35 | 20.42 | 86.78 | 64.11 | 87.48 | 83.96 | 93.82 | 62.00 | 89.88 | 82.77 |
| Baseline+ACFA | 82.04 | 21.46 | 86.54 | 65.73 | 87.69 | 84.27 | 93.38 | 65.17 | 90.12 | 83.42 |
| Baseline+ACFA+TFFA | 83.23 | 16.72 | 86.05 | 64.43 | 87.89 | 86.86 | 94.58 | 70.99 | 91.07 | 83.97 |
| Baseline+ACFA+TFFA+SMMM | 83.92 | 18.91 | 87.63 | 67.51 | 89.26 | 86.47 | 94.72 | 69.95 | 92.46 | 83.33 |

Table 5: Ablation study on the impact of different modules on the model on the Synapse dataset.

| Method | DSC | SE | SP | ACC |
|---|---|---|---|---|
| Baseline | 85.95 | 83.68 | 98.96 | 96.05 |
| Baseline+ACFA | 87.82 | 85.12 | 98.16 | 95.92 |
| Baseline+ACFA+TFFA | 89.15 | 89.83 | 97.29 | 96.85 |
| Baseline+ACFA+TFFA+SMMM | 91.40 | 92.75 | 97.78 | 97.26 |

Table 6: Study on the ablation of the impact of different modules on the ISIC 2017 dataset.

| Fourier | Wavelet | | DSC | SE | SP | ACC |
|---|---|---|---|---|---|---|
| | Mexican Hat | DoG | | | | |
| NO | NO | NO | 90.32 | 89.31 | 97.73 | 95.37 |
| YES | NO | NO | 90.48 | 91.43 | 97.27 | 95.74 |
| YES | YES | NO | 90.57 | 92.63 | 96.41 | 96.15 |
| YES | YES | YES | 90.71 | 93.34 | 96.18 | 96.62 |

Table 7: Ablation study on the effects of different components within the TFFA module.

## Ablation Studies

To validate the effectiveness of the proposed decoder, we conducted ablation experiments on the Synapse and ISIC 2017 datasets, as shown in Tables 5, 6, 7and 8.

### Effect of Different Components on Synapse

Results on the Synapse dataset confirm that each component contributes to performance improvement. Introducing ACFA increased parameters by only 5.6M and computation by 1.5 GMac, while boosting DSC to 82.04%. This demonstrates that directional awareness and channel–spatial fusion strengthen structural perception. Incorporating TFFA further improved DSC to 83.2% and reduced HD95 to 16.72, indicating that joint spatial–frequency modeling complements global semantics and local boundary learning. When all modules were integrated, the decoder achieved 83.92% DSC with 42.52M parameters and 18.29 GMac, delivering the best segmentation of complex organs such as the liver and kidneys while maintaining efficiency.

### Effect of Different Components on ISIC 2017

We also conducted ablation studies on the ISIC 2017 dataset. The baseline achieved 85.95% DSC. Adding ACFA raised it to 87.82%, highlighting enhanced directional and regional discrimination. When TFFA was incorporated, DSC reached 89.15% and SE increased to 89.83%, verifying that spatial–frequency fusion improves edge and texture representation. Integrating SMMM yielded the best performance (DSC 91.4%, SE 92.75%, ACC 97.26%), demonstrating that joint multi-module learning enhances lesion delineation and boundary accuracy.

| ACFA | TFFA | TFFA | Params (M) | Complexity (GMac) |
|---|---|---|---|---|
| NO | NO | NO | 25.07 | 11.85 |
| YES | NO | NO | 30.67 | 13.35 |
| YES | YES | NO | 32.01 | 13.75 |
| YES | YES | YES | 42.52 | 18.29 |

Table 8: Computational cost and complexity of modules on synapse dataset.

