Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DDSF-Net: A Dual-Domain Synergistic Fusion Network with Spatial-Frequency Interaction for Medical Image Segmentation

Yuliang Ma

mayuliang@hdu.edu.cn

Hangzhou Dianzi University

Tianqi Chen

Hangzhou Dianzi University

Lijun Liang

Hangzhou Dianzi University

Qingshan She

Hangzhou Dianzi University

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11671075/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DDSF-Net: A Dual-Domain Synergistic Fusion Network with Spatial-Frequency Interaction for Medical Image Segmentation

Yuliang Ma[1,2*], Tianqi Chen[1], Lijun Liang[1], Qingshan She[1,2]

[1]School of Automation, Hangzhou Dianzi University, Hangzhou, 310018, Zhejiang, China.
[2]Zhejiang Provincial Key Laboratory of Brain Computer Collaborative Intelligence Technology and Applications, Hangzhou, 310018, China.

*Corresponding author(s). E-mail(s): mayuliang@hdu.edu.cn;
Contributing authors: tianqi.chen@hdu.edu.cn; llj@hdu.edu.cn; qsshe@hdu.edu.cn;

Abstract

Medical image segmentation remains challenging when blurred boundaries, heterogeneous textures, and cross-scale anatomical variation must be handled simultaneously. Although hybrid CNN-Transformer models improve global context modeling, their skip fusion and decoder reconstruction stages often treat spatial structure and frequency cues in a coupled and content-agnostic manner. To address this limitation, we propose DDSF-Net, a U-shaped segmentation network that combines dual-domain skip fusion with dynamic upsampling. A Dual-Domain Feature Extraction Module (DDFEM) contains a spatial structural path (SSP) and a frequency texture path (FTP), together with internal gating and fusion operations. Boundary-Residual Synergistic Skip (BRS-Skip) connections apply DDFEM to encoder and decoder features, then integrate a boundary-feature branch and encoder-residual mixing. In the decoder, a Dynamic Spatial-adaptive Upsampler (DSAU) performs content-aware offset sampling to better preserve boundary details during resolution recovery. Experiments on the Synapse, ACDC, and ISIC datasets show that DDSF-Net consistently improves the CSWin-UNet baseline and achieves competitive performance across different imaging modalities. On Synapse, DDSF-Net reaches 82.37% Dice and 17.91 mm HD95. Ablation studies further indicate that the combination of dual-domain skip fusion and dynamic upsampling contributes to the final gains.

Keywords: Medical image segmentation , Transformer , Dual-Domain Fusion , Adaptive Feature Fusion

1 Introduction

Medical image segmentation supports quantitative analysis and treatment planning, but low contrast, heterogeneous texture, and variable anatomy remain challenging [1]. U-Net and its successors combine encoder context with decoder detail through skip connections [2–4]. Transformer-based and hybrid architectures extend contextual modeling through self-attention [5–9]. However, generic skip fusion and fixed upsampling can limit the interaction between local structure, texture, and semantic context.

DDSF-Net revisits these two components of a CSWin-based U-shaped network. DDFEM processes features through spatial and frequency branches, followed by internal gating and fusion. BRS-Skip applies DDFEM to both encoder and decoder streams, then integrates local boundary features and an encoder residual. DSAU uses learned sampling offsets for resolution recovery. Experiments evaluate the combined design on

Synapse, ACDC, and ISIC, with ablations separating the contributions of spatial processing, frequency processing, and upsampling.

2 Related work

CNN-based segmentation methods, including FCN, U-Net, and DeepLab, use hierarchical features and local convolutional structure [2, 10, 11]. TransUNet, Swin-UNet, and CSWin-UNet introduce attention-based context modeling within encoder-decoder pipelines [9, 12, 13]. Attention gating and channel recalibration further improve feature selection [4, 14, 15]. DDSF-Net retains the CSWin backbone and focuses on skip integration and reconstruction.

Frequency-aware approaches provide complementary representations of texture and signal composition [16–20]. Their benefit is not automatic when multi-scale spatial filtering is already available. We therefore treat FTP as a complementary branch and assess its empirical contribution alongside SSP. Our claims are restricted to the evaluated methods, datasets, and settings.

3 Method

3.1 Network overview and notation

DDSF-Net combines a hierarchical CSWin encoder, a progressive decoder, DDFEM blocks, and Boundary-Residual Synergistic Skip (BRS-Skip) connections. DDFEM contains SSP and FTP branches, followed by internal gating, concatenation, and channel mixing. BRS-Skip applies separate DDFEM blocks to encoder and decoder features, followed by feature fusion, a boundary-feature branch, and encoder-residual mixing.

Let $I \in \mathbb{R}^{B\times C_{\text{in}}\times H\times W}$ denote the input image batch. Encoder stage l produces $E_l \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$, where $l = 1,\dots,4$, $C_l = 2^{l-1}C$, and $(H_l, W_l) = (H/2^{l+1}, W/2^{l+1})$. The supplied skip implementation uses $C = 64$. All equations use image-format tensors; token-to-image reshaping is omitted.

