

Dual-path adaptive frequency-aware network for SAR image segmentation

Zihan Xia¹, Jin Liu¹, Zhiwen Zheng¹, Xingru Huang¹, Svetlana Illarionova², Jinfeng Liu⁵, Wangmeng Zuo⁴, Changqing Zou³, Evgeny Burnaev²

¹ Hangzhou Dianzi University, Hangzhou 310018, China, ² Applied AI Institute, Moscow, Russia, ³ Zhejiang Lab, Hangzhou 311100, China, ⁴ Harbin Institute of Technology, Harbin 150001, China, ⁵ Zhengzhou University, Zhengzhou 450001, China

Corresponding author: Jin Liu, jinliu@hdu.edu.cn

Synthetic Aperture Radar (SAR) image segmentation aims to recover coherent semantic structures from observations that are often corrupted by speckle noise. This creates a persistent tension for segmentation methods: broad contextual modeling is needed to judge whether local responses belong to a complete target structure, while fine local evidence is still required to recover boundaries, narrow objects, and small regions. We introduce a Dual-Path Adaptive Frequency-Aware Segmentation Network (DPAFNet) designed around this tension. DPAFNet uses a global-local complementary dual-path encoder to decouple long-range structural reasoning from local detail extraction, a hierarchical adaptive representation bridging module to fuse the two streams with reliability-aware branch weights, and a frequency-perceptive skip modulation module to condition high resolution skip features before decoding. Experiments demonstrate that DPAFNet achieves competitive segmentation performance across overlap, agreement, and boundary metrics. Ablation studies further show the proposed encoder design, adaptive bridging strategy, and frequency-aware skip modulation all contribute to the final performance. These results demonstrate the potential of global-local coordination for improving dense prediction in SAR segmentation tasks.

Keywords: Global-local learning, frequency modulation, image segmentation, SAR imagery, speckle noise

1. INTRODUCTION

Synthetic Aperture Radar (SAR) provides an all-weather and day-and-night imaging capability, making it widely used in ocean monitoring, disaster assessment, urban observation, agricultural mapping, and target surveillance [1]. As a fundamental task in SAR image interpretation, semantic segmentation assigns a category label to each pixel and supports fine-grained extraction of target regions [2]. Unlike image-level recognition or object-level detection, SAR image segmentation must recover accurate boundaries, shapes, and spatial distributions under speckle noise and multi-scale land cover structures. Achieving stable and precise segmentation therefore remains a critical problem in SAR-based remote sensing interpretation.

Recent deep learning methods, especially Convolutional Neural Networks (CNNs) and transformers have greatly advanced remote sensing image segmentation. CNN-based architectures are effective at capturing local edges, textures, and gradient variations, which are important for boundary localization and detail recovery [3]. Transformer-based architectures can model long-range spatial dependencies through self-attention and are better suited to capturing global shape, regional continuity, and contextual relations [4]. These models provide strong feature learning capacity for SAR segmentation. However, SAR images differ from optical remote sensing images because coherent imaging introduces multiplicative speckle noise. This noise causes strong intra-region intensity fluctuations and target-like pseudo-texture responses in background areas. As a result, models that rely excessively on local intensity or texture may split continuous targets into fragments or misclassify noisy background responses as foreground evidence.

At the same time, SAR segmentation cannot rely only on large context modeling to obtain stable predictions. Target boundaries, thin structures, small objects, and transitions between adjacent classes often appear as local high frequency changes. Although large receptive fields and global feature aggregation can improve semantic stability, they may also weaken edges, corners, fine contours, and local gradients during feature aggregation, leading to blurred boundaries, target adhesion, contour expansion, or missed small objects. Therefore, SAR image segmentation involves an inherent conflict: global structural discrimination requires broad contextual information to suppress local misclassification caused by speckle noise, while local detail representation requires fine spatial features to preserve boundaries and object contours. A single encoder usually forms a fixed compromise between global semantics and local details, which makes it difficult to achieve both structural completeness and boundary precision.

To address this problem, this paper proposes a Dual-Path Adaptive Frequency-Aware Segmentation Network (DPAFNet). The network introduces a Global-Local Complementary Dual-path Encoder (GLCDE) to decouple global structural modeling from local detail extraction. The global path uses a transformer-based structure to model long-range spatial dependencies and enhance structural discrimination, while the local path uses lightweight convolutional blocks to preserve texture and fine contour information. A Hierarchical Adaptive Representation Bridging Module (HARB) is then designed to dynamically evaluate the relative reliability of global and local features at different levels, avoiding feature conflicts caused by direct concatenation. Finally, a Frequency-Perceptive Skip Modulation Module (FPSM) uses frequency domain structural cues and decoder semantic states to selectively modulate skip features, reducing the unconditional injection of noisy responses and improving the recovery of object contours.

The main contributions are summarized as follows:

1. We propose DPAFNet, a dual-path adaptive frequency-aware network for SAR image segmentation, to improve both structural completeness and boundary precision under speckle noise.
2. We design GLCDE to decouple global structural discrimination and local detail representation, alleviating the fixed compromise of a single encoder between contextual modeling and fine boundary preservation.
3. We propose HARB to adaptively bridge global structural features and local detail features at multiple encoder levels through reliability-driven dynamic fusion.
4. We propose FPSM to selectively modulate skip features with frequency domain structural cues and decoder semantic guidance, reducing the interference of noisy details.

2. RELATED WORK

2.1 CNN-based segmentation methods

Fully convolutional networks transformed classification backbones into pixel-wise predictors [5], while U-Net-style encoder-decoder architectures used skip connections to recover spatial details. Later methods, such as PSPNet [6], further improved context aggregation, multi-scale representation, real-time inference, and lightweight segmentation design. These studies show that convolutional structures are effective for preserving local edges, textures, gradients, and object boundaries. In SAR image segmentation, such local detail modeling is important for targets such as ships and buildings, where boundary errors may affect footprint, shape, or target count [7]. However, speckle noise and acquisition geometry make local intensity and texture evidence unstable [1, 2, 8]. Thus, CNN-based methods can recover fine spatial details, but their local sensitivity may also propagate speckle-induced pseudo-textures and clutter responses, causing fragmented regions, false foreground predictions, and weak structural consistency.

2.2 Transformer-based segmentation methods

Transformer-based segmentation methods improve dense prediction by modeling long-range dependencies. A vision transformer and its dense-prediction variants showed that self-attention can replace purely convolutional spatial aggregation [9]. Hierarchical backbones such as PVT and PVTv2 further adapt transformers to segmentation by producing multi-scale feature maps [10], while SegFormer and related designs demonstrate the effectiveness of hierarchical transformer representations for semantic segmentation [11]. In SAR segmentation, long-range context helps determine whether local responses are consistent with the overall target structure, especially when speckle noise causes intra-target fluctuations or target-like background responses. However, transformer-based aggregation mainly strengthens global semantic consistency and may weaken local gradients, thin contours, and boundary details. Recent graph convolutional and state-space models also explore non-local or long-range modeling [12, 13, 14], but such representations alone still cannot fully address precise boundary recovery in SAR segmentation.