### Attention Heatmap Analysis on the ISIC 2018 Dataset

Attention heatmaps on the ISIC 2018 dataset (Fig. 4) show that the proposed decoder captures lesion boundaries, textures, and internal structures effectively. ACFA strengthens directional edge awareness, TFFA balances global and local cues through Fourier–wavelet fusion, and SMMM alleviates spatial detail loss during feature aggregation. Their synergy enables precise focus on lesion regions, smooth boundary prediction, and improved generalization.

### Internal Ablation Study of the TFFA Module

We conducted an internal analysis of the TFFA module and validated the contribution of each frequency component. Fourier modeling enhances global structural representation, while adding the Mexican Hat wavelet improves local edge sensitivity. Combining both DoG and Mexican Hat wavelets achieved the best results (DSC 90.71%, SE 93.34%), confirming that multi-scale spatial–frequency fusion effectively refines boundary perception.

## Conclusion

We propose a novel decoder for medical image segmentation that addresses challenges in edge detail modeling, long-range dependency capture, and multi-scale feature fusion. The framework integrates ACFA for directional and structural awareness, TFFA for joint Wavelet–Fourier–Spatial feature modeling, and SMMM for multi-scale skip fusion with saliency masking. Overall, the decoder achieves significant segmentation accuracy gains through joint directional–frequency–structural modeling, providing an effective and practical solution for high-precision medical image segmentation.

## Ethical Statement

All datasets used in this study are publicly available and de-identified. Our experiments comply with the terms of use provided by the data providers, ensuring no personally identifiable information is disclosed. This research adheres to standard ethical guidelines for the use of medical imaging data.

## Acknowledgements

This work was supported in part by the following: the National Natural Science Foundation of China under Grant Nos. U24A20219, 62272281, U24A20328, U22A2033, 62576193, the Special Funds for Taishan Scholars Project under Grant Nos. tsqn202306274, tsqn202507240, the Yantai Natural Science Foundation under Grant No. 2024JCYJ034, the Youth Innovation Technology Project of Higher School in Shandong Province under Grant No. 2023KJ212, and the Natural Science Foundation of Shandong Province under grant No. ZR20250C712.

## References

Azad, R.; Arimond, R.; Aghdam, E. K.; Kazerouni, A.; and Merhof, D. 2023. Dae-former: Dual attention-guided efficient transformer for medical image segmentation. In *International Workshop on PRedictive Intelligence In MEdicine*, 83–95. Springer.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2481–2495.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, Z.; Hu, Y.; Li, Z.; Fu, Z.; Song, X.; and Nie, L. 2025a. OFFSET: Segmentation-based Focus Shift Revision for Composed Image Retrieval. In *Proceedings of the ACM International Conference on Multimedia*, 6113–6122.

Chen, Z.; Hu, Y.; Li, Z.; Fu, Z.; Wen, H.; and Guan, W. 2025b. HUD: Hierarchical Uncertainty-Aware Disambiguation Network for Composed Video Retrieval. In *Proceedings of the ACM International Conference on Multimedia*, 6143–6152.

Fu, Z.; Li, Z.; Chen, Z.; Wang, C.; Song, X.; Hu, Y.; and Nie, L. 2025. PAIR: Complementarity-guided Disentanglement for Composed Image Retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S.-M. 2023. Visual attention network. *Computational Visual Media*, 9(4): 733–752.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heidari, M.; Kazerouni, A.; Soltany, M.; Azad, R.; Aghdam, E. K.; Cohen-Adad, J.; and Merhof, D. 2023. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 6202–6212.

Huang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; and Tong, R. 2022a. ScaleFormer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. *arXiv preprint arXiv:2207.14552*.

Huang, Q.; Chen, Z.; Li, Z.; Wang, C.; Song, X.; Hu, Y.; and Nie, L. 2025. MEDIAN: Adaptive Intermediate-grained Aggregation Network for Composed Image Retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Huang, X.; Deng, Z.; Li, D.; Yuan, X.; and Fu, Y. 2022b. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5): 1484–1494.

