SF-Net: Spatial-Frequency Feature Synthesis for Semantic Segmentation of High-Resolution Remote Sensing Imagery

Chenxu Ge, Qiangkui Leng, Ting Zhang, Samiullah, Abdallah Namoun, *Senior Member, IEEE,* Syed Mudassir Hussain, *Member, IEEE,* Hisham Alfuhaid, *Member, IEEE,* Muhammad Waqas, *Senior Member, IEEE,*

***Abstract*—Precise semantic segmentation of High-Resolution Remote Sensing(HRRS) images is essential for robust environmental surveillance and detailed land use mapping. Despite substantial advances in deep learning, most conventional approaches focus on the spatial domain. This focus often neglects the rich textural and structural nuances found in the frequency domain, which reduces the representation of comprehensive data. Addressing this issue, we introduce SF-Net. This network synthesizes features across spatial and frequency domains, aiming for seamless and effective integration. The core of SF-Net employs a multiscale Convolutional Grouping Fusion Module (CGFM) to extract spatial features at varying resolutions. Following this, the Haar Wavelet Transform decomposes these features into distinct low-frequency components (structure) and high-frequency components (detail). Subsequently, a Mamba-enhanced Global Spatial Feature Extraction Module (GSFEM) reinforces low-frequency semantic information with global context, while a Spatial-Frequency Fusion Module (S-FFM) applies targeted attention to sharpen high-frequency details. Experimental results on the ISPRS Vaihingen, LoveDA, and Potsdam benchmarks confirm SF-Net's superior performance, achieving state-of-the-art mean Intersection over Union (mIoU) scores of 83.12%, 53.28%, and 83.35%, respectively, validating its effectiveness and superority.**

***Index Terms*—Spatial-frequency fusion, state-space model, wavelet transform, semantic segmentation, remote sensing.**

This work is supported by the National Natural Science Foundation of China (61906005). The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number RGP.2/637/46. (*Corresponding author: Ting Zhang.*)

C. Ge and Q. Leng are with the School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China (e-mail: 18803131235@163.com; qkleng@126.com).

T. Zhang is with the School of Computer Science, Beijing University of Technology, Beijing 100124, China (e-mail: zhangting@bjut.edu.cn).

Samiullah is with the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Dir Upper, Pakistan (e-mail: sami@sbbu.edu.pk).

A. Namoun is with the AI Center, Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia (e-mail: a.namoun@iu.edu.sa).

S. M. Hussain is with the School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK (e-mail: s.m.hussain2@herts.ac.uk).

M. Waqas is with the School of Computing and Mathematical Science, Faculty of Engineering and Science, University of Greenwich, SE10 9LS, London, UK (e-mail: engr.waqas2079@gmail.com)

H. Alfuhaid is with the Department of Computer Science, King Khalid University, Abha, Saudi Arabia (e-mail: alasmary@kku.edu.sa).

I. Introduction

Semantic segmentation of remote sensing imagery stands as a cornerstone technology for applications ranging from environmental stewardship to societal development [1]–[3]. The pixel-level classification is one of the most important classifications that assigning an accurate land-cover category to every single pixel. Particularly, it is crucial for automated land-cover mapping and sustainable urban planning [4]. Besides its significance in mapping and urban planning, this classification also supports large-scale environmental monitoring [5], [6] and facilitates prompt decision-making during disaster response [7], [11]. Recently, remote sensing platforms have produced images with increasingly higher spatial resolution. This phenomenon offers remarkable detail of Earth's surface [8]–[10]. However, despite these advancements in application and resolution, obtaining correct segmentation remains difficult. This challenge arises because algorithms must handle diverse scene contents, extreme scale variations, and often blurry or ambiguous boundaries between land-cover classes.

The Convolutional Neural Networks (CNNs) [12] in this field achieved better performance. These convolution operations enable these models to extract spatial features. They are gradually moving from simple patterns to more complex representations. Moreover, some architectures, including the Fully Convolutional Network (FCN) [13], U-Net [14], and context-aware models including PSPNet [15], and DeepLabV3+ [16], provide enhanced spatial feature learning. This can be achieved through important techniques such as pyramid pooling and atrous convolutions. These models generally operate in the spatial domain. Therefore, they frequently neglect frequency-related texture and structural information. This limitation is noticeable when discriminating between visually identical groups, such as "grassland" and "agricultural land" . Although their spatial layouts appear to be identical, their frequency-based textures show slight but significant variances. Consequently, the models that rely purely on spatial information struggle to capture these distinctions accurately. The self-attention processes of the Transformer architecture [17]–[19] represent a revolutionary advancement by enabling the capture of long-range pixel interactions regardless of distance. Although this feature offers a high degree of expressiveness and modeling flexibility, but computing cost is expensive and grows quadratically with the number of pixels. Thus, it is

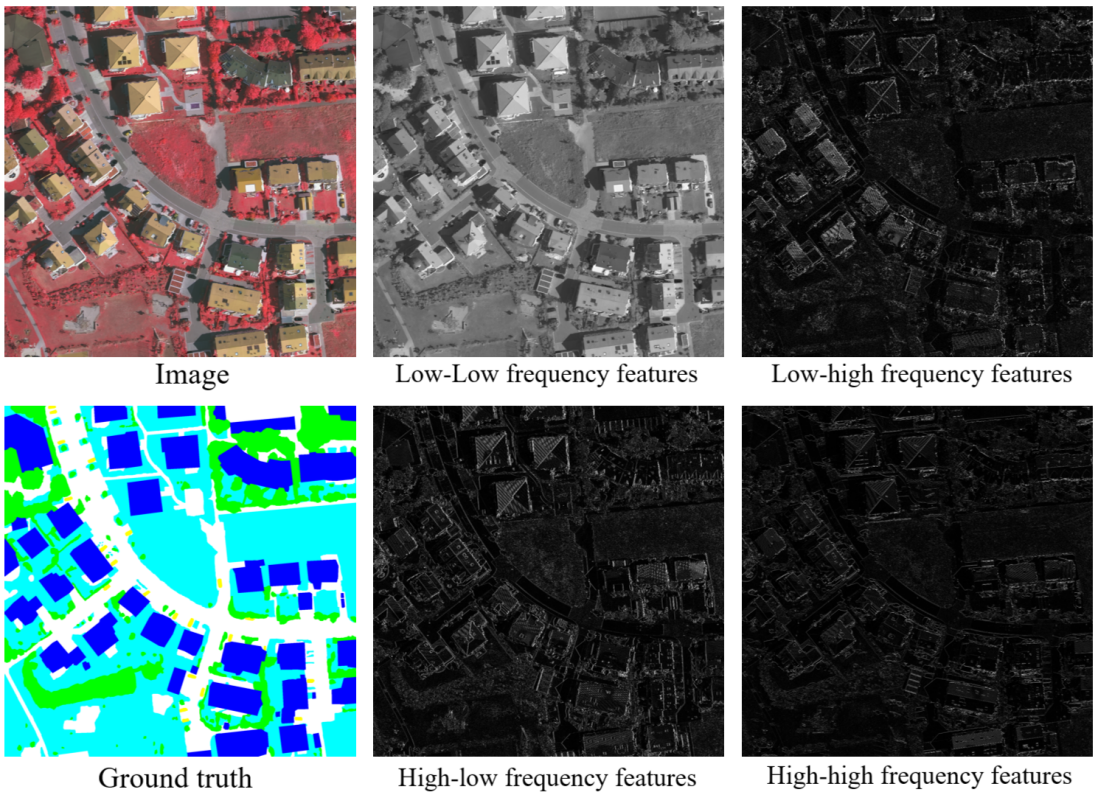

Fig. 1. The original image and its different featuers. The Haar DWT breaks down a remote sensing image into high-frequency features (details and edges) and low-frequency features (global semantics)

impractical for use at the pixel level on high-resolution remote sensing imagery, which restricts its applicability in real-world scenarios [20]–[22].

Recently, the researchers have identified a solution to fundamental trade-off between performance and efficiency. This is achieved via State-space models (SSMs). The Mamba [23], [24] architecture is distinguished among these new designs as a highly influential and effective approach. Mamba effectively models large dependencies throughout the image within linear time by using a selective scanning approach, yielding an optimal combination of a wide receptive field and feasible computational costs. This has opened the door to a new paradigm for HRRS image analysis that is highly efficient. However, even these sophisticated models are still mostly limited to the spatial domain. Frequency-domain analysis, through transforms like the Haar Discrete Wavelet Transform (DWT), provides a complementary perspective [25]–[27]. As depicted in Fig. 1, DWT decomposes an image into low-frequency elements that represent the macroscopic structure and high-frequency elements that capture detailed and textured fine-grained information. This kind of decomposition provides precious information for an all-around understanding of the scene. For instance, when distinguishing between "forest" and "urban areas", the spatial context may reveal their general layout, while frequency analysis can differentiate the repetitive patterns of tree canopies (high-frequency) from the more structured arrangements of buildings and roads (low-frequency) [28]–[30].

Existing methods that incorporate frequency information such as, SFFNet [31], MIFNet [36], and SF-Unet [38], often employ simple concatenation or addition for feature fusion [39]–[41]. This approach treats frequency information as a monolithic block, failing to fully exploit the distinct semantic roles of different frequency bands. Our work hypothesizes that optimal segmentation performance requires a more nuanced and synergistic integration of spatial and frequency information. Accordingly, we propose the deep Spatial-Frequency feature synthesis network (SF-Net). Our approach is not a simple feature concatenation, but a principled framework that orchestrates a deep, cross-domain synthesis. Specifically, SF-Net first employs a CGFM to extract multi-scale spatial features. Following this, it strengthens the long-range context with the Global Spatial Feature Extraction Module (GSFEM) and decouples features into their constituent frequency bands. Subsequently, it performs a differential fusion, enriching the low-frequency structural information with global context, while using an attention mechanism to refine the high-frequency details [42]–[44].

- We present SF-Net, a deep Spatial-Frequency feature synthesis network for HRRS imagery that establishes a new performance benchmark by synergistically integrating global spatial context with discrete wavelet frequency features.
- We design the Convolutional Grouping Fusion Module (CGFM), which employs a progressive fusion strategy to efficiently capture hierarchical scale and detail information while mitigating feature degradation.
- We construct the GSFEM, which combines the Mamba

SSM with a multi-pooling strategy to jointly enhance global semantic understanding and local feature saliency.
- We devise the Spatial-Frequency Fusion Module (S-FFM), which performs a differential fusion of spatial and frequency-domain features. By decoupling and selectively enhancing frequency components, it provides the decoder with a uniquely discriminative representation.
- Extensive experimental results on three publicly available benchmarks demonstrate that the SF-Net outperforms the existing state-of-the-art approaches, thereby validating the efficacy of our cross-domain fusion strategy.

II. Related Work

This section conducts an in-depth review of previous research in three main aspects: semantic segmentation of remote sensing images, multi-scale feature extraction, and feature fusion techniques.