DDFEM blocks enhance the deepest encoder output and the decoder entrance. For $l = 3, 2, 1$, DSAU upsamples D_{l+1} to U_l, BRS-Skip combines E_l and U_l into S_l, and CSWin blocks produce D_l. The final decoder feature is normalized, upsampled by four, and projected to segmentation logits. Throughout this section, brackets denote channel concatenation, $\odot$ denotes broadcast multiplication, and δ denotes GELU.

$$U_l = \mathcal{U}_{2,l}(D_{l+1}), \quad S_l = \mathcal{B}_l(E_l, U_l), \quad D_l = \mathcal{T}_l(S_l). \tag{1}$$

3.2 Dual-Domain Feature Extraction Module (DDFEM)

For an input X and an even target channel count C_o, SSP and FTP each generate $C_o/2$ channels. DDFEM internally gates and concatenates these branch features, then returns a C_o-channel representation:

$$\begin{aligned} F_s &= \mathcal{S}(X), \qquad F_f = \mathcal{F}(X; K), \\ Y &= \mathcal{D}(X) = \mathcal{G}(F_s, F_f). \end{aligned} \tag{2}$$

The operators $\mathcal{S}$ and $\mathcal{F}$ denote SSP and FTP, respectively. The operator $\mathcal{G}$ denotes the internal gating and fusion operations of DDFEM, rather than a separate module. Different DDFEM instances use separate parameters.

3.2.1 Spatial Structural Path (SSP)

SSP combines three grouped convolution branches. A global descriptor produces Softmax weights, and a grouped 1×1 modulation branch is added to the weighted sum:

$$\begin{aligned} F_s &= \sum_{j=1}^{3} w_j(X) \odot \mathcal{C}_j(X) + M_s(X), \\ \mathbf{w} &= \text{Softmax}(\text{Conv}(\text{GAP}(X))), \\ M_s(X) &= \delta(\text{BN}(\text{GConv}_{1\times1}(X))). \end{aligned} \tag{3}$$

The three branches and the modulation path are shown in Fig. 2. The branch weights sum to one; the implementation uses adaptive weighting rather than runtime prediction of rectangular kernel sizes.

3.2.2 Frequency Texture Path (FTP)

FTP first separates low- and high-frequency responses using a real 2-D Fourier transform and a low-pass mask. The high-frequency response is

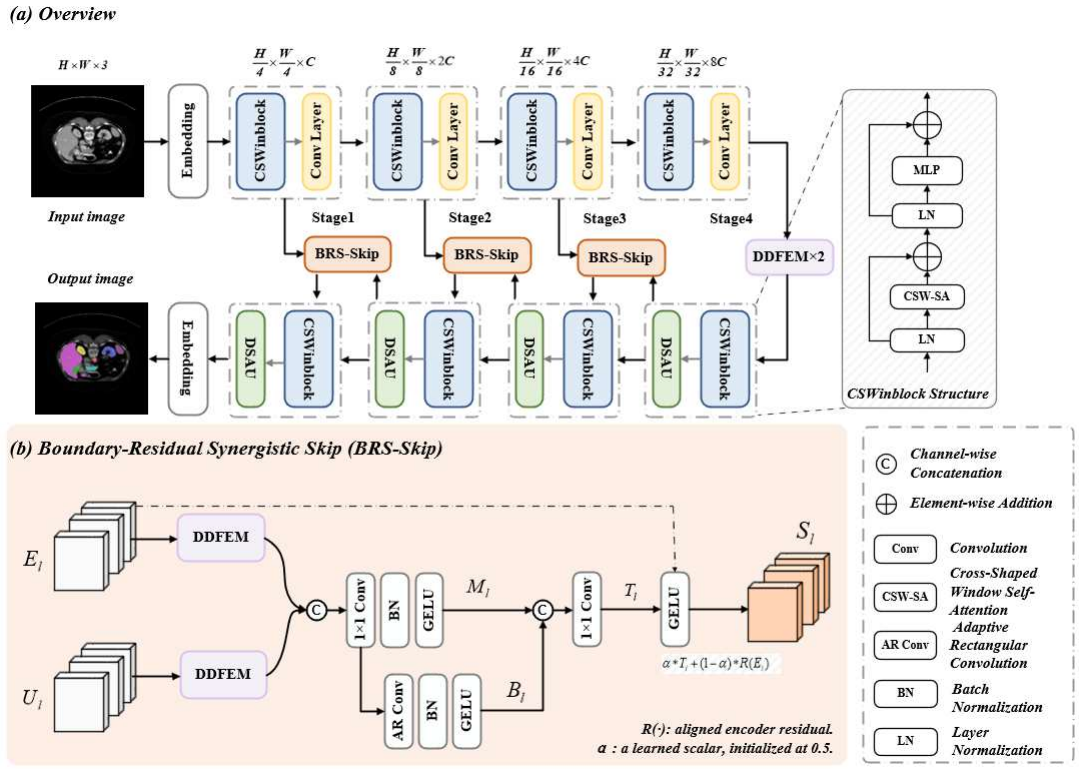

Fig. 1: Architecture of DDSF-Net. (a) The four-stage CSWin encoder and progressive decoder exchange features through BRS-Skip connections, while DSAU restores spatial resolution. The inset shows the CSWin block. (b) BRS-Skip applies separate DDFEM instances to encoder feature E_l and upsampled decoder feature U_l. Their outputs form M_l, which is integrated with boundary features B_l and projected to T_l. Learned residual mixing produces S_l.