2.3 Hybrid segmentation methods

Hybrid segmentation methods aim to combine local detail modeling with global context representation. Convolution-transformer networks, such as TransFuse, MISSFormer, UNExT, and related gated attention architectures, show

that convolutional inductive bias and global representations can complement each other in segmentation tasks [15, 16]. Multi-scale structures, including feature pyramid networks, high resolution networks, and redesigned skip connections, further improve dense prediction by combining features across resolutions [17]. Attention mechanisms, such as channel attention, spatial-channel attention, attention gates, and dual attention modules, are also widely used to recalibrate feature responses [18, 19]. In addition, frequency and wavelet-domain methods provide structural cues by exposing information that is less explicit in the raw spatial stream [20, 21, 22, 23, 24].

Although these methods improve feature interaction, many hybrid designs rely on fixed fusion operations or general attention recalibration. They do not explicitly distinguish global structural evidence from local detail evidence, which is important for SAR segmentation. Global features help maintain target continuity and suppress noise-induced local errors, whereas local features support boundary recovery, thin structures, and small targets. DPAFNet follows the hybrid direction, but it organizes the network into an explicit global-local-frequency flow that separates global and local representations, bridges them adaptively, and modulates skip feature transfer with frequency-aware decoder guidance.

3. METHOD

Fig. 1 illustrates the overall architecture of the proposed Dual-Path Adaptive Frequency-Aware Segmentation Network (DPAFNet). Given a SAR image $\mathbf{X}$, DPAFNet first uses a Global-Local Complementary Dual-path Encoder (GLCDE) to extract global structural features $\{\mathbf{G}_i\}_{i=1}^4$ and local detail features $\{\mathbf{L}_i\}_{i=1}^4$. The global path captures long-range contextual relations, while the local path preserves boundary-sensitive spatial evidence. At each encoder level, a Hierarchical Adaptive Representation Bridging Module (HARB) adaptively fuses the two feature streams into a bridged representation $\mathbf{B}_i$. The decoder then progressively reconstructs the segmentation map from $\{\mathbf{B}_i\}_{i=1}^4$. During decoding, Frequency-Perceptive Skip Modulation (FPSM) uses frequency-aware structural cues and decoder semantics to selectively transfer high resolution skip features. The final prediction head produces the segmentation result.

3.1 Global-local complementary dual-path encoder

GLCDE is designed to decouple global structural discrimination and local detail representation in SAR image segmentation. Given an input SAR image $\mathbf{X} \in \mathbb{R}^{C_0 \times H \times W}$, the encoder extracts two aligned feature streams:

$$\{\mathbf{G}_i\}_{i=1}^4 = \mathcal{E}_g(\mathbf{X}), \quad \{\mathbf{L}_i\}_{i=1}^4 = \mathcal{E}_l(\mathbf{X}) \quad (1)$$

where $\mathcal{E}_g(\cdot)$ denotes the global structural path, $\mathcal{E}_l(\cdot)$ denotes the local detail path, $\mathbf{G}_i$ and $\mathbf{L}_i$ are the global and local features at the i -th encoder level, and C_0 , H , and W denote the input channel number, height, and width, respectively. The transformer-based global structural path is referred to as the Global Structural Encoder (GSE), and the convolution-based local detail path is referred to as the Local Detail Encoder (LDE) in the ablation experiments.

The global path is implemented with a hierarchical transformer backbone to capture long-range spatial relations. It produces multi-scale structural representations:

$$\mathbf{G}_i = \mathcal{T}_i(\mathbf{G}_{i-1}), \quad \mathbf{G}_i \in \mathbb{R}^{C_i \times H_i \times W_i} \quad (2)$$

where $\mathcal{T}_i(\cdot)$ denotes the i -th transformer stage, and C_i , H_i , and W_i are the channel number, height, and width of the global feature at level i . This path relates local responses to broader target shape, regional continuity, and contextual structure, which helps suppress local misclassification caused by speckle noise.

The local path uses convolutional downsampling blocks to preserve boundary-sensitive spatial evidence. Each block is defined as

$$\mathcal{D}_{a \rightarrow b}(\mathbf{U}) = \eta \left(\text{Conv}_{3 \times 3}^b \left(\eta \left(\text{Conv}_{3 \times 3}^{a \rightarrow b} \left(\text{MaxPool}_2(\mathbf{U}) \right) \right) \right) \right) \quad (3)$$

where $\mathbf{U}$ is the input feature map, $\mathcal{D}_{a \rightarrow b}(\cdot)$ maps the channel number from a to b , $\text{MaxPool}_2(\cdot)$ denotes 2×2 downsampling, $\text{Conv}_{3 \times 3}^{a \rightarrow b}(\cdot)$ denotes a 3×3 convolution from a channels to b channels, and $\eta(\cdot)$ denotes batch normalization followed by ReLU activation. The local features are then obtained as

$$\mathbf{L}_i = \mathcal{D}_i(\mathbf{L}_{i-1}), \quad \mathbf{L}_i \in \mathbb{R}^{C_i \times H_i \times W_i} \quad (4)$$

Here, $\mathcal{D}_i(\cdot)$ denotes the local downsampling block at level i . The local path retains edges, gradients, textures, and fine contours that are essential for accurate boundary localization.

The global path provides structural constraints for stable target recognition, while the local path provides fine spatial evidence for boundary recovery. The paired features $(\mathbf{G}_i, \mathbf{L}_i)$ are then passed to HARB for adaptive representation bridging.

3.2 Hierarchical adaptive representation bridging

Although GLCDE produces paired global and local features, the two streams are not homogeneous. The global