A. Semantic Segmentation of Remote Sensing Images

CNN-based architectures serve as the groundwork for contemporary image segmentation techniques. Pioneering works such as FCN [13], U-Net [14], PSPNet [15], and DeepLabV3+ [16] have continuously advanced the progress of spatial feature extraction by introducing encoder-decoder structures, multi-scale pooling, and atrous convolutions. In contrast, EAAH-Net [45] generalizes convolutional neural network operators in hyperbolic space by leveraging the Lorentz model, so as to learn low-distortion representations of remote sensing images. DTUML [46] proposes to leverage the theory of dilated convolution to expand the receptive field without increasing parameter overhead, thereby achieving more accurate reconstruction of spatial and spectral information. XINet [47] adopts a coupled design of two disjoint U-Nets to achieve powerful multi-scale and multi-depth feature representation. Although proficient at recognizing localized patterns, their intrinsically limited receptive fields impede the modeling of non-local dependencies. This deficiency often yields disjointed predictions for large objects and an incomplete grasp of the overall scene context [48]–[50].

Transformer-based architectures subsequently arose to overcome this restriction on local context. By utilizing self-attention, models such as MAResU-Net [51] and DC-Swin [52] effectively model global relationships, resulting in more structurally coherent segmentation. Nevertheless, this expressive capability is inextricably linked to quadratic computational complexity, posing a restrictive challenge for processing megapixel HRRS images and thus compromising their feasibility in practical settings.

Mamba-based architectures aims to address this efficiency challenge. Through the incorporation of a state-space model and a selective scan mechanism, Mamba [23] and its variants [53], [54] are capable of efficiently modeling long-range dependencies with linear complexity, thereby creating a new frontier for both efficient and powerful spatial modeling. These models still don't fully use frequency features. Frequency analysis is very good at capturing the textures and repeating patterns prevalent in remote sensing images. Early studies, such as SFFNet [31], MIFNet [36], [37], and SFUnet [38], have shown that their fusion algorithms are usually simple, involving only the concatenation or addition of spatial features. In contrast, FDNet [55] independently refines high-frequency and low-frequency components in the frequency domain through the Sparse-aware Spectrum Enhancement Module, Frequency Decoupling Attention Module, and Attention Frequency Context Module to enhance feature representation, focusing on improving the utilization efficiency of frequency-domain features. DDFNet [56] integrates spatial and frequency domain features via the DDFF module, and improves boundary accuracy and geometric consistency using the HGPG module with gradient and curvature information. FreDNet [57] suppresses frequency-domain noise and preserves edge structures via the Dual-path Residual Block integrating FDM and FFM, and achieves adaptive encoder-decoder feature fusion with the Frequency-aware Cross-level Fusion Module. GPINet [58] leverages the local-global interaction module to enhance feature communication across diverse data types in RSI, enabling a more refined understanding of spatial and spectral properties. SF-Net takes a step further by using Haar discrete wavelet transform to initially separate frequency features, and then adopting a sophisticated multi-step method to combine them with spatial features enhanced by Mamba. This new method of fusing spatial and frequency domains significantly improves the accuracy and efficiency of the final segmentation map in high-resolution remote sensing tasks.

B. Multi-scale Feature Extraction

It requires an advanced method to analyze high-resolution remote sensing imagery to effectively extract and synthesize features at various levels [59]–[61]. The initial Feature Pyramid Network (FPN) [62] laid down a fundamental framework for the integration of multi-scale features, thus allowing subsequent innovations. Considering this, researchers have adopted attention mechanisms more frequently to enable feature aggregation that is more dynamic and adaptive. For instance, Attentional Feature Fusion (AFF) assigns weights to features from various scales intelligently. Similarly, MSCAM [63] and MSAF [64] integrate multi-scale contextual features into their attention-based modules without any disruption. However, the core of these strategies usually relies on fundamental element-wise addition or combination at the time of fusion. The small, fine details from the early layers can be destroyed by this simple approach. To address these limitations, our SF-Net employs the Convolutional Grouping Fusion Module (CGFM). It promotes a gradual and controlled exchange between features at different scales. The CGFM contributes to managing complex segmentation tasks with better fidelity. While ensuring that complex details are not only retained but also effectively utilized throughout the entire fusion operation.

C. Feature Fusion

Advanced segmentation networks primarily use feature fusion. Well-known techniques often fall into two broad categories: *intra-domain* and *cross-domain*. The goal of intra-domain fusion is to improve the overall representational power

of features limited to the spatial domain. This is achieved through a variety of methods, ranging from global pooling mechanisms [65] to advanced graph-based architectures [31], [32]. The methods enable integration of wide-ranging scene-level insights. In this context, our Global Spatial Feature Enhancement Module (GSFEM) advances the limits further by creating a synergistic relationship between Mamba's effective, linear-time management of long-range dependencies and the additional insights obtained from combined global average and max pooling. By means of this strategic integration, a more sophisticated and powerful global feature set can be produced with a strikingly low computational cost [33]–[35]. Hence, optimizing performance while maintaining efficiency.

Cross-domain fusion fuses and synergistically optimizes heterogeneous information to obtain a global representation with acceptable computational overhead. Modules like CCM [53] and FFM [66] adopt attention mechanisms for flexible cross-branch feature fusion, while recent methods (e.g., FAFF [67], HFFM [68]) integrate frequency information into spatial feature learning. However, these spatial-frequency models share a critical limitation: their fusion relies on a simple concatenation strategy. FAFF directly concatenates Fourier-extracted frequency features with spatial features along the channel dimension without differentiating frequency components' semantic contributions; HFFM employs a concatenation-based layer for multi-scale frequency-spatial fusion but fails to establish targeted interactions between complementary information. This undifferentiated approach overlooks the distinct semantic roles of low-frequency structures and high-frequency textures, leading to suboptimal complementary feature fusion.

To address this issue, we propose the S-FFM module for deep, targeted spatial-frequency fusion. Unlike the simple concatenation in FAFF and HFFM, S-FFM first decomposes spatial features into semantically distinct low- and high-frequency components via Haar wavelet transform. It then deploys separate enhancement pathways: low-frequency components are refined to boost global structural consistency, while high-frequency components are enhanced to accentuate discriminative local textures, thus enabling precise, adaptive fusion of complementary spatial-frequency features.

III. METHOD

In order to fully utilize the complementary characteristics of spatial and frequency-domain information for HRRS image segmentation, we put forward the Spatial-Frequency feature synthesis network (SF-Net). Its overall architecture is depicted in Fig. 2. As presented in Fig. 2, SF-Net is structured as an encoder-decoder framework. We adopt ResNet18 [69] as the encoder - initialized with ImageNet pre-trained weights and not frozen during training; instead, we continue fine-tuning these weights to extract a hierarchy of spatial features. These features are then processed by three core modules developed in this work: the CGFM, the GSFEM, and the S-FFM. In the proposed work, the CGFM module enhances the encoder's output by integrating information across multiple scales, thereby providing more details of the input data. The S-FFM successfully combines spatial and frequency-domain characteristics to provide a more comprehensive feature set, whereas the GSFEM component efficiently extracts a comprehensive global context vector that includes crucial overarching qualities. The decoder receives this refined feature map and uses the UNetFormer architecture [70] to produce the final high-resolution pixel-wise segmentation map.

A. Convolutional Grouping Fusion Module (CGFM)

We suggest the CGFM to extract multi-level features while preserving fine-grained details. Its architecture, as shown in Fig. 3, includes a progressive fusion technique to ensure thorough information exchange within a parallel, multi-branch structure that captures features across multiple receptive fields.The module takes in a feature map $\boldsymbol{F}_{R_r}$ from the r-th stage of the encoder. A set of parallel depthwise separable convolutions with varying kernel sizes $k \in \{1, 3, 5, 7\}$ process $\boldsymbol{F}_{R_r}$, producing:

$$\boldsymbol{F}^k_{R_r} = \text{ReLU}(DWConv_{k\times k}(\boldsymbol{F}_{R_r})), \quad k \in \{1, 3, 5, 7\} \quad (1)$$

Each resulting feature map is split along the channel dimension:

$$\boldsymbol{A}^k_{R_r}, \boldsymbol{B}^k_{R_r} = \text{Split}(\boldsymbol{F}^k_{R_r}), \quad k \in \{1, 3, 5, 7\} \quad (2)$$

where $DWConv_{k\times k}(\cdot)$ denotes a depthwise separable convolution with kernel size k, and Split$(\cdot)$ divides the feature map into two channel-wise groups.

These split features are then fused in a progressive, grouped manner. The process is carried out in the following way. The first group $\boldsymbol{A}^k_{R_r}$ and the second group $\boldsymbol{B}^k_{R_r}$ are independently concatenated and then passed through a 1×1 convolution:

$$\boldsymbol{F}^1_{R_r} = \text{ReLU}(\text{BN}(DWConv_{1\times1}(\text{Cat}(\boldsymbol{A}^1_{R_r}, \boldsymbol{A}^3_{R_r}, \boldsymbol{A}^5_{R_r}, \boldsymbol{A}^7_{R_r})))) \quad (3)$$

$$\boldsymbol{F}^2_{R_r} = \text{ReLU}(\text{BN}(DWConv_{1\times1}(\text{Cat}(\boldsymbol{B}^1_{R_r}, \boldsymbol{B}^3_{R_r}, \boldsymbol{B}^5_{R_r}, \boldsymbol{B}^7_{R_r})))) \quad (4)$$

where the notation Cat$(\cdot)$ represents channel-wise concatenation, BN$(\cdot)$ stands for batch normalization, and ReLU$(\cdot)$ is the activation function.

Ultimately, the two merged groups are concatenated once again and then passed through a final 1×1 convolution. The CGFM produces output by creating a residual link with the initial input:

$$\boldsymbol{F}_{C_r} = \text{ReLU}(\text{BN}(DWConv_{1\times1}(\text{Cat}(\boldsymbol{F}^1_{R_r}, \boldsymbol{F}^2_{R_r})))) + \boldsymbol{F}_{R_r} \quad (5)$$

where $\boldsymbol{F}_{C_r} \in \mathbb{R}^{C\times H\times W}$ represents the final output of CGFM at stage r.