spatially modulated and then passed through the input-conditioned dynamic convolution:

$$\begin{aligned} L &= \mathrm{irFFT2}(\mathrm{rFFT2}(X) \odot \mathcal{M}_K), \\ H &= X - L, \quad X_{\mathrm{fb}} = A(X) \odot H + L, \\ F_f &= \mathrm{Conv}(X_{\mathrm{fb}}; W_{\mathrm{dyn}}(X)). \end{aligned} \tag{4}$$

Here, $A(X) = 2\,\mathrm{sigmoid}(\mathrm{GConv}_{3\times3}(X))$, and $\mathcal{M}_K$ denotes the selected low-pass mask. For the configuration shown in Fig. 2, $K = [2]$.

3.2.3 Internal gating and fusion

DDFEM concatenates F_s and F_f, predicts two independent gates, and mixes the gated branches with 1×1 convolution, batch normalization, and GELU:

$$\begin{aligned} [a_s, a_f] &= \mathrm{sigmoid}(\mathrm{MLP}(\mathrm{GAP}([F_s, F_f]))), \\ Y &= \delta(\mathrm{BN}\,(\mathrm{Conv}_{1\times1}([a_s F_s, a_f F_f]))). \end{aligned} \tag{5}$$

The MLP uses two pointwise convolutions separated by ReLU. The gates are independent and need not sum to one.

3.3 Boundary-Residual Synergistic Skip (BRS-Skip)

At levels $l = 1, 2, 3$, BRS-Skip processes E_l and spatially aligned U_l using separate DDFEM instances. Their outputs are concatenated and projected through 1×1 convolution, BN, and GELU to form M_l. A lightweight spatial branch extracts $C_l/4$-channel boundary features B_l from M_l. Concatenating M_l and B_l, followed by 1×1 projection, produces C_l-channel feature T_l. Encoder-residual mixing yields

$$S_l = \delta(\lambda_l T_l + (1 - \lambda_l) R_l(E_l)). \tag{6}$$

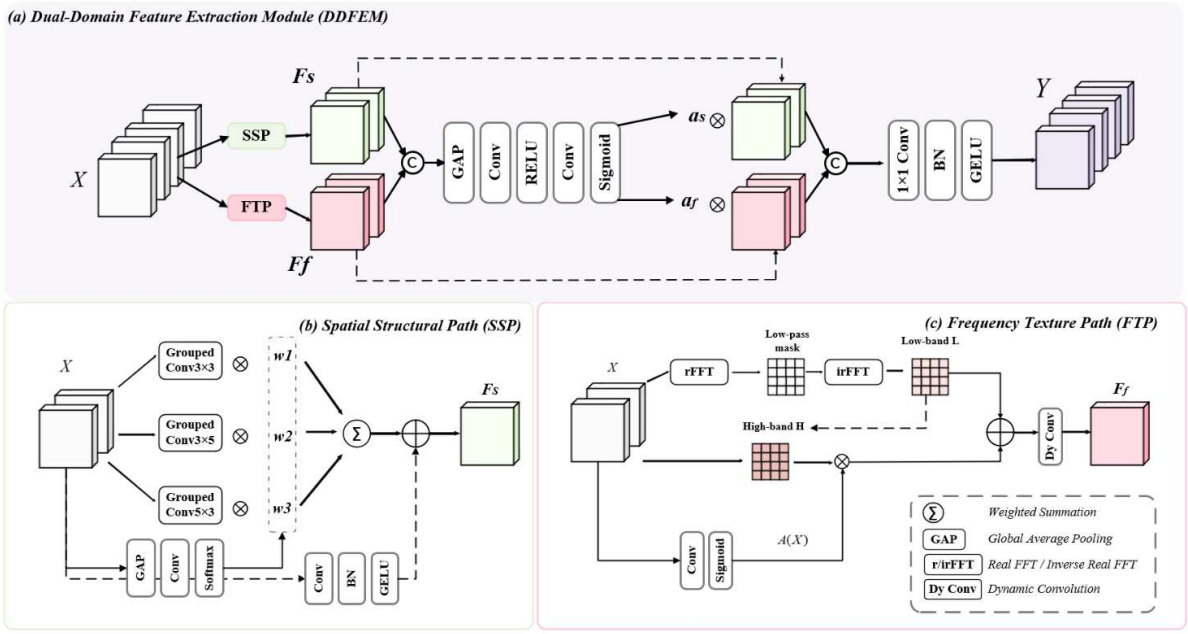

Fig. 2: Dual-Domain Feature Extraction Module (DDFEM). (a) DDFEM separates X into the spatial structural path (SSP) and frequency texture path (FTP), then uses independent gates a_s and a_f to combine F_s and F_f. (b) SSP uses three weighted grouped-convolution branches and a modulation branch. (c) FTP obtains a low-frequency component by real Fourier transformation and low-pass masking, modulates the complementary high-frequency component with $A(X)$, and applies dynamic convolution to produce F_f.