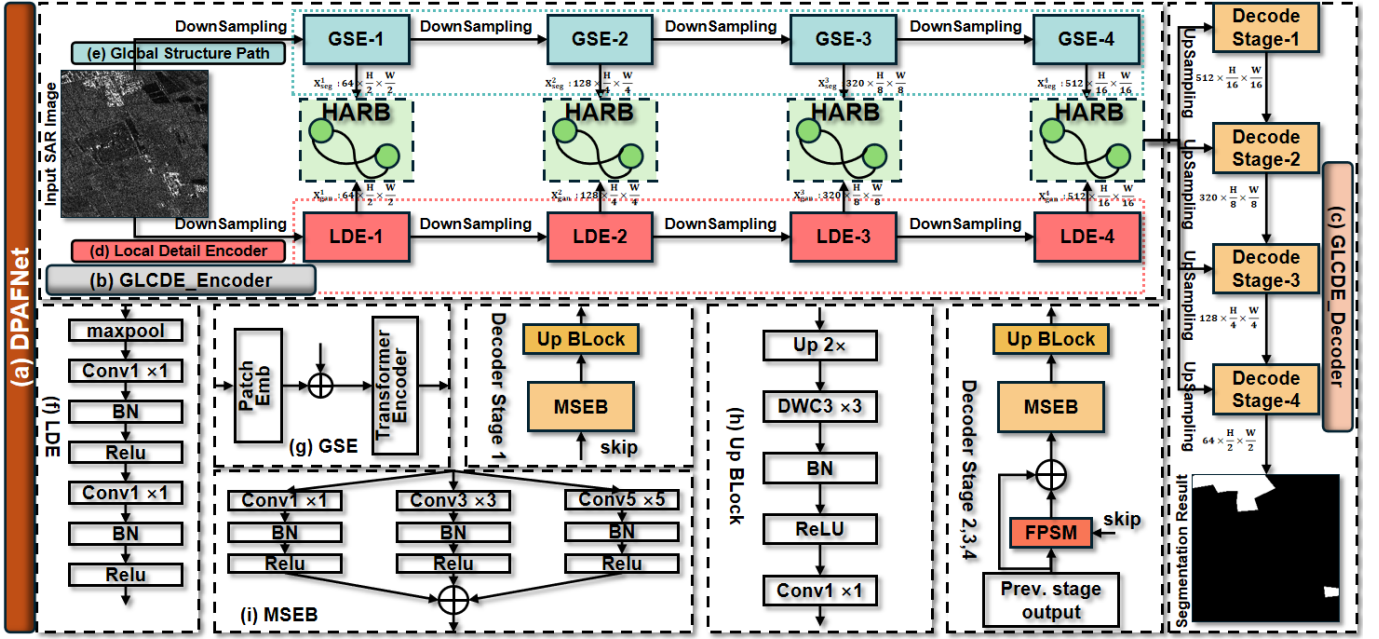


Figure 1 – Overall framework of the Dual-Path Adaptive Frequency-Aware Segmentation Network (DPAFNet).

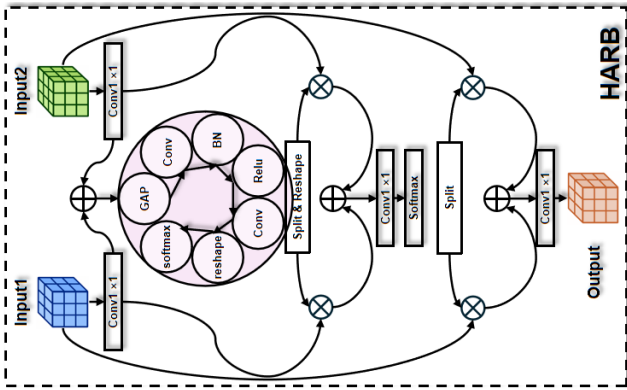


Figure 2 – Structure of the Hierarchical Adaptive Representation Bridging Module (HARB).

feature mainly encodes structural consistency and contextual relations, whereas the local feature focuses on edges, gradients, and fine contours. Direct addition assumes equal reliability of the two streams, and concatenation only stacks them without explicit reliability estimation. To address this problem, HARB adaptively bridges the two representations at each encoder level.

At level i , HARB takes the global feature G_i and the local feature L_i as inputs. First, the two features are projected into a compact comparison space:

$$Q_i^g = \mathcal{P}_i^g(G_i), \quad Q_i^l = \mathcal{P}_i^l(L_i) \quad (5)$$

where $\mathcal{P}_i^g(\cdot)$ and $\mathcal{P}_i^l(\cdot)$ are 1×1 projection layers for the global and local streams. $Q_i^g, Q_i^l \in \mathbb{R}^{C_c \times H_i \times W_i}$ are the compressed comparison features, and C_c is the channel number of the comparison space. In our implementation, C_c is set to 8.

Then, HARB estimates branch-level reliability in the comparison space:

$$s_i = \text{GAP}(Q_i^g + Q_i^l) \quad (6)$$

$$[\beta_i^g, \beta_i^l] = \text{Softmax}_{m \in \{g, l\}}(\mathcal{M}_i(s_i)) \quad (7)$$

where $\text{GAP}(\cdot)$ denotes global average pooling, $\mathcal{M}_i(\cdot)$ denotes a lightweight nonlinear mapping, and β_i^g and β_i^l are branch-level weights for the global and local comparison features.

The weighted comparison representation is computed as

$$R_i = \beta_i^g \odot Q_i^g + \beta_i^l \odot Q_i^l \quad (8)$$

where $\odot$ denotes element-wise multiplication with broadcasting. Based on R_i , HARB further generates spatial reliability maps:

$$[\alpha_i^g, \alpha_i^l] = \text{Softmax}_{m \in \{g, l\}}(\mathcal{W}_i(R_i)) \quad (9)$$

$$\alpha_{i,u,v}^g + \alpha_{i,u,v}^l = 1 \quad (10)$$

where $\mathcal{W}_i(\cdot)$ is the spatial weight generator, and $\alpha_i^g, \alpha_i^l \in \mathbb{R}^{1 \times H_i \times W_i}$ denote the spatial reliability maps for the global and local features. At each spatial position (u, v) , the two weights are normalized to sum to one.

Finally, the learned reliability maps are applied to the original global and local feature carriers:

$$B_i = \phi_i(\alpha_i^g \odot G_i + \alpha_i^l \odot L_i) \quad (11)$$

where $B_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ is the bridged representation, and $\phi_i(\cdot)$ denotes a 3×3 convolution for feature refinement.

In this way, HARB uses the compressed comparison space only to infer the relative reliability of the two streams, while keeping the original global and local features as the information carriers. This allows the model to strengthen global structural evidence in noisy homogeneous regions and preserve local detail evidence near boundaries, thin structures, and small targets.

3.3 Frequency-perceptive skip modulation

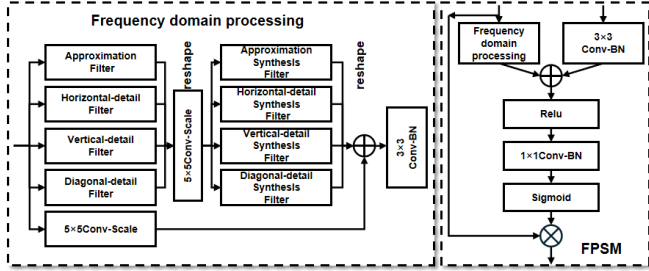


Figure 3 – Structure of the Frequency-Perceptive Skip Modulation Module (FPSM).

As shown in Fig. 3, FPSM controls the transfer of bridged skip features during decoding. At decoder level i , the module takes the upsampled decoder feature $\mathbf{U}_i$ and the bridged skip feature $\mathbf{B}_i$ as inputs. Here, $\mathbf{U}_i$ represents the current semantic reconstruction state, while $\mathbf{B}_i$ provides high resolution spatial information from the encoder.