B. Global Spatial Feature Extraction Module (GSFEM)

As shown in Fig. 4, the GSFEM module preserves both global contextual information and local features. By combining the Mamba State Space Model (SSM) with the dual pooling attention mechanism, it achieves a thorough understanding of global scene semantics. Inspired by the 2D Selective Scan (SS2D) module of VMamba [75], the GSFEM first processes the input feature Fcr through a Feed-Forward Mamba (F-FM) module, as illustrated in Fig. 5. The normalized 2D

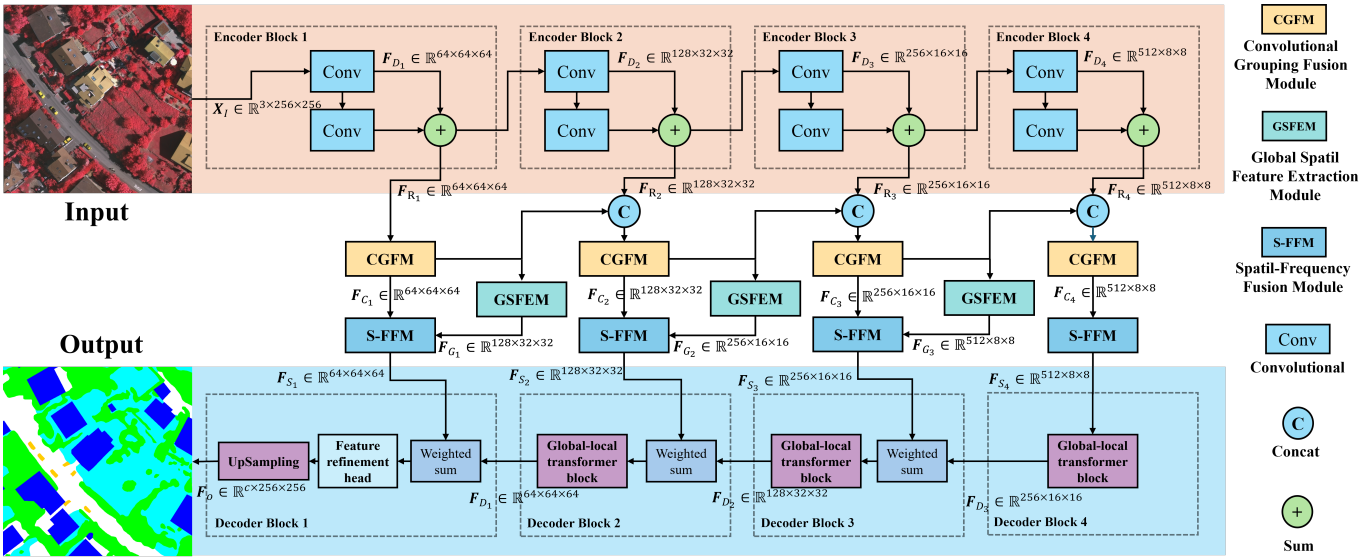

Fig. 2. The overall SF-Net architecture, showing the path from the ResNet18 encoder to the final decoder via the central CGFM, GSFEM, and S-FFM modules.

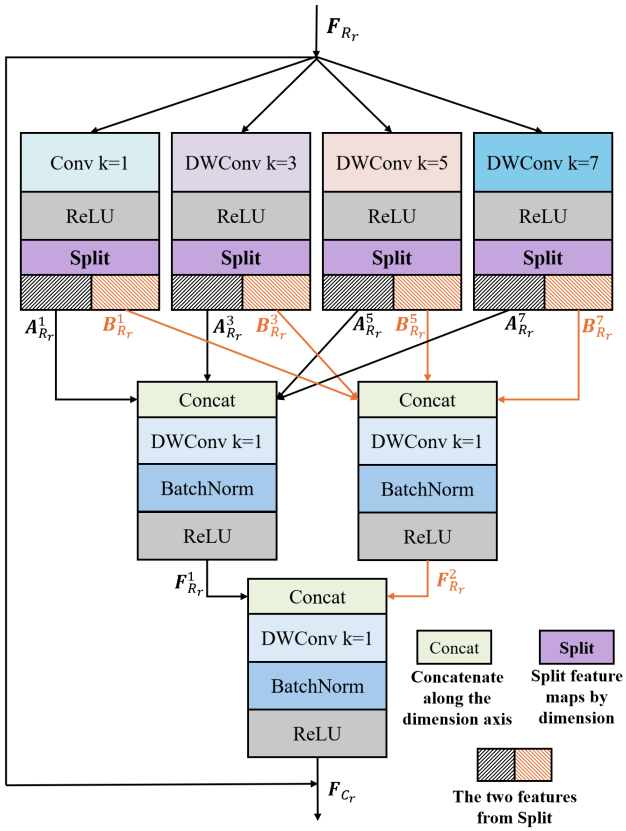

Fig. 3. Detailed structure of the CGFM, showcasing the parallel convolutions, channel splitting, and progressive fusion.

feature map is split into 4 one-dimensional (1D) sequences along four distinct scanning paths (row-wise left-to-right, row-wise right-to-left, column-wise top-to-bottom, and column-wise bottom-to-top) via the expand$(\cdot)$ operation. Subsequently, the S6 module processes these four independent sequences using scans in the four different directions, yielding $\boldsymbol{F}'_{1_r}$, $\boldsymbol{F}'_{2_r}$, $\boldsymbol{F}'_{3_r}$ and $\boldsymbol{F}'_{4_r}$. Finally, the results of the four paths are merged through the merge$(\cdot)$ operation to reconstruct $\boldsymbol{F}'_{1_r}$, $\boldsymbol{F}'_{2_r}$, $\boldsymbol{F}'_{3_r}$ and $\boldsymbol{F}'_{4_r}$ back into a 2D feature map, and the features are refined via the Feed-Forward Networks(FFN). This block significantly reduces long-range dependencies with linear complexity using the following steps:

$$\boldsymbol{F}_{\nu_r} = \text{expand}(\text{LN}(\boldsymbol{F}_{C_r}), v), \quad v \in \{1,2,3,4\} \tag{6}$$

$$\boldsymbol{F}'_{v_r} = S6(\boldsymbol{F}_{v_r}) \tag{7}$$

$$\boldsymbol{F}'_r = \text{LN}(\text{merge}(\boldsymbol{F}'_{1_r}, \boldsymbol{F}'_{2_r}, \boldsymbol{F}'_{3_r}, \boldsymbol{F}'_{4_r})) \tag{8}$$

$$\boldsymbol{F}'_{f_r} = \text{FFN}(\boldsymbol{F}'_r) + \boldsymbol{F}'_r \tag{9}$$

where LN$(\cdot)$ represents layer normalization, $S6(\cdot)$ is the selective scan operator, expand$(\cdot)$ and merge$(\cdot)$ are directional scan expansion and fusion, respectively, v represents the 4 distinct scanning expansion directions.

To supplement Mamba's global modeling, the output $\boldsymbol{F}'_{f_r}$ is processed via two parallel pooling branches: Global Average Pooling (GAP) and Global Max Pooling (GMP), resulting:

$$\boldsymbol{F}_r^A = \text{GAP}(\boldsymbol{F}'_{f_r}), \quad \boldsymbol{F}_r^M = \text{GMP}(\boldsymbol{F}'_{f_r}) \tag{10}$$

$$\boldsymbol{F}_r^{AC}, \boldsymbol{F}_r^{MC} = \text{Split}(\text{Sigmoid}(Conv_{1\times1}(\text{Cat}(\boldsymbol{F}_r^A, \boldsymbol{F}_r^M)))) \tag{11}$$

These adaptive attention maps are applied to $\boldsymbol{F}'_{f_r}$, yielding the final globally enhanced representation:

$$\boldsymbol{F}_{G_r} = \boldsymbol{F}'_{f_r} + \boldsymbol{F}'_{f_r} \otimes \boldsymbol{F}_r^{AC} + \boldsymbol{F}'_{f_r} \otimes \boldsymbol{F}_r^{MC} \tag{12}$$

where $\otimes$ denotes element-wise multiplication and $\boldsymbol{F}_{G_r}$ is the output of GSFEM at stage r.

C. Spatial-Frequency Fusion Module (S-FFM)

Features in the spatial and frequency domains often exhibit different distributions and semantic properties [71], [72]. Directly fusing these heterogeneous features can result in information loss or ambiguity, thus deteriorating the performance of the model [73], [74]. To tackle this issue, we put forward

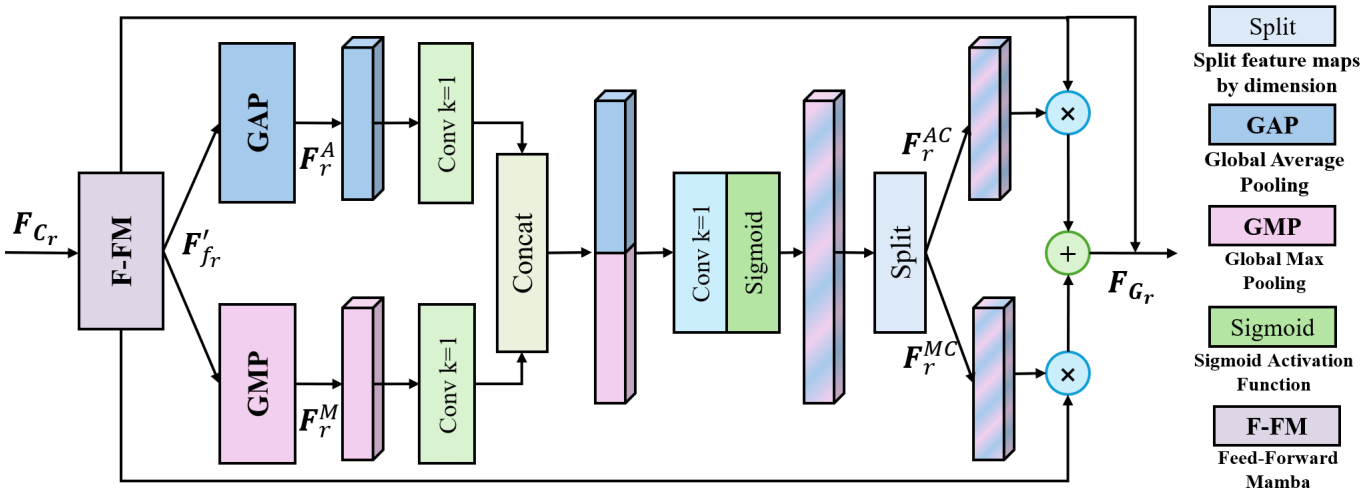

Fig. 4. The GSFEM's detailed structure, emphasizing the dual-pooling attention mechanism and the Feed-Forward Mamba (F-FM) block.