Here, R_l aligns the encoder residual and is the identity when dimensions match. Each λ_l is an unconstrained learned scalar initialized to 0.5. The boundary branch has no independent boundary supervision.

3.4 Dynamic Spatial-Adaptive Upsampler (DSAU)

DSAU uses DySample with pixel-shuffle-first offset prediction, four sampling groups, and a scope branch. After channel projection when needed, pixel-shuffled features predict offsets and Sigmoid scope weights. Their product is pixel-unshuffled, scaled by 0.5, and added to initialized subpixel offsets. The resulting normalized grid guides grouped bilinear sampling of the projected input. Border padding and `align_corners=False` are used. Intermediate decoder stages upsample by two, and the final stage upsamples by four.

4 Experiments

4.1 Datasets

Synapse contains 30 abdominal CT volumes, split into 18 training and 12 evaluation cases. We report DSC and Hausdorff-based distance for eight abdominal organs. ACDC contains cardiac cine MRI from 100 subjects [21], split into 70 training, 10 validation, and 20 evaluation cases; mean DSC is reported for RV, MYO, and LV. Following the HiFormer protocol, ISIC2017 uses 1,400/200/400 training/validation/test images, while ISIC2018 uses 1,815/259/520 [22, 23]. Skin-lesion evaluation reports DSC, sensitivity, specificity, and accuracy.

4.2 Experimental settings

DDSF-Net was implemented in Python using the PyTorch framework and trained on an NVIDIA® GeForce RTX™ 4090 GPU with 24GB of VRAM. The CSWin backbone was initialized with ImageNet pre-trained weights to accelerate convergence and stabilize optimization. In the current

implementation, the network is instantiated with an input image size of 224×224. During training, standard data augmentation operations, including flipping and rotation, were used to improve robustness to spatial variation. The batch size was set to 24, and optimization was performed with SGD using a learning rate of 0.05, a momentum coefficient of 0.9, and a weight decay of 10^{-4}. Unless otherwise stated, all reported DDSF-Net results were obtained under this training setting.

To jointly optimize region overlap and pixel-wise classification accuracy, we used a hybrid loss function that combines Dice loss and cross-entropy loss:

$$\text{Loss} = \alpha \cdot \text{Loss}_{\text{Dice}} + \beta \cdot \text{Loss}_{\text{cross}} \tag{7}$$

where α and β are balancing coefficients for the Dice and cross-entropy terms, respectively. This combined objective encourages both accurate region overlap and stable optimization across different segmentation targets.

4.3 Experimental result

4.3.1 Comparative results on Synapse dataset

As shown in Table 1, DDSF-Net achieved the highest DSC among the methods listed for the Synapse multi-organ segmentation dataset, reaching 82.37% and exceeding CSWin-UNet by 1.25 percentage points. Its reported distance value was 17.91, lower than that of CSWin-UNet (18.86) but higher than that of HiFormer-B (14.70). At the organ level, DDSF-Net did not improve every category: its scores for gallbladder and pancreas were below those of CSWin-UNet. These results indicate an improvement in overall Dice over the listed baseline, but not uniform gains in boundary distance or across all organs.

At the organ level, DDSF-Net scored higher than CSWin-UNet on six of the eight categories, with gains on aorta (+2.99), left kidney (+0.68), right kidney (+1.04), liver (+1.28), spleen (+3.64), and stomach (+1.56). Its scores for gallbladder and pancreas were lower than the baseline. Thus, the overall DSC improvement does not extend uniformly to all organs, particularly the two small-organ categories.

The qualitative examples in Fig. 3 are consistent with the quantitative results. Compared with the reference methods, DDSF-Net produced more continuous contours for spleen and stomach and fewer obvious leakage regions around complex abdominal boundaries. However, the visual results for small or low-contrast targets remained less stable, which is consistent with the organ-wise quantitative analysis.

4.3.2 Comparative results on ACDC dataset

As shown in Table 2, DDSF-Net achieved the best average DSC on ACDC (91.58%), slightly exceeding CSWin-UNet (91.46%). At the class level, DDSF-Net obtained the best RV and LV scores and the second-best MYO score. The gain over the baseline was therefore modest but consistent with the view that the revised skip fusion and decoder reconstruction are also effective on cardiac MRI.

The qualitative examples in Fig. 4 show the same trend. DDSF-Net generated contours that were generally closer to the reference annotations for RV and LV, especially in regions where several compared methods showed visible blurring or local under-segmentation. Residual errors were still present in thin myocardium regions, which is consistent with the remaining gap on MYO.

4.3.3 Comparative results on skin lesion segmentation datasets

Table 3 and Fig. 5 present the comparative results of DDSF-Net on the ISIC2017 and ISIC2018 skin lesion datasets.