FPSM first extracts a frequency-aware structural cue from the bridged skip feature:

$$\mathbf{C}_i = \mathcal{P}_{\text{freq}}(\mathbf{B}_i), \quad \mathbf{C}_i \in \mathbb{R}^{C_i \times H_i \times W_i} \quad (12)$$

where $\mathcal{P}_{\text{freq}}(\cdot)$ denotes the frequency-aware processing operator, $\mathbf{C}_i$ is the structural cue used for skip modulation.

The frequency-aware operator consists of a wavelet domain branch and a spatial base branch:

$$\mathcal{P}_{\text{freq}}(\mathbf{B}_i) = \mathcal{W}_{\text{db1}}^{-1} \left(\text{Unstack} \left(\Gamma_i^w \odot \text{DWConv}_{5 \times 5}^{4C_i} \left(\text{Stack} \left(\mathcal{W}_{\text{db1}}(\mathbf{B}_i) \right) \right) \right) + \Gamma_i^b \odot \text{DWConv}_{5 \times 5}^{C_i}(\mathbf{B}_i) \right) \quad (13)$$

Here, $\mathcal{W}_{\text{db1}}(\cdot)$ and $\mathcal{W}_{\text{db1}}^{-1}(\cdot)$ denote the fixed one-level db1 wavelet transform and inverse transform. $\text{Stack}(\cdot)$ concatenates the four wavelet subbands along the channel dimension, and $\text{Unstack}(\cdot)$ restores them to the subband form for inverse transformation. $\text{DWConv}_{5 \times 5}^{4C_i}(\cdot)$ and $\text{DWConv}_{5 \times 5}^{C_i}(\cdot)$ denote depthwise convolutions in the wavelet and spatial branches, respectively. Γ_i^w and Γ_i^b are learnable scale parameters. The wavelet domain branch captures localized frequency structures, while the spatial base branch preserves the original spatial response.

The db1 (Haar) wavelet is selected because its compact support provides localized responses to abrupt intensity

changes and object boundaries with low computational cost. Since the wavelet branch in FPSM is used to provide structural cues at multiple decoder levels, db1 offers a suitable balance between boundary sensitivity, computational efficiency, and implementation simplicity.

Then, the decoder feature and the frequency-aware cue are projected into a shared modulation space to estimate a spatial mask:

$$\mathbf{M}_i = \sigma \left(\text{BN}_i^\psi \left(\omega_i^\psi * \rho \left(\text{BN}_i^d \left(\omega_i^d * \mathbf{U}_i \right) + \text{BN}_i^s \left(\omega_i^s * \mathbf{C}_i \right) \right) \right) \right), \quad \mathbf{M}_i \in [0, 1]^{1 \times H_i \times W_i} \quad (14)$$

Here, $*$ denotes convolution, ω_i^d and ω_i^s are learnable convolution kernels for the decoder feature and the frequency-aware cue, ω_i^ψ is the kernel for mask generation, $\text{BN}(\cdot)$ denotes batch normalization, $\rho(\cdot)$ denotes ReLU, and $\sigma(\cdot)$ denotes sigmoid.

FPSM does not rely on a manually fixed coefficient to balance the frequency cue and the spatial response. Their relative contributions are controlled by learnable scale and projection parameters that are jointly optimized with the entire network. Therefore, the fusion weights are adaptively determined during end-to-end training rather than selected as fixed hyperparameters.

The spatial mask is applied to the original bridged skip feature:

$$\tilde{\mathbf{B}}_i = \mathbf{B}_i \odot \mathbf{M}_i, \quad \tilde{\mathbf{B}}_i \in \mathbb{R}^{C_i \times H_i \times W_i} \quad (15)$$

where $\odot$ denotes element-wise multiplication with broadcasting, and $\tilde{\mathbf{B}}_i$ is the FPSM-modulated skip feature. In this way, the frequency-aware response describes the structural content of the skip feature, and the decoder state provides the current reconstruction demand. The original bridged skip feature remains the carrier of high resolution spatial information, while the learned mask controls its selective transfer.

After skip modulation, the decoder progressively reconstructs the segmentation representation from the deepest bridged feature:

$$\mathbf{D}_4 = \text{MSCB}_4(\mathbf{B}_4) \quad (16)$$

$$\mathbf{U}_i = \text{UpBlock}_i(\mathbf{D}_{i+1}), \quad \mathbf{D}_i = \text{MSCB}_i(\mathbf{U}_i + \tilde{\mathbf{B}}_i) \quad (17)$$

Here, $\mathbf{D}_i$ denotes the decoder feature at level i , $\text{UpBlock}_i(\cdot)$ denotes the upsampling block, and $\text{MSCB}_i(\cdot)$ denotes the multi-scale convolutional refinement block. Finally, a 1×1 prediction head followed by upsampling produces the segmentation result:

$$\hat{\mathbf{Y}} = \text{Up}_4 \left(\omega_{1 \times 1}^o * \mathbf{D}_1 + \mathbf{b}^o \right) \quad (18)$$

where $\omega_{1 \times 1}^o$ and $\mathbf{b}^o$ are the weight and bias of the output classifier, and $\text{Up}_4(\cdot)$ restores the prediction to the target resolution.

Table 1 – Quantitative comparison of different methods on the SARBuD and HRSID datasets. The best results are highlighted in bold.