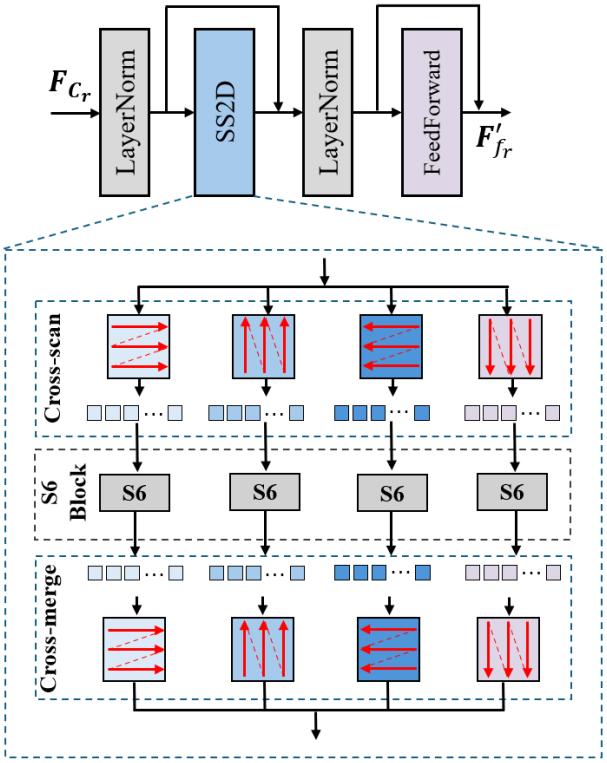

Fig. 5. Structure of the Feed-Forward Mamba (F-FM) block employs the 2D Selective Scan technique. SS2D establishes global-local feature connections with linear complexity via a pipeline of cross-scan splitting, directional feature encoding, and cross-merge reconstruction, where cross-scan splits the 2D feature map into four 1D sequences along distinct directions, S6 Block encodes each directional 1D sequence, and cross-merge merges and reconstructs the encoded 1D sequences back into a 2D feature map.

a strategy that effectively integrates cross-domain features to make full use of their complementary nature. At the heart of this strategy is the S-FFM. As illustrated in Fig. 6, the S-FFM employs the Haar DWT as its core operator. The module receives multi-scale features from CGFM and the global context vector from GSFEM. First, the discrete wavelet transform (DWT) (refer to Fig. 7) disassembles the spatial feature map $\boldsymbol{F}_{C_r} \in \mathbb{R}^{C\times H\times W}$ into four frequency sub-bands, namely one low-frequency approximation ($\boldsymbol{F}_r^{LL} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$) and three high-frequency detail components ($\boldsymbol{F}_r^{LH} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$, $\boldsymbol{F}_r^{HL} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$, $\boldsymbol{F}_r^{HH} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$). The low - frequency sub-band encapsulates the structural core of the image, whereas the high-frequency bands seize the fine-grained edge and texture details.The relevant formulas are as follows:

$$\boldsymbol{F}_r^{LL}, \boldsymbol{F}_r^{LH}, \boldsymbol{F}_r^{HL}, \boldsymbol{F}_r^{HH} = DWT\left(\boldsymbol{F}_{C_r}\right) \tag{13}$$

In order to obtain these components, we employ a differential approach. The three high-frequency components are first combined and then successively refined by a channel attention mechanism (CAM) and a spatial attention mechanism (SAM), which results in an enhanced high-frequency representation:

$$\boldsymbol{F}_r^{High} = SAM\left(CAM\left(Cat\left(\boldsymbol{F}_r^{LH}, \boldsymbol{F}_r^{HL}, \boldsymbol{F}_r^{HH}\right)\right)\right) \tag{14}$$

where $Cat(\cdot)$ denotes channel-wise concatenation, $CAM(\cdot)$ and $SAM(\cdot)$ denote the channel and spatial attention mechanisms, respectively, and $\boldsymbol{F}_r^{High} \in \mathbb{R}^{3C\times \frac{H}{2}\times \frac{W}{2}}$ is the attention-refined high-frequency representation. In parallel, the semantically rich low-frequency sub-band $\boldsymbol{F}_r^{LL}$ is deeply fused with the global context vector $\boldsymbol{F}_{G_r}$ from GSFEM. These are concatenated and passed through a 1×1 convolution to facilitate structural enhancement:

$$\boldsymbol{F}_r^{Low} = Conv_{1\times 1}\left(Cat\left(\boldsymbol{F}_r^{LL}, \boldsymbol{F}_{G_r}\right)\right) \tag{15}$$

Finally, the fused low-frequency feature $\boldsymbol{F}_r^{Low} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$ and the enhanced high-frequency representation $\boldsymbol{F}_r^{High} \in \mathbb{R}^{3C\times \frac{H}{2}\times \frac{W}{2}}$ (split back into three sub-bands) are recombined using the Haar Inverse Discrete Wavelet Transform (IDWT).This reconstructs a spatial domain feature map that includes both detailed textures and global contextual signals. To improve segmentation accuracy, the S-FFM's cross-domain synthesis creates a discriminative and semantically rich representation for the decoder.

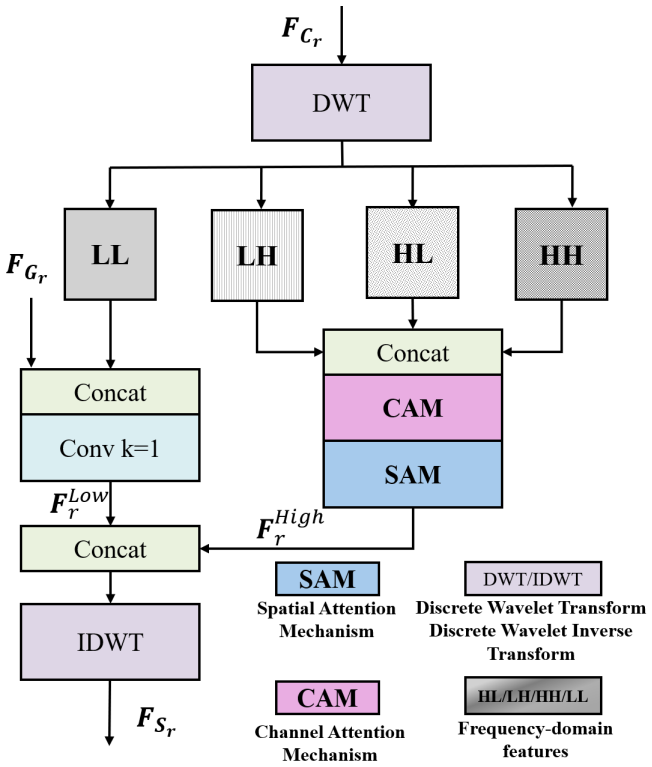

Fig. 6. Detailed structure of the S-FFM, illustrating the DWT decomposition, differential enhancement, and IDWT reconstruction.

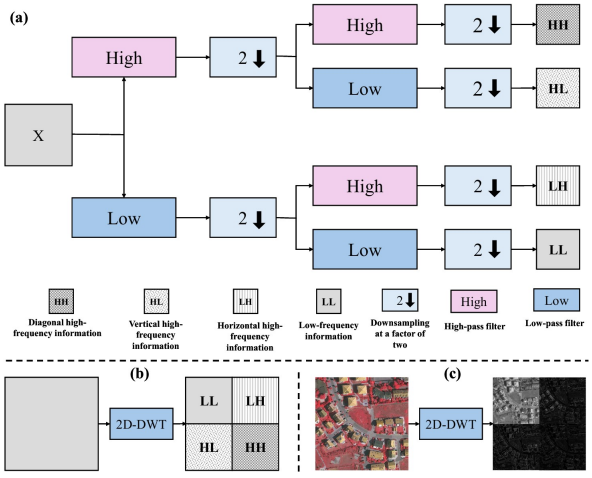

Fig. 7. The Haar DWT process. (a) Basic workflow. (b) Sub-band decomposition. (c) Example decomposition of an image.

D. Loss Function

The network is optimized by reducing the conventional pixel-wise cross-entropy loss, which quantifies the difference between predicted and true class distributions. The loss is defined as follows:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} \boldsymbol{y}_{n,c}\log(\boldsymbol{p}_{n,c}), \tag{16}$$

here, N denotes the total number of pixels in the input image, C represents the count of semantic classes, $\boldsymbol{y}_{n,c}$ denotes the one-hot encoded ground truth for pixel n and class c, and $\boldsymbol{p}_{n,c}$ shows the corresponding predicted class probability.

IV. Experimental Results

We evaluated SF-Net's performance with a comprehensive series of experiments. This section describes the details considered metrics, chosen datasets, and the implementation parameters. After providing an introduction, we show quantitative and qualitative results compared to current approaches.

A. Datasets

ISPRS Vaihingen [76]: This benchmark contains 16 high-resolution aerial images annotated with 6 land-cover classes. Following standard protocol, 12 images were used for training and 4 for testing. Images were cropped into 256×256 patches with a stride of 32 pixels for training.

LoveDA [77]: This large-scale dataset comprises 5,987 images of size 1024×1024 pixels, covering diverse urban and rural scenes with 7 annotated land-cover classes. We used the official training and testing split and extracted 256×256 patches with a stride of 64 pixels for training.

ISPRS Potsdam [78]: This dataset consists of 38 ultra-high-resolution aerial images (5 cm GSD), annotated with 6 land-cover classes. We followed the standard split, using 24 images for training and 14 for testing, and applied a 256×256 patch size with a stride of 32 pixels.

B. Implementation Details

All experiments were implemented using the PyTorch framework and executed on an NVIDIA GeForce RTX 3090 GPU. We used the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. We employed a polynomial learning rate decay scheduler with power = 0.9. The batch size was set to 8 for the Vaihingen and Potsdam datasets and 10 for the LoveDA dataset. All models were trained over 50 epochs. We used standard data augmentation techniques during training, including random horizontal flipping and random rotation within the range of $[-15°, 15°]$. We run the model five times, the reported their mean and standard deviation. Model performance was assessed using two commonly used metrics: (1) **Mean F1-score (mF1)**: the average F1-score across all classes, and (2) **Mean Intersection over Union (mIoU)**: the average IoU across all semantic classes.

C. Comparison with State-of-the-Art Methods

We compared SF-Net to leading segmentation models, including CNN, Transformer, and Mamba architectures. Tables I, II, and III show quantitative results, whereas Figures 8, 9, and 10 present qualitative representations.

ISPRS Vaihingen Dataset: SF-Net achieved a top-performing mIoU of 83.68% and mF1 of 91.57%,as shown in Table I. Performance on the challenging 'Car' class, a small-object category was particularly impressive, reaching 85.21%. This results in gain of 3.01% to 14.7% over the compared methods. This increase is attributed to the planned

TABLE I
RESULTS OF DIFFERENT METHODS ON THE ISPRS VAIHINGEN DATASET.