Quantitatively, DDSF-Net achieved the best DSC, SE, and SP on both ISIC2017 and ISIC2018. Compared with CSWin-UNet, the DSC increased from 91.47% to 92.55% on ISIC2017 and from 91.11% to 93.15% on ISIC2018. DDSF-Net also achieved the best ACC on ISIC2017, whereas HiFormer-B remained strongest on ISIC2018 ACC. These results indicate that the proposed design transferred well to dermoscopic segmentation and was particularly effective on overlap- and sensitivity-related metrics.

4.4 Ablation studies

We conducted ablation experiments on the Synapse dataset to analyze the individual and joint contributions of SSP, FTP, and DSAU. Unless otherwise stated, we report DSC, HD95,

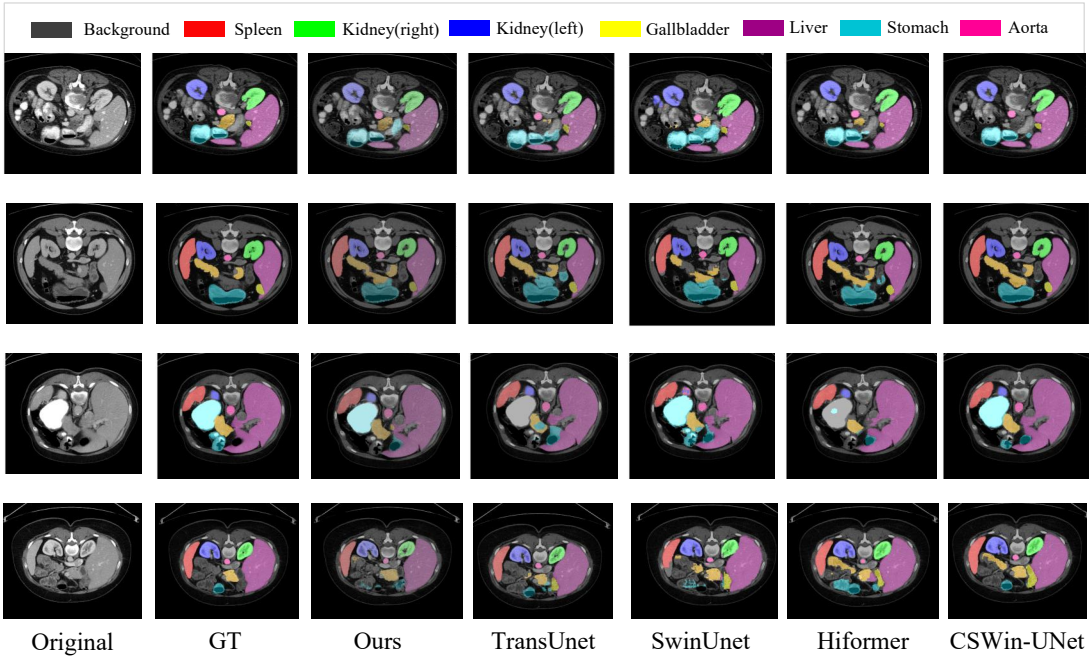

Fig. 3: Segmentation results of 8 abdominal organs (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, stomach) in the Synapse dataset using different methods.

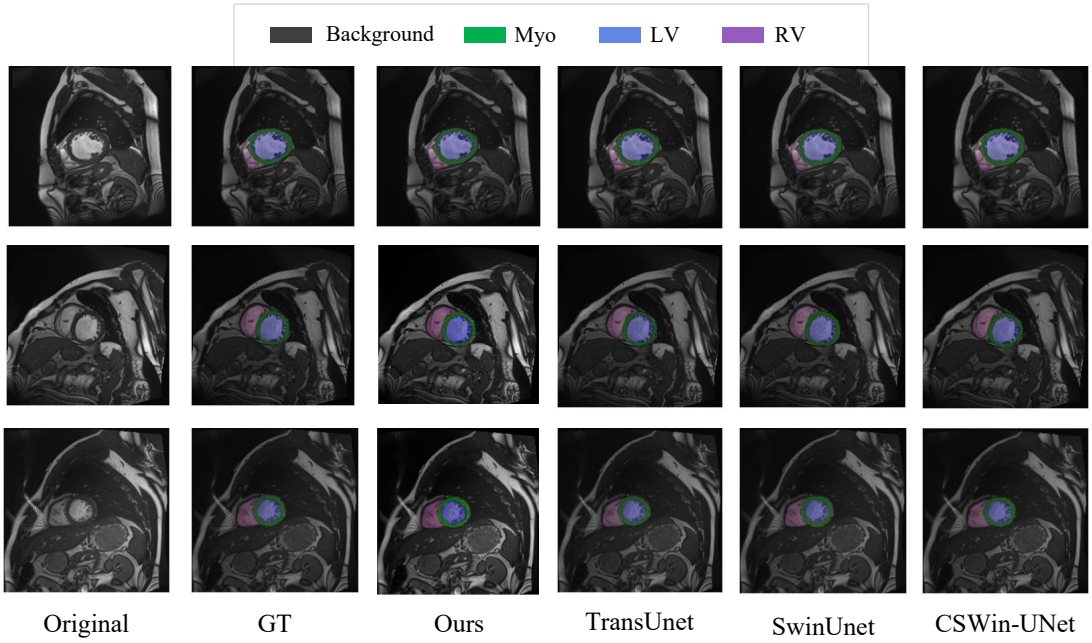

Fig. 4: Qualitative segmentation results of different methods on the ACDC dataset. The color labels indicate background (dark gray), myocardium (Myo, green), left ventricle (LV, blue), and right ventricle (RV, purple).