Method	Type	HD95↓		Dice(%)↑		mIoU(%)↑		Kappa(%)↑		MCC(%)↑	
		SARBuD	HRSID	SARBuD	HRSID	SARBuD	HRSID	SARBuD	HRSID	SARBuD	HRSID
U-Net	CNN	5.96(±2.00)	2.85(±1.54)	74.37(±19.74)	89.97(±8.95)	62.33(±20.33)	82.73(±12.05)	70.05(±19.73)	89.88(±9.15)	71.36(±18.66)	90.12(±8.68)
PSPNet		6.15(±2.07)	2.84(±1.50)	71.08(±19.99)	90.16(±9.80)	58.29(±20.66)	83.21(±12.76)	65.99(±20.07)	90.12(±9.81)	67.54(±18.84)	90.34(±9.29)
LEDNet		6.17(±2.05)	3.10(±1.65)	70.19(±20.57)	86.55(±12.22)	57.34(±20.99)	77.87(±14.80)	64.76(±21.08)	86.47(±12.34)	66.52(±19.73)	86.84(±11.56)
ICNet		6.29(±2.11)	3.14(±1.60)	66.17(±22.67)	85.45(±11.58)	53.19(±22.50)	76.04(±14.35)	60.33(±22.37)	85.40(±11.60)	62.55(±20.94)	85.77(±11.05)
TransUNet	CNN+Transformer	6.11(±2.01)	2.84(±1.59)	69.52(±20.66)	89.89(±10.95)	56.57(±21.18)	83.05(±14.31)	64.32(±20.57)	89.84(±10.97)	66.10(±19.20)	90.08(±10.50)
UNext		6.17(±2.00)	2.96(±1.61)	70.45(±21.68)	88.00(±10.96)	58.00(±22.00)	79.90(±13.81)	65.31(±21.59)	87.94(±10.99)	67.06(±20.26)	88.22(±10.50)
U-Net+Att	CNN+Att	5.87(±1.99)	2.81(±1.61)	75.30(±19.09)	90.82(±9.63)	63.42(±20.26)	84.27(±12.52)	70.60(±19.32)	90.78(±9.63)	71.97(±18.06)	90.97(±9.26)
PAttUNet		6.01(±2.08)	2.82(±1.54)	74.10(±18.72)	90.46(±9.55)	61.78(±20.01)	83.66(±12.51)	69.02(±19.40)	90.42(±9.55)	70.64(±17.87)	90.60(±9.23)
DAttUNet		6.13(±2.12)	2.82(±1.55)	72.88(±19.78)	90.48(±9.43)	60.52(±20.87)	83.66(±12.35)	67.88(±20.29)	90.44(±9.42)	69.62(±18.77)	90.60(±9.18)
DANet		6.06(±2.06)	2.89(±1.57)	73.28(±18.63)	89.88(±10.15)	60.72(±19.99)	82.80(±13.00)	68.00(±19.13)	89.84(±10.16)	69.69(±17.60)	90.08(±9.61)
SGCNet	CNN+GCN	6.04(±2.01)	2.75(±1.55)	72.63(±19.68)	91.42(±8.30)	60.13(±20.49)	85.06(±11.50)	67.67(±19.91)	91.33(±8.53)	69.60(±18.25)	91.54(±7.98)
BAT	Transformer	6.18(±2.04)	3.34(±1.37)	72.93(±18.66)	85.14(±9.46)	60.22(±19.68)	75.13(±12.25)	67.69(±19.02)	85.06(±9.56)	68.98(±17.88)	85.40(±9.20)
MISSFormer		6.11(±2.07)	2.92(±1.63)	70.87(±20.33)	89.17(±9.79)	58.18(±21.28)	81.59(±13.00)	65.54(±20.67)	89.08(±10.01)	67.61(±19.00)	89.35(±9.46)
MambaHSI	Mamba	6.04(±1.90)	3.12(±1.66)	64.93(±20.63)	84.97(±15.10)	51.18(±20.66)	76.18(±17.83)	58.88(±20.01)	84.88(±15.23)	60.44(±19.03)	85.22(±14.66)
VMUNet		6.06(±2.06)	2.81(±1.48)	72.47(±20.00)	90.36(±9.07)	60.06(±21.02)	83.39(±11.99)	67.45(±20.23)	90.28(±9.25)	69.12(±18.76)	90.52(±8.59)
Ours	–	5.84(±2.01)	2.66(±1.52)	77.55(±16.37)	92.15(±8.33)	66.21(±18.60)	86.34(±11.96)	73.29(±17.02)	92.09(±8.62)	75.02(±15.63)	92.28(±8.41)

4. EXPERIMENTS

4.1 Experimental settings

Datasets. The experiments are conducted on two SAR image segmentation datasets: SARBuD and HRSID. SAR-BuD 1.0 is used for building extraction from GF-3 fine-mode SAR imagery. It covers diverse terrain conditions, including plains, mountains, and plateau regions, and the image samples are cropped to 256×256 for training and testing. HRSID is adopted for high resolution SAR ship segmentation in sea, nearshore, harbor, and other complex maritime scenes [7]. It contains 136 panoramic SAR images, 5 604 images of size 800×800 , and 16 951 ship instances, with image resolutions ranging from approximately 1 m to 5 m.

Implementation details. All methods are evaluated under the same preprocessing and training configurations to ensure reproducibility and fair comparison. The input images are processed with consistent normalization and resizing. Random cropping, random intensity perturbation, and affine transformations, including rotation, translation, and scaling, are applied for data augmentation. The initial learning rate is set to 0.0001, and AdamW is used as the optimizer. All networks are trained for 100 epochs with a batch size of 2. All experiments are conducted on a workstation with two NVIDIA RTX 3090 GPUs, an AMD Ryzen 9 5950X CPU, and 128 GB RAM. The software environment consists of Python 3.10.0, PyTorch 2.0.0, and CUDA 11.8.

Evaluation metrics. The segmentation performance is evaluated using HD95, Dice, mIoU, Kappa, and MCC. HD95 measures boundary distance error, where a lower value indicates better performance. Dice and mIoU measure region-level overlap between predictions and

Table 2 – Ablation experiments of the SARBuD dataset on key modules of FPSM, LDE, GSE and HARB

FPSM	LDE	GSE	HARB	HD95↓	Dice(%)↑	mIoU(%)↑	Kappa(%)↑	MCC(%)↑
✓	✓	✓	✓	5.84(±2.01)	77.55(±16.37)	66.21(±18.60)	73.29(±17.02)	75.02(±15.63)
✓	✓	✓	✓	5.86(±2.01)	76.26(±17.75)	64.40(±19.33)	71.83(±18.06)	73.12(±16.75)
✓	✓	✓	✓	5.88(±2.03)	77.02(±17.84)	65.35(±19.42)	72.66(±18.17)	74.42(±16.11)
✓	✓	✓	✓	5.85(±2.02)	76.47(±17.41)	64.57(±19.26)	71.87(±17.94)	73.36(±16.10)
✓	✓	✓	✓	6.01(±2.09)	75.58(±18.32)	63.59(±19.74)	71.05(±18.71)	72.75(±16.93)

ground truth. Kappa and MCC measure prediction agreement and are useful for evaluating segmentation under class imbalance. For Dice, mIoU, Kappa, and MCC, higher values indicate better performance.

4.2 Comparison with existing methods

Table 1 compares DPAFNet with representative SAR image segmentation methods on the SARBuD and HRSID datasets. The compared methods cover CNN-based methods, such as U-Net, PSPNet, LEDNet, and ICNet [3, 6, 25, 26]; CNN+Transformer-based methods, such as TransUNet and UNext [27, 28]; transformer-based methods, such as BAT and MISSFormer [4, 29]; CNN+attention-based methods, such as U-Net+Att, PAttUNet, DAttUNet, and DANet [30, 31]; a CNN+GCN-based method, namely SGCNet [32]; and Mamba-based methods, including MambaHSI and VMUNet [33, 34].