Type	Methods	Imp.surf.	Building	Low.Veg.	Trees	Cars	mF1	mIoU	FLOPs	Parameters
Spatial	FCN [13]	83.76±0.93	91.35±1.10	65.45±1.32	81.58±0.53	79.59±1.75	88.53±1.31	80.35±1.28	69.45	32.96
	UNet [14]	84.73±1.62	91.52±0.75	64.82±2.37	80.75±1.55	77.26±2.31	88.42±1.77	80.18±1.86	109.52	31.04
	PSPNet [15]	84.71±1.13	91.81±0.58	65.33±0.77	82.20±0.96	82.18±0.43	89.45±0.41	81.25±0.38	92.45	46.71
	DeepLabV3+ [16]	85.63±0.42	92.44±1.13	65.38±1.52	82.49±0.89	78.75±1.94	89.07±1.15	80.94±1.36	13.23	5.81
	MACUNet [79]	83.62±0.43	90.75±1.21	65.44±1.17	82.20±1.01	74.58±0.78	88.39±0.54	79.32±0.66	16.80	**5.15**
	A2FPN [80]	85.66±0.55	93.28±0.22	66.45±0.31	82.79±0.37	81.11±1.53	90.01±0.49	81.86±0.32	20.92	22.82
	MAResU-Net [51]	85.37±0.88	92.40±0.92	66.28±0.26	82.33±1.14	81.57±0.83	89.95±0.74	81.59±0.81	17.55	26.28
	MANet [81]	84.31±0.59	90.75±1.10	64.77±1.58	82.20±0.58	81.64±0.69	88.82±0.88	80.73±0.72	38.90	35.86
	CMTFNet [82]	85.39±0.45	92.67±0.71	66.38±0.74	82.53±0.56	80.25±1.32	89.72±0.72	81.44±0.53	17.14	30.07
	DC-Swin [52]	82.01±1.06	89.27±1.51	62.83±0.75	81.44±1.21	73.03±0.91	86.56±0.57	77.72±0.45	25.29	45.61
	PyramidMamba [54]	85.24±1.11	92.16±0.83	66.58±0.15	82.55±0.37	80.57±1.43	89.78±0.76	81.42±0.73	59.14	109.99
	RS3Mamba [53]	85.26±0.53	91.97±0.41	66.27±0.21	81.09±1.05	80.33±2.13	89.18±0.92	80.98±0.85	31.65	49.66
Spatial-Frequency	SFUNet [38]	84.51±0.81	92.10±1.42	64.35±1.24	81.62±1.13	76.61±1.06	87.53±0.78	79.84±0.71	151.61	28.84
	MFMSNet [83]	85.41±0.67	91.76±1.31	65.73±0.53	82.53±0.38	80.26±0.55	89.62±0.55	81.14±0.73	46.79	58.06
	TexLiverNet [84]	81.72±0.63	92.11±0.28	66.18±1.12	80.77±0.63	70.51±1.71	86.55±0.72	78.23±0.83	77.95	40.07
	MEW-UNet [85]	84.95±1.28	90.76±0.76	67.21±0.13	82.57±1.23	73.88±1.48	88.05±0.39	79.87±0.40	81.05	138.89
	MIFNet [36]	85.57±0.89	92.79±1.13	65.05±0.87	82.20±0.53	78.79±2.78	89.07±1.37	80.88±1.32	17.71	25.35
	SFFNet [31]	84.73±1.02	91.43±0.31	66.45±0.31	81.75±0.85	79.85±1.02	89.12±0.52	80.84±0.65	25.84	34.16
	WaveMix [86]	84.64±1.14	89.79±1.78	66.13±0.42	81.55±1.10	81.47±1.36	89.07±1.28	80.72±1.42	37.52	38.85
	AFENet [87]	85.27±0.44	90.85±1.01	65.38±1.75	80.46±0.72	80.11±1.58	88.75±1.16	80.41±1.07	12.80	20.23
	SF-Net(Ours)	**86.67±0.52**	**93.80±1.21**	**68.75±0.85**	**83.95±0.43**	**85.21±0.72**	**91.57±0.59**	**83.68±0.67**	**9.48**	23.01

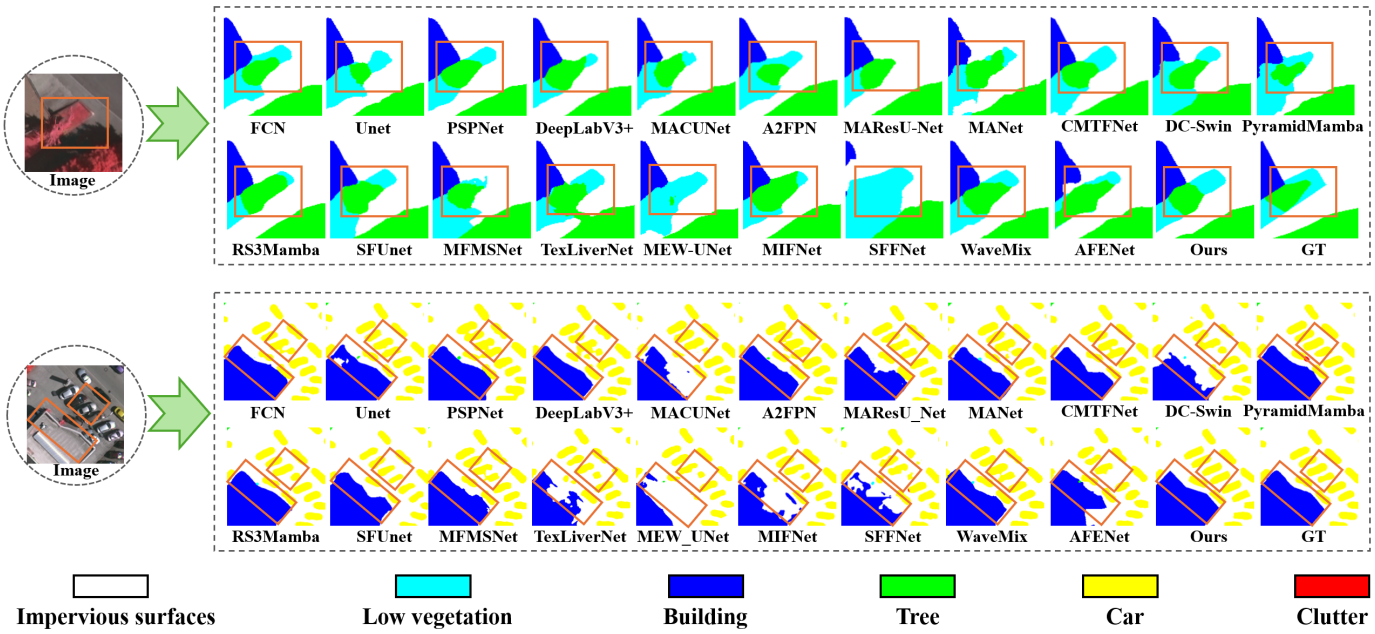

Fig. 8. Comparison of segmentation examples using various approaches on the ISPRS Vaihingen dataset.

TABLE II
RESULTS OF DIFFERENT METHODS ON THE LOVEDA DATASET.

Type	Methods	Back.	Buil.	Road	Water	Bar.	For.	Agri.	mF1	mIoU	FLOPs	Parameters
Spatial	FCN [13]	40.33±1.32	61.08±0.91	54.51±1.23	56.30±1.48	36.65±2.53	47.20±2.19	52.12±0.20	66.01±1.41	49.74±1.35	69.45	32.96
	UNet [14]	40.64±0.56	59.47±0.39	53.82±1.52	59.17±1.36	31.88±0.82	49.99±0.16	48.20±1.76	65.20±1.04	49.02±1.16	109.52	31.04
	PSPNet [15]	42.41±0.83	58.62±1.28	55.12±0.39	61.51±0.26	40.95±1.18	49.73±0.62	51.70±0.30	67.69±0.39	51.43±0.41	92.45	46.71
	DeepLabV3+ [16]	41.62±1.27	60.57±0.53	56.70±0.43	62.29±0.32	25.78±2.55	44.87±1.36	49.33±0.84	63.96±1.22	48.74±1.37	13.23	5.81
	MACUNet [79]	39.50±0.38	57.31±1.37	53.19±0.47	59.75±0.33	29.83±2.39	48.42±0.25	44.43±0.55	63.50±0.94	47.49±0.77	16.80	**5.15**
	A2FPN [80]	41.51±0.55	60.78±0.83	57.21±0.29	61.72±0.38	40.25±1.17	49.28±0.39	51.11±0.88	67.88±0.69	51.69±0.56	20.92	22.82
	MAResU-Net [51]	40.71±0.75	61.00±0.48	55.29±1.01	61.31±0.22	40.21±0.38	48.53±0.63	45.71±2.25	66.37±1.27	50.39±0.89	17.55	26.28
	MANet [81]	42.25±0.58	59.93±0.37	56.26±0.44	60.77±0.31	40.41±0.17	46.15±1.41	48.63±0.50	66.82±1.35	50.63±1.26	38.90	35.86
	CMTFNet [82]	41.55±1.07	60.27±0.53	51.83±1.34	61.38±0.44	35.73±1.78	50.05±0.25	50.56±0.61	66.35±0.80	50.20±0.69	17.14	30.07
	DC-Swin [52]	41.73±1.11	59.32±0.42	55.20±0.77	59.86±1.52	41.37±0.51	49.88±0.43	46.67±1.81	66.68±1.02	50.58±0.95	25.29	45.61
	PyramidMamba [54]	42.72±0.33	54.90±1.27	54.12±1.78	59.95±0.38	40.27±0.93	46.81±1.56	47.56±0.83	65.26±0.94	49.48±1.10	59.14	109.99
	RS3Mamba [53]	41.78±1.22	61.63±0.39	57.72±0.62	61.86±0.70	39.64±1.14	48.30±0.66	50.11±1.21	67.83±0.82	51.58±0.73	31.65	49.66
Spatial-Frequency	SFUNet [38]	42.02±0.91	61.45±0.65	56.92±1.38	59.18±1.71	32.65±2.12	50.33±0.82	47.17±1.18	65.78±0.92	49.96±1.07	151.61	28.84
	MFMSNet [83]	40.20±1.31	53.40±1.55	55.78±1.24	59.83±1.32	41.07±0.72	40.45±0.65	50.02±0.70	64.39±1.38	48.68±0.65	46.79	58.06
	TexLiverNet [84]	42.15±0.79	58.61±1.42	57.79±0.47	60.11±1.59	40.46±0.43	45.36±1.70	48.61±1.86	66.57±0.51	50.44±0.41	77.95	40.07
	MEW-UNet [85]	43.02±0.35	61.23±0.62	53.86±1.84	61.77±0.53	38.45±1.97	47.95±1.42	51.16±0.27	67.18±0.65	51.06±0.42	81.05	138.89
	MIFNet [36]	42.61±0.82	61.50±0.64	57.21±0.35	61.39±0.57	39.15±1.53	50.56±0.31	48.23±1.41	67.69±0.81	51.52±0.93	17.71	25.35
	SFFNet [31]	41.86±1.37	60.45±1.23	55.20±1.71	59.26±0.53	41.33±0.69	48.76±0.93	47.21±0.85	66.78±1.26	50.58±1.42	25.84	34.16
	WaveMix [86]	42.26±0.61	61.63±0.35	57.11±0.26	61.28±0.29	39.96±1.62	49.84±0.52	47.80±0.79	67.53±0.57	51.41±0.63	17.71	25.35
	AFENet [87]	40.72±1.16	59.58±0.73	54.65±0.92	59.26±1.37	41.43±0.38	48.72±1.01	47.29±1.92	66.36±0.76	50.24±0.81	12.80	20.23
	SF-Net(Ours)	**43.35±1.02**	**62.51±0.86**	**58.76±0.41**	**62.72±0.26**	**42.10±0.33**	**51.34±0.71**	**52.58±1.36**	**69.28±0.39**	**53.34±0.47**	**9.48**	23.01