Table 1: Comparison of segmentation results on the Synapse dataset

Method	DSC	HD	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
UNet	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
HiFormer-B[24]	80.39	**14.70**	86.21	65.69	**85.23**	**79.77**	94.61	59.52	90.99	81.08
MT-UNet[25]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
TransUNet	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-UNet[13]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
CSWin-UNet	81.12	18.86	87.13	67.85	83.51	78.53	95.23	**65.94**	89.05	81.74
DDSF-Net	**82.37**	17.91	**90.12**	66.80	84.19	79.57	**96.51**	65.75	**92.69**	**83.30**

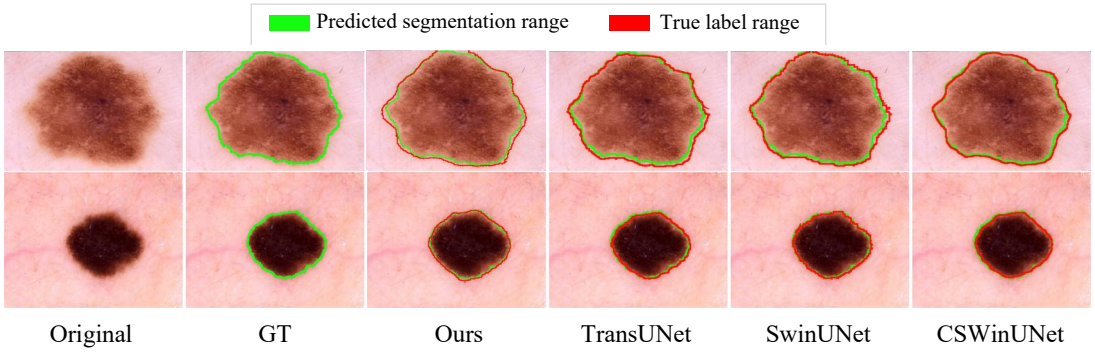

Fig. 5: Qualitative results of different segmentation methods on the ISIC dataset. The green boundaries represent the ground truth, and the red boundaries indicate the predictions.

Table 2: Performance comparison with representative methods on the ACDC dataset. Bold black data indicate the best results, and underlined black data indicate the second-best results.

Method	DSC	RV	MYO	LV
UNet	87.55	87.10	80.63	94.92
Att-UNet	86.75	87.58	79.20	93.47
nnUNet	90.91	89.21	**90.20**	93.35
UNetR	88.61	85.29	86.52	94.02
TransUNet	89.71	88.86	84.53	95.73
Swin-UNet	90.00	88.55	85.62	95.83
CSWin-UNet	91.46	89.68	88.94	95.76
DDSF-Net	**91.58**	**90.00**	89.35	**96.39**

parameter count, and FLOPs so that accuracy changes can be interpreted together with model complexity.

4.4.1 Impact of Different Components in DDSF-Net

Table 4 shows the effect of gradually adding FTP, SSP, and DSAU to the CSWin-UNet baseline. The comparison is intended to clarify which gains come from single modules and which require joint use of the proposed design.

When interpreted together with the baseline result in Table 1, the first two rows show that neither FTP alone nor SSP alone is sufficient to outperform the original CSWin-UNet. However, SSP alone performs better than FTP alone, indicating that explicit morphology-sensitive spatial modeling contributes more strongly than using the frequency-aware branch in isolation. This also suggests that the main value of FTP lies in complementing SSP rather than replacing spatial modeling.

When FTP and SSP are used together, performance increases to 82.03% DSC and 18.44 mm HD95, exceeding both single-branch variants and surpassing the baseline model. This result supports the view that the two branches capture complementary information: SSP emphasizes morphology-sensitive local structure, whereas FTP provides additional frequency-aware texture cues that become more useful when the spatial branch is already present.

Table 3: Comparison results of the proposed DDSF-Net with other representative methods on two skin lesion segmentation datasets. Bold black data indicate the best results, and underlined black data indicate the second-best results.

Method	ISIC2017				ISIC2018			
	DSC	SE	SP	ACC	DSC	SE	SP	ACC
UNet	81.59	81.72	96.80	91.64	85.45	88.00	96.97	94.04
Att-UNet	80.82	79.98	97.76	91.45	85.66	86.74	<u>98.63</u>	93.76
HiFormer-B	<u>92.53</u>	91.55	98.40	<u>97.02</u>	<u>91.02</u>	<u>91.19</u>	97.55	**96.21**
TransUNet	81.23	82.63	95.77	92.07	84.99	85.78	96.53	94.52
CSWin-UNet	91.47	<u>93.79</u>	<u>98.56</u>	97.26	91.11	92.31	97.88	95.25
DDSF-Net	**92.55**	**95.81**	**98.88**	**97.82**	**93.15**	**94.33**	**98.85**	<u>95.56</u>

Table 4: The impact of integrating FTP, SSP, and DSAU modules into the baseline-CSWin-UNet model. The bold data indicates the best results, and the underlined data indicates the second best results.