On the SARBuD dataset, DPAFNet achieves the best performance across all five metrics, with 5.84 HD95, 77.55% Dice, 66.21% mIoU, 73.29% Kappa, and 75.02% MCC. Compared with the strongest competing method, U-Net+Att, DPAFNet improves Dice from 75.30% to 77.55% and mIoU from 63.42% to 66.21%, while reducing HD95 from 5.87 to 5.84. These gains are consistent with the motivation of GLCDE and HARB: SAR building extraction requires both global structural consistency to suppress speckle-induced false responses and local detail preser-

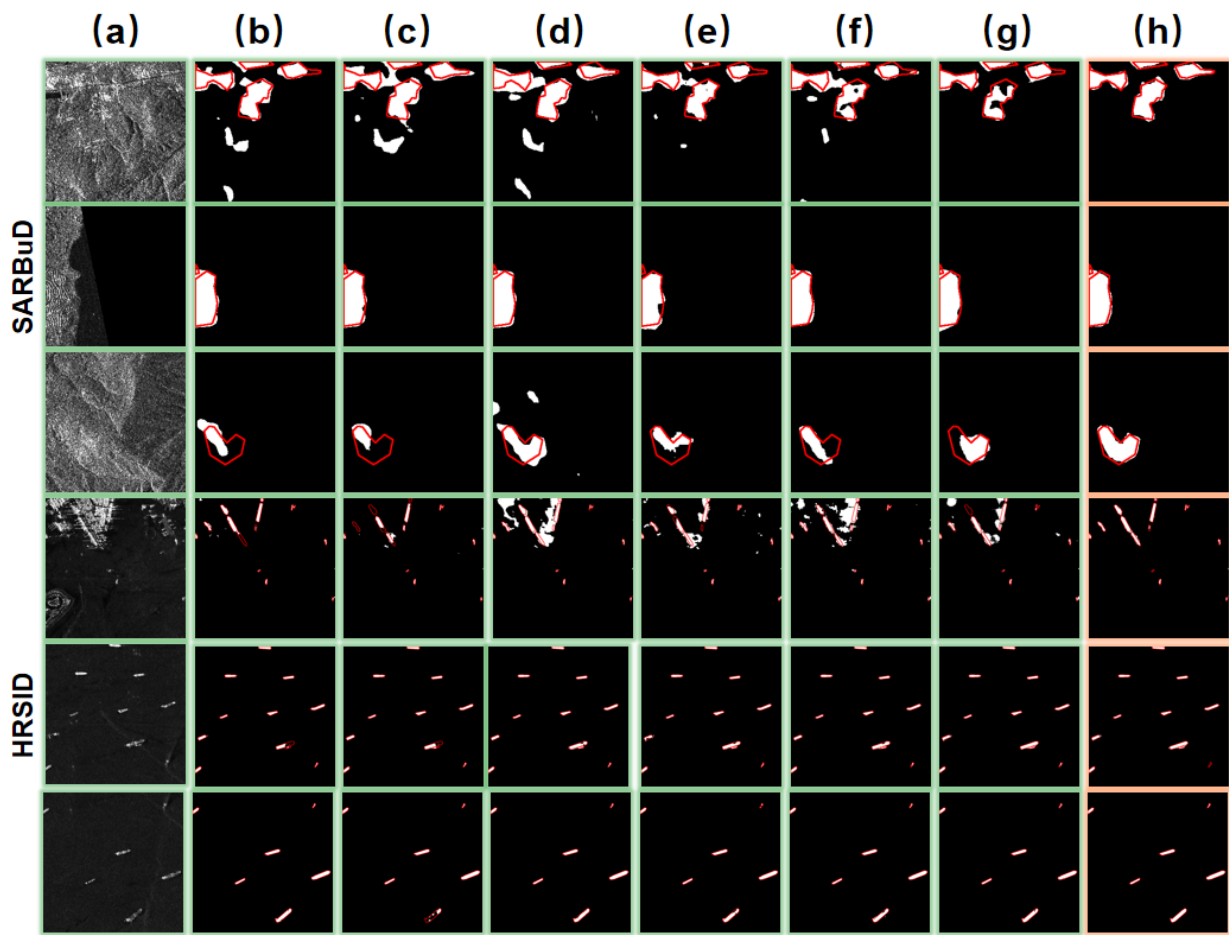


Figure 4 – Visual comparison of multi-model segmentation results on the SARBuD and HRSID datasets.

vation to recover unstable building boundaries. The improvement in overlap metrics and boundary distance indicates that the proposed global-local representation and adaptive bridging strategy are beneficial for complex building segmentation scenes.

On the HRSID dataset, DPAFNet also obtains the best overall performance, achieving 2.66 HD95, 92.15% Dice, 86.34% mIoU, 92.09% Kappa, and 92.28% MCC. Compared with SGCNet, the closest competing method, DPAFNet improves Dice from 91.42% to 92.15% and mIoU from 85.06% to 86.34%. Since ships in HRSID are often small, elongated, and affected by maritime clutter, the consistent improvement suggests that DPAFNet can better preserve target contours and suppress noisy local responses. This is in line with the role of FPSM, which uses frequency-aware structural cues and decoder semantics to selectively transfer high resolution skip information.

Overall, the results on SARBuD and HRSID show that DPAFNet is effective for both SAR building extraction and SAR ship segmentation. The improvements across overlap, agreement, and boundary metrics support the joint contribution of global structural modeling, local detail preservation, reliability-aware feature bridging, and frequency-perceptive skip modulation.

Table 3 – Comparison of model complexity and inference latency across different segmentation methods.

Model	Params (M)↓	FLOPs (G)↓	Latency (ms)↓
U-Net	31.04	437.94	28.70
PSPNet	46.70	369.47	22.80
LEDNet	2.31	12.64	5.60
ICNet	26.24	73.93	7.70
TransUNet	66.82	260.82	25.30
UNet++	1.47	4.59	2.60
U-Net+Att	7.98	116.19	11.40
PAttUNet	11.17	134.72	13.30
DAttUNet	11.25	197.84	20.30
DANet	47.56	410.70	25.90
SGCNet	30.07	66.11	11.60
BAT	21.43	50.52	7.70
MISSFormer	35.45	14.51	11.80
MambaHSI	42.73	623.19	75.20
VMUNet	0.05	15.34	192.30
Ours	37.28	16.13	18.21

Table 3 compares the model complexity and inference latency of DPAFNet with the representative segmentation methods. The comparison includes trainable parameters, FLOPs, and inference latency. DPAFNet contains 37.2894M parameters, requires 16.1310G FLOPs, and achieves an inference latency of 18.21 ms per image.