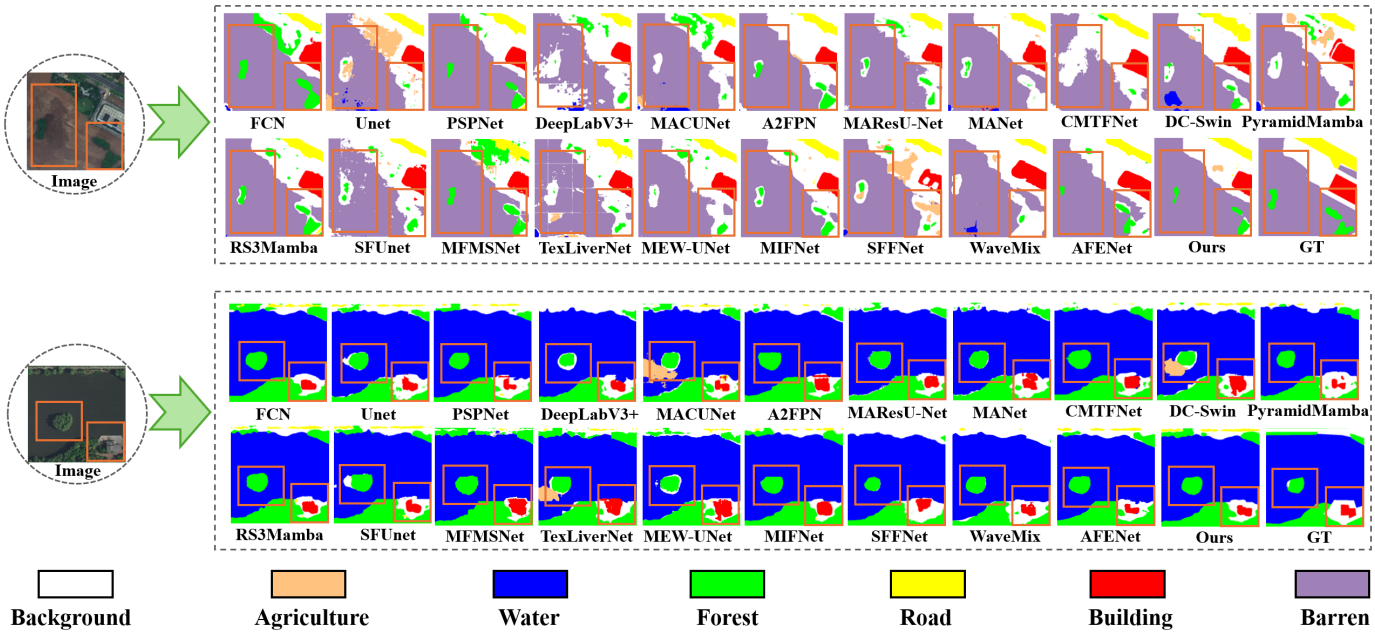

Fig. 9. Segmentation examples of different methods on the LoveDA dataset.

TABLE III
RESULTS OF DIFFERENT METHODS ON THE ISPRS POTSDAM DATASET.

Type	Methods	Imp.surf.	Building	Low.Veg.	Trees	Cars	mF1	mIoU	FLOPs	Parameters
Spatial	FCN [13]	79.13±1.25	83.17±2.04	73.75±0.90	76.12±0.81	88.24±0.37	88.51±1.41	80.08±1.23	69.45	32.96
	UNet [14]	78.32±2.17	84.26±1.93	71.21±1.51	71.52±2.20	85.71±2.31	87.12±1.98	78.20±2.10	109.52	31.04
	PSPNet [15]	81.71±1.07	87.79±0.64	74.30±0.82	76.11±0.52	86.79±0.90	89.58±1.17	81.34±0.86	92.45	46.71
	DeepLabV3+ [16]	82.30±1.05	88.42±1.20	73.70±1.45	76.15±0.57	86.62±1.66	89.85±1.08	81.44±1.17	13.23	5.81
	MACUNet [79]	81.21±1.21	87.38±1.81	72.81±2.02	74.39±2.14	86.03±1.30	88.93±1.77	80.36±1.85	16.80	**5.15**
	A2FPN [80]	82.65±1.55	90.32±0.42	74.21±0.81	75.73±1.14	87.25±0.69	90.56±0.55	82.03±0.78	20.92	22.82
	MAResU-Net [51]	82.34±1.30	87.93±2.18	74.15±0.33	76.48±0.51	87.53±1.31	90.21±0.67	81.69±0.94	17.55	26.28
	MANet [81]	81.34±2.27	88.45±1.04	73.74±1.16	75.79±0.71	86.28±0.95	89.53±1.37	81.12±1.32	38.90	35.86
	CMTFNet [82]	83.30±0.43	90.08±0.23	74.21±0.77	75.81±1.26	86.63±1.51	90.45±0.49	82.01±0.72	17.14	30.07
	DC-Swin [52]	82.96±0.83	88.25±1.75	74.26±0.38	75.34±1.21	86.12±2.13	89.73±1.10	81.39±1.26	25.29	45.61
	PyramidMamba [54]	83.03±0.73	86.92±1.14	72.58±1.53	64.59±1.73	85.26±0.75	86.93±1.33	78.48±1.47	59.14	109.99
	RS3Mamba [53]	82.36±0.77	89.17±0.53	73.58±1.32	75.58±0.73	87.67±0.80	90.11±0.55	81.67±0.73	31.65	49.66
Spatial-Frequency	SFUNet [38]	82.26±1.20	88.50±0.53	74.23±0.82	75.85±1.15	88.02±0.93	90.12±0.96	81.77±0.87	151.61	28.84
	MFMSNet [83]	79.38±0.43	88.26±0.83	74.42±0.21	73.56±1.03	85.10±1.19	88.65±0.43	80.14±0.39	46.79	58.06
	TexLiverNet [84]	81.87±2.13	87.65±1.55	73.91±1.19	74.76±1.11	86.58±0.58	89.52±1.28	80.95±1.49	77.95	40.07
	MEW-UNet [85]	83.11±0.75	88.28±0.86	71.73±0.79	72.27±1.28	86.30±1.38	88.90±1.21	80.34±1.37	81.05	138.89
	MIFNet [36]	82.57±0.78	89.18±1.40	73.71±1.02	75.63±0.88	87.53±0.63	89.99±0.46	81.72±0.58	17.71	25.35
	SFFNet [31]	82.83±1.13	87.71±2.16	73.96±0.77	74.55±1.37	87.70±1.62	89.58±1.19	81.35±1.31	25.84	34.16
	WaveMix [86]	83.05±0.22	87.55±1.43	72.53±1.32	73.79±0.63	87.60±0.81	89.41±0.50	80.90±0.52	37.52	38.85
	AFENet [87]	82.35±1.20	90.11±0.21	73.55±1.10	75.23±1.06	87.46±0.83	90.10±0.74	81.74±0.98	12.80	20.23
	SF-Net(Ours)	**84.61±0.82**	**91.22±0.45**	**75.85±0.63**	**77.36±1.37**	**89.93±1.55**	**92.03±0.65**	**83.79±0.57**	**9.48**	23.01

CGFM module's ability to preserve details. SF-Net consistently produces clean and well-delineated predictions, while other methods struggle to segment densely packed cars, as shown in Figure 8.

LoveDA Dataset: As shown in Table II, SF-Net achieved the best performance, with an mIoU of 53.34% and an mF1 of 69.28%, outperforming all other methods across nearly every class in this more diverse and challenging dataset. SF-Net achieved an mIoU improvement of up to 4.66% as ompared to purely spatial-domain models. Moreover, SF-Net has the ability to resolve semantic ambiguity along complex boundaries, particularly between visually similar classes such as 'Water' and 'Forest', as illustrated in Figure 9. GSFEM and S-FFM work together to augment feature representations with global contextual cues and fine-grained frequency information, resulting in improved discriminative capabilities.

ISPRS Potsdam Dataset: Table III shows that SF-Net has the highest mIoU of 83.79%. It consistently outperformed spatial and frequency-based techniques, with mIoU and mF1 increases up to 5.59% and 5.1%, respectively. Moreover, the results presented in Fig. 10 show the credibility of proposed SF-Net to maintain the geometric integrity of large structures like 'Buildings'. This also produces sharp, regular boundaries with adjacent classes. The model is able to effectively captures both fine features and macro-structural background.

Complexity Analysis: Considering computational complexity, SF-Net is lightweight because it uses only 9.48G FLOPs and 23.01M parameters. SF-Net outperforms the existing models, such as PSPNet, DC-Swin, and SF-Unet in terms of efficiency. This model produces high accuracy while having a low computational footprint. This ability of SF-Net making it one of the most efficient methods. Since SF-Net effectively

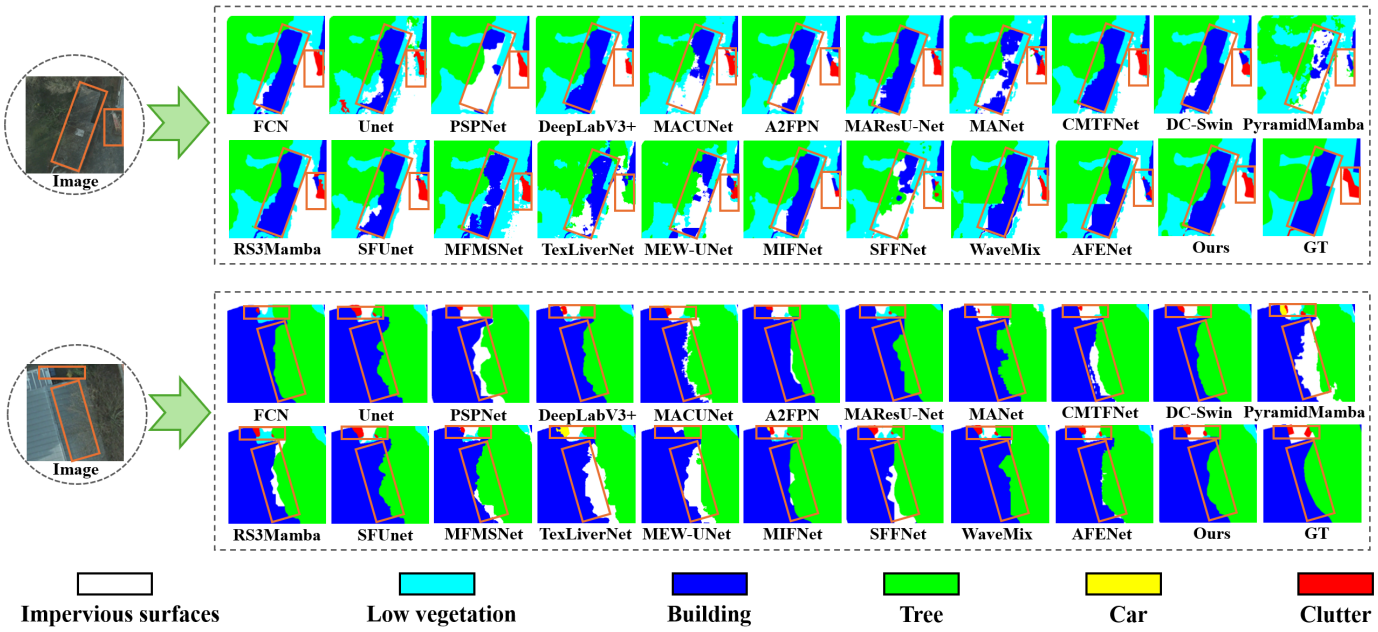

Fig. 10. Segmentation examples of different methods on the ISPRS Potsdam dataset.

balances performance and efficiency, which makes it a feasible solution for real-world applications.

D. Significance analysis

All experimental results are presented as mean ± standard deviation. Statistical significance was analyzed by paired t-tests on five independent replicates, Among them, $p<0.05$ indicates a significant difference compared with the comparative methods, $p<0.01$ indicates a highly significant difference, and $p<0.001$ indicates an extremely significant difference, as shown in Table IV. SF-Net achieved the best mIoU performance on the ISPRS Vaihingen, LoveDA and ISPRS Potsdam datasets, with an extremely significant improvement at $p<0.001$ over all spatial and spatial-frequency comparative methods. This fully verifies the effectiveness and superiority of the designed spatial-frequency fusion strategy, and the performance gain is not due to random experimental factors but has reliable statistical significance.