Modules			Synapse Multi-organ Segmentation	
FTP	SSP	DSAU	DSC (%)	HD95 (mm)
✓	×	×	79.80	22.26
×	✓	×	80.96	20.05
✓	✓	×	<u>82.03</u>	<u>18.44</u>
✓	✓	✓	**82.37**	**17.91**

After adding DSAU on top of FTP+SSP, the complete DDSF-Net achieves 82.37% DSC and 17.91 mm HD95. This indicates that dynamic offset-based upsampling provides a further improvement during decoder reconstruction, especially for recovering fine structural detail. In our setting, DSAU also keeps the computational cost competitive relative to the compared upsampling choices.

4.4.2 Impact of Adaptive Convolution Kernels in the SSP Module

Table 6 presents the ablation study on the impact of multi-scale adaptive convolution kernels in the depthwise convolution of the SSP module on the Synapse multi-organ dataset. Bold black data indicate the best results, and underlined black data indicate the second-best results.

Table 6 evaluates different kernel combinations in SSP. As the branch design expanded from a single 3×3 kernel to the three-branch setting $3 \times 3 + 3 \times 5 + 5 \times 3$, DSC increased from 80.12% to 82.24% and HD95 decreased from 20.08 mm to 18.22 mm within the same $[1, 5]$ range. Extending the kernel range from $[1, 5]$ to $[1, 7]$ further improved performance to 82.37% DSC and 17.26 mm HD95.

4.4.3 Ablation Experiment of FTP Frequency Decomposition

Table 7 summarizes the ablation results for the FTP decomposition configuration. As the configuration expanded from $[2]$ to $[2, 4]$ and then to $[2, 4, 8]$, DSC increased from 80.05% to 82.37% and HD95 decreased from 20.40 mm to 17.26 mm. This trend indicates that combining multiple decomposition granularities was more effective than using only a coarse setting.

When the configuration was shifted toward higher-frequency settings, namely $[4, 8, 16]$, performance decreased to 82.14% DSC and 18.54 mm HD95. Under similar complexity, the balanced $[2, 4, 8]$ setting therefore provided a better trade-off than a configuration biased toward finer components. The additional cost from richer FTP settings was moderate, so $[2, 4, 8]$ was used as the default configuration in the final model.

Figure 6 shows changes in the spatial distribution of feature responses after FTP at Stages 1 and 2. These examples complement the configuration ablation in Table 7, but do not establish a specific spectral mechanism or quantify changes in activation magnitude.

4.4.4 The Impact of Different Upsampling Methods

To evaluate DSAU under matched conditions, we fixed the backbone and dual-domain skip modules

Table 5: Ablation study on different upsampling strategies on the Synapse dataset. Bold black data indicate the best results, and underlined black data indicate the second-best results.

Strategy	DSC (%)	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach	Params (M)	FLOPs (G)
Bilinear interpolation	80.78	89.15	64.23	83.50	77.81	95.10	**65.92**	88.45	83.10	**26.35**	5.38
Transposed convolution	81.56	89.88	65.55	**84.88**	78.90	94.88	62.10	90.10	82.48	29.03	7.88
CARAFE	<u>82.13</u>	<u>90.05</u>	65.10	84.05	<u>79.25</u>	<u>96.20</u>	62.55	<u>91.80</u>	83.04	26.77	<u>5.12</u>
DSAU (ours)	**82.37**	**90.12**	**66.80**	<u>84.19</u>	**79.57**	**96.51**	<u>65.75</u>	**92.69**	**83.30**	<u>26.59</u>	**4.82**

Table 6: Ablation study of multi-scale adaptive convolution kernels in SSP module (Synapse dataset)

Convolution Kernel Configuration	Kernel Size Range	DSC (%)	HD95 (mm)	#Params (M)	#FLOPs (G)
3×3	[1,5]	80.12	20.08	**24.15**	5.02
$3 \times 3 + 3 \times 5$	[1,5]	81.18	19.05	<u>25.76</u>	4.95
$3 \times 3 + 3 \times 5 + 5 \times 3$	[1,5]	<u>82.24</u>	<u>18.22</u>	26.43	<u>4.82</u>
$3 \times 3 + 3 \times 5 + 5 \times 3$	[1,7]	**82.37**	**17.26**	26.59	**4.73**

Table 7: Ablation study on different frequency component configurations.

k_list	Freq Components	DSC (%)	HD95 (mm)	Params (M)	FLOPs (G)
[2]	2	80.05	20.40	**24.34**	**4.18**
[2,4]	3	81.12	19.32	<u>25.10</u>	<u>4.32</u>
[2,4,8]	4	**82.37**	**17.26**	26.78	4.70
[4,8,16]	4	<u>82.14</u>	<u>18.54</u>	26.89	4.73