Although DPAFNet adopts a dual-path encoder to model global structure and local details, its computational cost remains moderate compared with several high-complexity methods. These results indicate that DPAFNet achieves a reasonable balance between segmentation per-

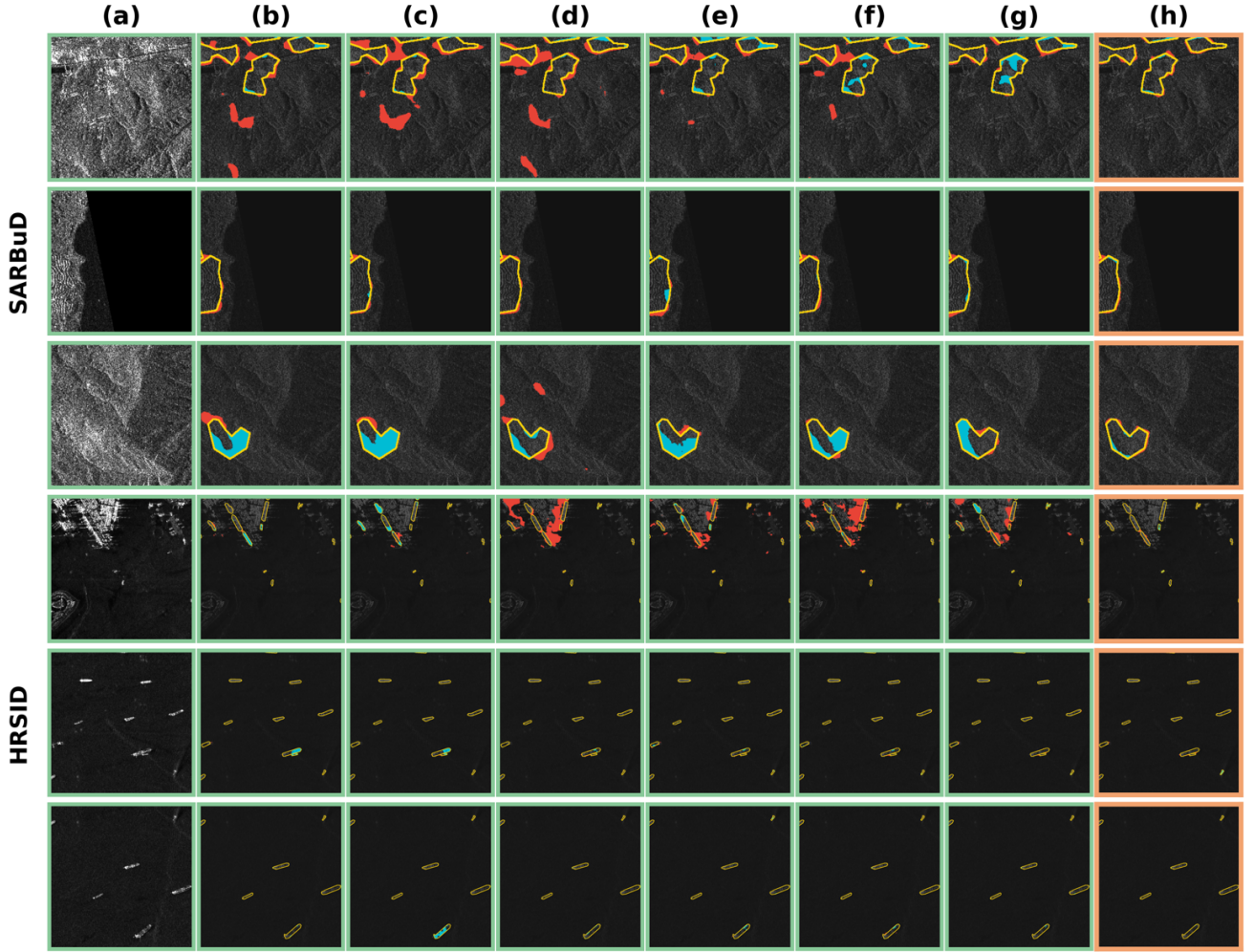


Figure 5 – Spatial error comparison of multi-model segmentation results on the SARBuD and HRSID datasets.

formance and computational efficiency, improving SAR segmentation accuracy without introducing excessive inference overhead.

To provide a boundary-oriented qualitative comparison, we further construct spatial error maps between the predictions and ground truth for representative samples from SARBuD and HRSID. False-positive regions are marked in red, false-negative regions are marked in cyan, and ground truth boundaries are marked in yellow. Uncolored regions indicate correct predictions over the darkened SAR images. This representation distinguishes over-segmentation and missed target regions more clearly than the final binary masks alone.

As shown in Fig. 5, DPAFNet generally produces fewer and more localized error regions than the representative comparison methods. In particular, the error maps show reduced false responses and missed regions around irregular building contours and small or elongated ship targets. These visual results complement the overlap and boundary metrics by directly presenting the spatial distribution of segmentation errors in complex boundary and high frequency detail regions.

4.3 Cross-dataset zero-shot evaluation

To evaluate the cross-dataset generalizability of the learned features, we directly applied the DPAFNet model trained on SARBuD to the HRSID test set. The SARBuD-trained model and its parameters remained unchanged, and no HRSID samples were used for retraining, fine-tuning, or model selection. The HRSID ground-truth masks were used only for calculating the evaluation metrics. This setting represents a stringent zero-shot transfer from SAR building segmentation to SAR ship segmentation.

Under the SARBuD-to-HRSID zero-shot setting, DPAFNet achieves an HD95 of 54.73 ± 101.54 , a Dice score of $60.11\% \pm 32.49\%$, an mIoU of $49.90\% \pm 30.22\%$, a Kappa score of $59.43\% \pm 32.96\%$, and an MCC of $68.13\% \pm 28.90\%$. These results show that the learned representation retains partial transferability under substantial dataset and target-category shifts. The performance gap relative to the in-domain HRSID evaluation also quantifies the remaining sensitivity to dataset-specific characteristics and provides direct evidence for assessing the cross-dataset overfitting risk.

4.4 Robustness to speckle noise

To evaluate the robustness of DPAFNet to varying speckle-noise intensities, simulated multiplicative Gamma speckle noise with different numbers of looks was added to the SARBuD test set. The trained model was tested directly under all noise conditions without retraining or fine-tuning.

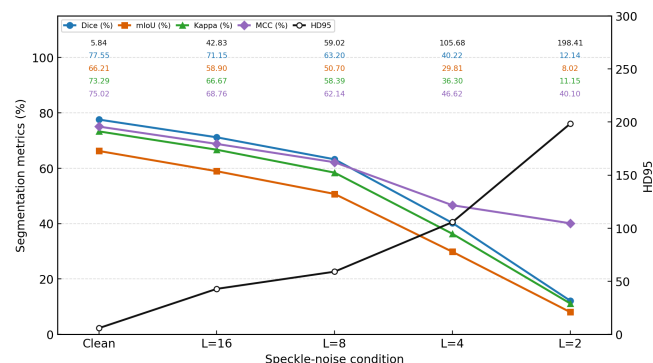


Figure 6 – Performance degradation of DPAFNet under different intensities of simulated multiplicative Gamma speckle noise on the SARBuD test set. A smaller number of looks L indicates stronger speckle noise.