E. Ablation Studies

We performed a series of rigorous ablation studies on the ISPRS Vaihingen dataset to evaluate the role of each core component in SF-Net.

1) Impact of Core Modules: To show the impact of core modules, the proposed work began with a basic encoder-decoder architecture and gradually added the CGFM, GSFEM, and S-FFM modules. The benchmarks model obtained a mIoU of 80.39%, as shown in Table V. Introducing individual modules resulted in significant performance improvements, showing their independent effectiveness. Integrating different modules produced additional benefits due to their complementary characteristics. The complete SFNet model with all three components exhibits the greatest mIoU of 83.68%, which shows an absolute improvement of 3.29% over the benchmarks. These results support the synergistic integration of CGFM, GSFEM, and S-FFM for better segmentation performance.

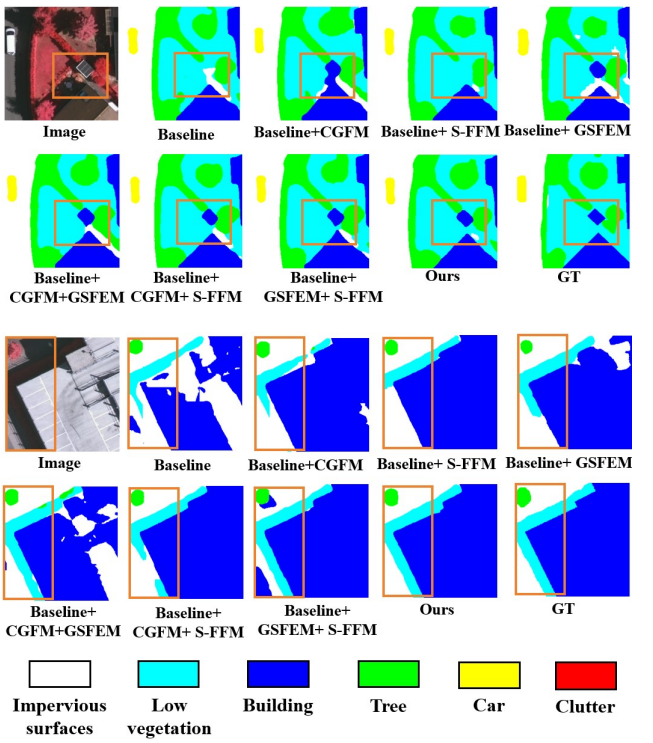

Fig. 11. Segmentation examples of different methods on the ISPRS Vaihingen dataset.

2) Analysis of the CGFM: To assess the efficiency of the CGFM, we substituted it with a number of well utilized multiscale fusion modules. The results, presented in Table VI, demonstrate that the CGFM-equipped model greatly outperformed all other models, obtaining an mIoU of 83.68%, a

TABLE IV
RESULTS OF PAIRED T-TEST

Dataset	Method	mF1	p-Value(mF1)	mIoU	p-Value(mIoU)
ISPRS Vaihingen	FCN	88.53±1.31	< 0.001	80.35±1.28	< 0.001
	Unet	88.42±1.77	0.00287	80.18±1.86	0.00359
	PSPNet	89.45±0.41	< 0.001	81.25±0.38	< 0.001
	DeepLabV3+	89.07±1.15	0.00421	80.94±1.36	0.00518
	MACUNet	88.39±0.54	< 0.001	79.32±0.66	< 0.001
	A2FPN	90.01±0.49	< 0.001	81.86±0.32	< 0.001
	MAResU-Net	89.95±0.74	0.00634	81.59±0.81	< 0.001
	MANet	88.82±0.88	< 0.001	80.73±0.72	< 0.001
	CMTFNet	89.72±0.72	< 0.001	81.44±0.53	< 0.001
	DC-Swin	86.56±0.57	< 0.001	77.72±0.45	< 0.001
	PyramidMamba	89.78±0.76	0.00729	81.42±0.73	< 0.001
	RS3Mamba	89.18±0.92	< 0.001	80.98±0.85	< 0.001
	SFUNet	87.53±0.78	< 0.001	79.84±0.71	< 0.001
	MFMSNet	89.62±0.55	< 0.001	81.14±0.73	< 0.001
	TexLiverNet	86.55±0.72	< 0.001	78.23±0.83	< 0.001
	MEW-UNet	88.05±0.39	< 0.001	79.87±0.40	< 0.001
	MIFNet	89.07±1.37	0.00815	80.88±1.32	0.00436
	SFFNet	89.12±0.52	< 0.001	80.84±0.65	< 0.001
	WaveMix	89.07±1.28	0.00362	80.72±1.42	0.00609
	AFENet	88.75±1.16	< 0.001	80.41±1.07	< 0.001
	SF-Net(Ours)	91.57±0.59		83.68±0.67	
LoveDA	FCN	66.01±1.41	< 0.001	49.74±1.35	< 0.001
	Unet	65.20±1.04	< 0.001	49.02±1.16	< 0.001
	PSPNet	67.69±0.39	< 0.001	51.43±0.41	< 0.001
	DeepLabV3+	63.96±1.22	< 0.001	48.74±1.37	< 0.001
	MACUNet	63.50±0.94	< 0.001	47.49±0.77	< 0.001
	A2FPN	67.88±0.69	0.00257	51.69±0.56	< 0.001
	MAResU-Net	66.37±1.27	< 0.001	50.39±0.89	< 0.001
	MANet	66.82±1.35	0.00519	50.63±1.26	0.00472
	CMTFNet	66.35±0.80	< 0.001	50.20±0.69	< 0.001
	DC-Swin	66.68±1.02	< 0.001	50.58±0.95	< 0.001
	PyramidMamba	65.26±0.94	< 0.001	49.48±1.10	< 0.001
	RS3Mamba	67.83±0.82	0.00318	51.58±0.73	< 0.001
	SFUNet	65.78±0.92	< 0.001	49.96±1.07	< 0.001
	MFMSNet	64.39±1.38	< 0.001	48.68±0.65	< 0.001
	TexLiverNet	66.57±0.51	< 0.001	50.44±0.41	< 0.001
	MEW-UNet	67.18±0.65	< 0.001	51.06±0.42	< 0.001
	MIFNet	67.69±0.81	0.00487	51.52±0.93	0.00503
	SFFNet	66.78±1.26	0.00659	50.58±1.42	0.00381
	WaveMix	67.53±0.57	< 0.001	51.41±0.63	< 0.001
	AFENet	66.36±0.76	< 0.001	50.24±0.81	< 0.001
	SF-Net(Ours)	69.28±0.39		53.34±0.47	
ISPRS Potsdam	FCN	88.51±1.41	< 0.001	80.08±1.23	< 0.001
	Unet	87.12±1.98	< 0.001	78.20±2.10	< 0.001
	PSPNet	89.58±1.17	0.00243	81.34±0.86	< 0.001
	DeepLabV3+	89.85±1.08	0.00379	81.44±1.17	0.00625
	MACUNet	88.93±1.77	0.00508	80.36±1.85	0.00462
	A2FPN	90.56±0.55	0.00714	82.03±0.78	0.00883
	MAResU-Net	90.21±0.67	< 0.001	81.69±0.94	0.00547
	MANet	89.53±1.37	0.00426	81.12±1.32	0.00318
	CMTFNet	90.45±0.49	< 0.001	82.01±0.72	< 0.001
	DC-Swin	89.73±1.10	0.00675	81.39±1.26	0.00732
	PyramidMamba	86.93±1.33	< 0.001	78.48±1.47	< 0.001
	RS3Mamba	90.11±0.55	< 0.001	81.67±0.73	< 0.001
	SFUNet	90.12±0.96	0.00285	81.77±0.87	0.00409
	MFMSNet	88.65±0.43	< 0.001	80.14±0.39	< 0.001
	TexLiverNet	89.52±1.28	0.00347	80.95±1.49	0.00581
	MEW-UNet	88.90±1.21	< 0.001	80.34±1.37	< 0.001
	MIFNet	89.99±0.46	< 0.001	81.72±0.58	< 0.001
	SFFNet	89.58±1.19	0.00472	81.35±1.31	0.00615
	WaveMix	89.41±0.50	< 0.001	80.90±0.52	< 0.001
	AFENet	90.10±0.74	< 0.001	81.74±0.98	0.00364
	SF-Net(Ours)	92.03±0.65		83.79±0.57	

TABLE V
RESULTS OF SF-NET WITH DIFFERENT COMPONENTS.

CGFM	GSFEM	S-FFM	mF1	mIoU
			88.61±0.43	80.39±0.25
✓			89.43±0.37	81.45±0.14
	✓		89.69±0.27	81.51±0.40
		✓	89.61±0.51	81.47±0.39
✓	✓		90.11±0.65	81.90±0.73
✓		✓	90.73±0.38	82.34±0.62
	✓	✓	90.82±0.47	82.51±0.50
✓	✓	✓	**91.57±0.59**	**83.68±0.67**

TABLE VI
RESULTS OF DDFNET WITH DIFFERENT MULTI-SCALE FUSION MODELS.

Modules	mF1	mIoU
FPN [62]	88.89±1.13	80.32±1.21
TIF [88]	89.30±0.78	80.63±0.92
AFF [63]	89.13±0.63	81.25±0.72
IAFF [63]	89.32±0.28	80.75±0.49
MSCAM [63]	89.52±0.84	80.83±1.03
M2SNet [89]	89.47±0.61	81.53±0.55
TFAM [90]	89.39±0.41	81.49±0.29
M2SA [82]	89.90±0.39	82.09±0.50
MSAA [91]	89.57±0.81	81.67±0.75
MSAF [64]	90.22±0.57	82.31±0.61
CGFM (Ours)	**91.57±0.59**	**83.68±0.67**

TABLE VII
RESULTS OF DDFNET WITH DIFFERENT GLOBAL CONTEXT MODULES.

Methods	mF1	mIoU
GF [65]	89.88±0.47	81.69±0.60
GB [31]	88.75±1.12	80.93±1.26
LEGM [92]	90.10±0.39	82.16±0.42
LWGA [93]	89.61±0.71	81.35±0.53
GSFEM (Ours)	**91.57±0.59**	**83.68±0.67**

1.37%, higher than the second-best module, MSAF. These results indicate the effectiveness of our progressive, channel-grouped fusion strategy in integrating multi-scale features while preserving crucial spatial details. CGFM allows for progressive feature interaction across scales, reducing detail loss compared to direct or abrupt fusion procedures.