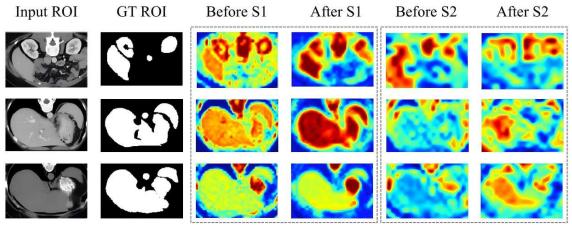

Fig. 6: Feature-response maps before and after FTP at Stages 1 and 2 for three Synapse cases with $k_list = [2, 4, 8]$. Each row shows the input ROI, ground-truth ROI, and paired before/after maps at each stage.

and replaced only the decoder upsampling strategy. The compared methods were bilinear interpolation, transposed convolution, CARAFE[26], and DSAU.

As shown in Table 5, DSAU achieved the best overall DSC (82.37%) while keeping the parameter count and FLOPs competitive. Compared with transposed convolution, DSAU improved overall accuracy with fewer parameters in this setting. Compared with CARAFE, the gain in DSC was smaller, but DSAU still produced the best average result and the lowest FLOPs among the non-bilinear learnable upsampling methods reported here.

5 Discussion

The ablations support complementary spatial and frequency processing, with SSP alone stronger than FTP alone and DSAU providing a further gain. These results indicate practical benefit within the evaluated backbone and settings, but do not establish a unique spectral mechanism or superiority over all recent methods. Gallbladder and pancreas remain weaker than CSWin-UNet, and boundary metrics are not uniformly improved. Broader matched comparisons and stronger feature analysis remain necessary.

6 Conclusions

DDSF-Net integrates spatial-frequency processing within DDFEM, encoder-decoder interaction within BRS-Skip, and dynamic resampling

through DSAU. It improves reported overall DSC over CSWin-UNet across the evaluated datasets and is competitive among the listed methods. These gains are not uniform across organs or metrics.

7 Acknowledgments

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LZ26F010007; in part by National Natural Science Foundation of China under Grant 62371171, 62371172, 62071161; in part by the Central Government-Guided Local Science and Technology Development Fund (Grant No. 2025ZY01045) and the Zhejiang Provincial Key Laboratory of Brain Computer Collaborative Intelligence Technology and Applications (Grant No. 2025E10015)

References

[1] Panayides, A.S., Amini, A., Filipovic, N.D., Sharma, A., Tsaftaris, S.A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K.S., Veasey, B.P., Zervakis, M., Saltz, J.H., Pattichis, C.S.: Ai in medical imaging informatics: Current challenges and future directions. IEEE Journal of Biomedical and Health Informatics **24**(7), 1837–1857 (2020)

[2] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (2015)

[3] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (2018)

[4] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., Mcdonagh, S., Hammerla, N.Y., Kainz, B.: Attention u-net: Learning where to look for the pancreas (2018)

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2023)

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021)

[7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

[8] Ren, P., Li, C., Wang, G., Xiao, Y., Du, Q., Liang, X., Chang, X.: Beyond Fixation: Dynamic Window Visual Transformer (2022)

[9] Liu, X., Gao, P., Yu, T., Wang, F., Yuan, R.-Y.: Cswin-unet: Transformer unet with cross-shaped windows for medical image segmentation. Information Fusion **113**, 102634 (2025)

[10] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (2015)

[11] Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation (2017)

[12] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A., Zhou, Y.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)

[13] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation (2021)

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

[15] Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3559–3568 (2021)

[16] Shengnan, Z., Xinming, C.: Fea-net: End-to-end frequency-edge automation for ambiguous lesion boundary segmentation in gastrointestinal and dermatologic imaging. Preprints (2025) https://doi.org/10.20944/preprints202505.0885.v1

[17] Li, Y., Wang, Y., Leng, T., Zhijie, W.: Wavelet u-net for medical image segmentation. In: Artificial Neural Networks and Machine Learning – ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I (2020)

[18] Saxton, W.O.: Chapter two - the discrete fourier transform reprinted from advances in electronics and electron physics, supplement 10, 1978. In: Computer Techniques for Image Processing in Electron Microscopy vol. 214, pp. 37–48. Elsevier, ??? (2020)

[19] Jena, S.K.: Discrete Fourier Transform, pp. 229–273 (2025)

[20] Li, Y., Wang, Y., Leng, T., Zhijie, W.: Wavelet u-net for medical image segmentation. In: Artificial Neural Networks and Machine Learning – ICANN 2020, Cham, pp. 800–810 (2020)

[21] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.-M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.-M.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018)

[22] Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172 (2018)

[23] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC) (2019)

[24] Heidari, M., Kazerouni, A., Kadarvish, M.S., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 6191–6201 (2022)

[25] Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.-H., Chen, Y.-W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2390–2394 (2022)

[26] Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: Carafe: Content-aware reassembly of features. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 3007–3016 (2019)