As shown in Fig. 6, the segmentation performance decreases progressively as the speckle noise becomes stronger. DPAFNet maintains relatively stable performance under mild and moderate noise, while a more evident degradation occurs under severe noise. Together with the component ablation results in Table 2, these results provide complementary evidence for the noise-robust design of HARB and FPSM.

4.5 Ablation study

Table 2 verifies the contribution of the main components, including FPSM, LDE, GSE, and HARB. The complete DPAFNet achieves the best overall result, with 77.55% Dice and 66.21% mIoU on SARBuD. Removing any component leads to performance degradation, which shows that all four modules contribute to the final segmentation performance. Among them, removing HARB causes the most evident decline, indicating that adaptive bridging between global structural features and local detail features is the key to effective feature integration.

5. CONCLUSION

This paper presents DPAFNet, a dual-path adaptive frequency-aware network for SAR image segmentation under speckle noise. The method addresses the conflict between global structural discrimination and local detail preservation by separating long-range context modeling and boundary-sensitive feature extraction into two complementary paths. It further uses reliability-aware representation bridging and frequency-perceptive skip modulation to adaptively fuse heterogeneous features and control high resolution information transfer during decoding. Experiments on SARBuD and HRSID show competitive performance, and ablation studies verify the effectiveness of the main components. Overall, the results indicate that reliability-aware coordination of structural context and local detail is effective for robust and precise SAR image segmentation. Future work will explore more efficient frequency-aware designs and extend the framework to broader SAR imaging modes, target categories, and cross-scene segmentation settings.

REFERENCES

- [1] Chris Oliver and Shaun Quegan. *Understanding synthetic aperture radar images*. SciTech Publishing, 2004.
- [2] Jong-Sen Lee and Eric Pottier. *Polarimetric radar imaging: from basics to applications*. CRC press, 2017.
- [3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.
- [4] Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin. “Boundary-aware transformers for skin lesion segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* 24. Springer. 2021, pp. 206–216.
- [5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.
- [6] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. “Pyramid Scene Parsing Network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).
- [7] Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. “HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation”. In: *Ieee Access* 8 (2020), pp. 120234–120254.
- [8] Joseph W Goodman. “Some fundamental properties of speckle”. In: *JOSA* 66.11 (1976), pp. 1145–1150.
- [9] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 6881–6890.

- [10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 568–578.
- [11] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.
- [12] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).
- [13] Albert Gu and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces". In: *arXiv preprint arXiv:2312.00752* (2023).
- [14] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. "Vmamba: Visual state space model". In: *Advances in neural information processing systems* 37 (2024), pp. 103031–103063.
- [15] Yundong Zhang, Huiye Liu, and Qiang Hu. "Transfuse: Fusing transformers and cnns for medical image segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2021, pp. 14–24.
- [16] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. "Missformer: An effective transformer for 2d medical image segmentation". In: *IEEE transactions on medical imaging* 42.5 (2022), pp. 1484–1494.
- [17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.
- [18] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.
- [19] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. "Cbam: Convolutional block attention module". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.
- [20] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.
- [21] Shin Fujieda, Kohei Takayama, and Toshiya Hachisuka. "Wavelet convolutional neural networks". In: *arXiv preprint arXiv:1805.08620* (2018).
- [22] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. "Multi-level wavelet-CNN for image restoration". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 773–782.
- [23] Shenfu Zhang, Qiang Liu, Zhenhua Zhang, Rui Zhao, Liang Chen, Feng Shao, and Xiangchao Meng. "Spatial-spectral heterogeneity-aware network for hyperspectral and LiDAR joint classification". In: *IEEE Transactions on Neural Networks and Learning Systems* 36.10 (2025), pp. 17766–17780.
- [24] Xiangchao Meng, Shenfu Zhang, Qiang Liu, Gang Yang, and Weiwei Sun. "Uncertain category-aware fusion network for hyperspectral and LiDAR joint classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–15.
- [25] Yu Wang, Quan Zhou, Jia Liu, Jian Xiong, Guangwei Gao, Xiaofu Wu, and Longin Jan Latecki. "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation". In: *2019 IEEE international conference on image processing (ICIP)*. IEEE. 2019, pp. 1860–1864.
- [26] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. "ICNet for Real-Time Semantic Segmentation on High-Resolution Images". In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Cham: Springer International Publishing, 2018, pp. 418–434. ISBN: 978-3-030-01219-9.
- [27] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. "Transunet: Transformers make strong encoders for medical image segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).
- [28] Jeya Maria Jose Valanarasu and Vishal M Patel. "Unetx: Mlp-based rapid medical image segmentation network". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2022, pp. 23–33.
- [29] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. "Missformer: An effective transformer for 2d medical image segmentation". In: *IEEE transactions on medical imaging* 42.5 (2022), pp. 1484–1494.
- [30] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999* (2018).
- [31] Fares Bougourzi, Cosimo Distanto, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. "PDAtt-Unet: Pyramid Dual-Decoder Attention Unet for Covid-19 infection segmentation from CT-scans". In: *Medical Image Analysis* (2023), p. 102797.
- [32] Gaodian Zhou, Weitao Chen, Qianshan Gui, Xianju Li, and Lizhe Wang. "Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), pp. 1–15.
- [33] Yapeng Li, Yong Luo, Lefei Zhang, Zengmao Wang, and Bo Du. "MambaHSI: Spatial-Spectral Mamba for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* (2024).
- [34] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. "Vm-unet: Vision mamba unet for medical image segmentation". In: *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

AUTHORS

ZIHAN XIA is currently studying at Hangzhou Dianzi University, Hangzhou, China. His research interests include SAR image processing, remote sensing image segmentation, and deep learning.

JIN LIU is affiliated with Hangzhou Dianzi University, Hangzhou, China. Dr Liu's research interests include SAR image interpretation, computer vision, and remote sensing image analysis.

ZHIWEN ZHENG is affiliated with Hangzhou Dianzi University, Hangzhou, China. His research interests include image segmentation, computer vision, and deep learning.

XINGRU HUANG is affiliated with Hangzhou Dianzi University, Hangzhou, China. His research interests include remote sensing image analysis, SAR image segmentation, and deep learning.

SVETLANA ILLARIONOVA is affiliated with the Applied AI Institute, Moscow, Russia. Her research interests include computer vision, remote sensing image analysis, and machine learning.

JINFENG LIU is affiliated with Zhengzhou University, Zhengzhou, China. Prof. Liu's research interests include SAR image processing, image segmentation, and computer vision.

WANGMENG ZUO is affiliated with Harbin Institute of Technology, Harbin, China. Prof. Zuo's research interests include computer vision, image restoration, and deep learning.

CHANGQING ZOU is affiliated with Zhejiang Lab, Hangzhou, China. Dr Zou's research interests include remote sensing, computer vision, and intelligent image analysis.

EVGENY BURNAEV is affiliated with the Applied AI Institute, Moscow, Russia. His research interests include machine learning, computer vision, and remote sensing image analysis.