Iqbal, S.; Khan, T. M.; Naqvi, S. S.; Naveed, A.; and Meijering, E. 2025. TBConvL-Net: A hybrid deep learning architecture for robust medical image segmentation. *Pattern Recognition*, 158: 111028.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, X.; Fu, C.; Wang, Q.; Zhang, W.; Sham, C.-W.; and Chen, J. 2024. DMSA-UNet: Dual Multi-Scale Attention makes UNet more strong for medical image segmentation. *Knowledge-Based Systems*, 299: 112050.

Li, Z.; Chen, Z.; Wen, H.; Fu, Z.; Hu, Y.; and Guan, W. 2025a. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5101–5109.

Li, Z.; Fu, Z.; Hu, Y.; Chen, Z.; Wen, H.; and Nie, L. 2025b. FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval. *https://arxiv.org/abs/2503.21309*.

Lin, X.; Yan, Z.; Deng, X.; Zheng, C.; and Yu, L. 2023. ConvFormer: Plug-and-play CNN-style transformers for improving medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 642–651. Springer.

Liu, X.; Gao, P.; Yu, T.; Wang, F.; and Yuan, R.-Y. 2025. CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation. *Information Fusion*, 113: 102634.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Rahman, M. M.; and Marculescu, R. 2023a. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6222–6231.

Rahman, M. M.; and Marculescu, R. 2023b. Medical Image Segmentation via Cascaded Attention Decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6222–6231.

Rahman, M. M.; and Marculescu, R. 2024. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Medical Imaging with Deep Learning*, 1526–1544. PMLR.

Rahman, M. M.; Munir, M.; and Marculescu, R. 2024. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11769–11779.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Ruan, J.; Xie, M.; Gao, J.; Liu, T.; and Fu, Y. 2023. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, 481–490. Springer.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, 459–479. Springer.

Wang, S.; Li, Z.; Li, Y.; Xiao, C.; Zhan, H.; Yao, Z.; Zhang, X.; Kang, J.; Li, L.; Liu, W.; et al. 2025. C3-OWD: A Curriculum Cross-modal Contrastive Learning Framework for Open-World Detection. *arXiv preprint arXiv:2509.23316*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3): 415–424.

Wang, Y.; Wang, H.; and Zhang, F. 2025. A Medical image segmentation model with auto-dynamic convolution and location attention mechanism. *Computer Methods and Programs in Biomedicine*, 261: 108593.

Wu, R.; Liu, Y.; Liang, P.; and Chang, Q. 2024. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*.

Xiao, C.; Hou, L.; Fu, L.; and Chen, W. 2025. Diffusion-Based Self-Supervised Imitation Learning from Imperfect Visual Servoing Demonstrations for Robotic Glass Installation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 10401–10407. IEEE.

Xiao, C.; et al. 2024. Confusion-resistant federated learning via diffusion-based data harmonization on non-IID data. *Advances in Neural Information Processing Systems*, 37: 137495–137520.

Xu, G.; Zhang, X.; He, X.; and Wu, X. 2023. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 42–53. Springer.

Yao, J.; Li, C.; Sun, K.; Cai, Y.; Li, H.; Ouyang, W.; and Li, H. 2023. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9421–9431. IEEE Computer Society.

Yao, J.; Li, C.; and Xiao, C. 2024. Swift sampler: Efficient learning of sampler by 10 parameters. *Advances in Neural Information Processing Systems*, 37: 59030–59053.

Zhang, M.; Yu, Y.; Jin, S.; Gu, L.; Ling, T.; and Tao, X. 2024. VM-UNET-V2: rethinking vision mamba UNet for medical image segmentation. In *International Symposium on Bioinformatics Research and Applications*, 335–346. Springer.

Zhang, X.; Zeng, F.; Quan, Y.; Hui, Z.; and Yao, J. 2025. Enhancing multimodal large language models complex reason via similarity computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10203–10211.