3) Analysis of the GSFEM: Table VII shows the comparative results of GSFEM against several established global context modules. The results indicate that GSFEM obtained the highest mF1 and mIoU scores as 91.57% and 83.68%, respectively. The reason of this increase performance is the dual design of GSFEM. The Mamba SSM does a linear-time scan at full resolution, establishing direct long-range dependencies throughout the spatial domain while keeping comprehensive local information. In addition, the multi-pooling attention method dynamically re-weights features from various geometric perspectives, effectively emphasizing semantically significant regions. This combination of global semantic modeling and localized attention enables GSFEM to provide both high-level consistency and structural precision, resulting in the large performance benefits seen. As shown in Table VIII, we further assessed the internal structure of GSFEM by isolating its individual components. A model that simply used the Feed-Forward Mamba (F-FM) block scored a mIoU of 81.95%. Moreover, using either Global Average Pooling (GAP) or Global Max Pooling (GMP) resulted in additional improvements. Nevertheless, The whole GSFEM configuration, including F-FM, GAP, and GMP, achieved the greatest mIoU of 83.68%. The results confirm our hypothesis that the three components are complementary: the Mamba SSM shows long-range dependencies, while GAP provides a global statistical summary, and GMP emphasizes salient local features. Their combined effect produces a more comprehensive and robust global representation.

4) Analysis of the S-FFM: We replaced the S-FFM with several advanced fusion modules to validate the effectiveness of our spatial-frequency fusion strategy. As shown in Table IX, the model equipped with S-FFM achieved the highest mIoU of

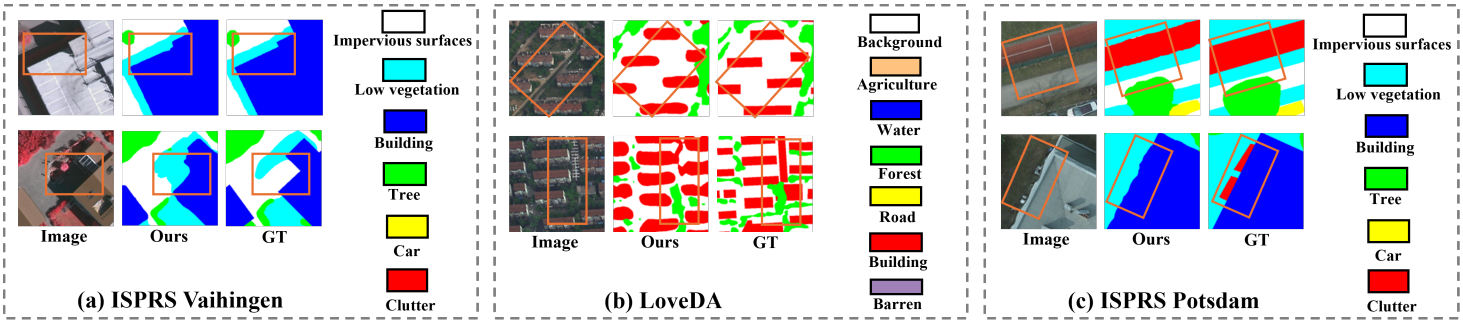

Fig. 12. Failure segmentation results of SF-Net in different complex scenarios across three datasets.

TABLE VIII
RESULTS OF SF-NET WITH DIFFERENT GLOBAL POOLING BRANCH COMBINATIONS.

Methods	mF1	mIoU
F-FM	89.82±0.63	81.95±0.52
F-FM+AvgPool	90.41±0.33	82.57±0.29
F-FM+MaxPool	90.32±0.40	82.43±0.37
F-FM+AvgPool+MaxPool	**91.57±0.59**	**83.68±0.67**

83.68%, significantly outperforming all alternatives. It significantly outperformed HFFM, a recent frequency-aware module, by 1.47% in mIoU. This demonstrates S-FFM's key advantage: instead of just combining features, it performs a principled, differential fusion based on DWT. S-FFM allows for more focused feature integration, which improves segmentation efficiency by explicitly separating structural (low-frequency) and textural (high-frequency) information.

TABLE IX
RESULTS OF DDFNET WITH DIFFERENT FUSION MODULES.

Methods	mF1	mIoU
CCM [53]	89.61±0.57	81.56±0.44
FFM [66]	88.75±1.31	80.76±1.28
EFFM [94]	90.03±0.31	82.03±0.51
FAFF [67]	89.44±0.93	81.49±1.02
HFFM [68]	90.12±0.42	82.21±0.57
S-FFM (Ours)	**91.57±0.59**	**83.68±0.67**

V. DISCUSSION ON THE LIMITATIONS OF SF-NET

To intuitively reveal the performance limitations of SF-Net in complex and challenging scenarios, this paper conducts qualitative visual analysis of segmentation results on three datasets for three typical challenging scenarios, namely shadow regions, dense target regions and overlapping regions, with the results presented in Fig. 12. For each dataset, the first visualization result demonstrates the model's segmentation performance in conventionally complex scenarios, while the second one corresponds to that in extremely complex scenarios. For shadow regions, SF-Net can still maintain favorable segmentation performance in scenarios where the contrast and brightness differences between shadow and non-shadow regions are significant (as illustrated in the first visualization result of the (a) ISPRS Vaihingen dataset). However, in scenarios with slight contrast and brightness differences between the two regions, the model's segmentation performance degrades markedly, accompanied by the misclassification of target pixels. Consistently, SF-Net exhibits a certain degree of robustness in segmenting dense target regions and overlapping regions under conventionally complex scenarios, whereas its performance deteriorates in extreme scenarios featuring extremely distributed dense targets and severely overlapping regions. Specifically, the model suffers from typical issues such as target missing and boundary adhesion in dense target regions, and presents failures in pixel attribution discrimination for overlapping regions. In summary, the visualization results clearly verify that SF-Net has good adaptability to conventionally complex scenarios involving shadow, dense target and overlapping regions, yet it still has non-negligible performance bottlenecks in the extremely challenging scenarios of the above three typical problematic regions.

VI. CONCLUSION

This paper presents SF-Net, a coherently designed deep learning architecture that deeply fuses spatial and frequency domain information for remote sensing image semantic segmentation. It consists of three core modules: CGFM for refined multi-scale feature extraction, GSFEM for efficient and comprehensive global context modeling, and S-FFM for systematic cross-domain feature fusion. Extensive experiments on three public benchmark datasets demonstrate that SF-Net consistently outperforms state-of-the-art methods, setting new benchmarks for computational efficiency and segmentation accuracy. It should be noted that SF-Net currently adopts fixed Haar wavelets for frequency domain processing. This design not only limits the model's adaptability to heterogeneous scenarios but also leads to its failure in scenes with complex frequency distributions, as the fixed wavelet basis cannot match diverse textural and frequency features, causing segmentation blurring. In addition, the model has performance limitations in shadowed areas and dense small target areas. Future work will explore adaptable and learnable time-frequency analysis techniques (focusing on spectral decomposition and adaptive wavelet transform), and design targeted evaluation protocols to verify the performance improvement of the improved methods in failure-prone and performance-limited scenarios.

REFERENCES

[1] H. Wu, J. Xie, W. Deng, A. Lin, A. R. M. Shariff, S. Akmalov, W. Wu, Z. Li, Q. Yu, Q. Wang, and J. Zhang, "CT-HiffNet: A contour-texture hierarchical feature fusion network for cropland field parcel extraction from high-resolution remote sensing images," Comput. Electron. Agric., vol. 239, p. 111010, 2025.

[2] L. Li, L. Liu, D. Huang, S. Wang, X. Wang, Y. He, and Z. Zhong, "KECS-Net: Knowledge-embedded CSwin-UNet with slicing-aided hypersegmentation for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 23, pp. 1–5, 2025.

[3] L. Li, L. Liu, F. Cheng, Y. He and Z. Zhong, "CN-UNet: ConvNeXt UNet With Slicing-Aided Hyper Segmentation for Infrared Small Target Detection," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 19, pp. 84-98, 2026,

[4] D. Hong, B. Zhang, and X. Li, "SpectralGPT: Spectral Remote Sensing Foundation Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 5227–5244, 2024.

[5] G. Gao, M. Wang, P. Zhou, L. Yao, X. Zhang, and H. Li, "A Multibranch Embedding Network With Bi-Classifier for Few-Shot Ship Classification of SAR Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2022.

[6] G. Zhou et al., "Adaptive Adjustment for Laser Energy and PMT Gain Through Self-Feedback of Echo Data in Bathymetric LiDAR," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-22, Art no. 4206422, 2024.

[7] J. Ding, C. Xu, W. Chen, J. Chen, J. Wang, Y. Zhang, L. Bai, T. Han, and Y. Xiong, "Impact of VAEformer Compression Algorithm Precision Loss on the Tropospheric Delays for Microwave Remote Sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–11, 2025.

[8] J. Jia, Y. Wang, X. Zheng, L. Yuan, C. Li, Y. Cen, F. Si, G. Lv, C. Wang, S. Wang, and C. Zhang, "Design, performance, and applications of AMMIS: A novel airborne multimodular imaging spectrometer for high-resolution Earth observations," *Engineering*, vol. 47, pp. 38–56, 2025.

[9] Y. Qiu, X. M. Li, L. Yan, and Z. Chen, "Synergic sensing of light and heat emitted by offshore oil and gas platforms in the South China Sea," *International Journal of Digital Earth*, vol. 17, no. 1, pp. 2441932, 2024.

[10] H. Liao, J. Xia, Z. Yang, F. Pan, Z. Liu and Y. Liu, "Meta-Learning Based Domain Prior With Application to Optical-ISAR Image Translation," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7041-7056, Aug. 2024.

[11] Z. Wang, Z. Luo, Q. Zhu, S. Peng, L. Ran, and Y. Zhang, "A Road-Detail Preserving Framework for Urban Road Extraction From VHR Remote Sensing Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025.

[12] Y. LeCun, L. Bottou, and Y. Bengio, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[15] H. Zhao, J. Shi, and X. Qi, "Pyramid Scene Parsing Network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

[16] L.-C. Chen, Y. Zhu, and G. Papandreou, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 833–851.

[17] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6000–6010.

[18] L. Yin, L. Wang, S. Lu, R. Wang, Y. Yang, B. Yang, S. Liu, A. AlSanad, S. A. AlQahtani, Z. Yin, and X. Li, "Convolution-transformer for image feature extraction," CMES-Comput. Model. Eng. Sci., vol. 141, no. 1, 2024.

[19] Y. Zhang et al., "S2DBFT: Spectral–Spatial Dual-Branch Fusion Transformer for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1-17, Art no. 5525517, 2025.

[20] Z. Guan et al., "Mixed Layer Depth Estimation From Multisource Remote Sensing Data Using Clustering-Machine Learning Method," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 11183-11197, 2025.

[21] Z. Liang, S. Bao, W. Zhang, H. Yan, B. Duan and H. Wang, "Super-Resolution Reconstruction of SMOS Sea Surface Salinity from Multivariate Satellite Observations Based on Deep Learning," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 24251-24266, 2025.

[22] Z. Gao, H. Zeng, X. Zhang, H. Wu, R. Zhang, Y. Sun, Q. Du, Z. Zhao, Z. Li, F. Zhao, and L. Liu, "Exploring tourist spatiotemporal behavior differences and tourism infrastructure supply–demand pattern fusing social media and nighttime light remote sensing data," *Int. J. Digit. Earth*, vol. 17, no. 1, pp. 2310723, 2024.

[23] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arXiv preprint arXiv:2312.00752*, 2024.

[24] Y. Zhu, J. Shu, W. Ding, C. Yue, Y. Lu, and J. Zhang, "A crack detection and quantification framework for high-resolution images using Mamba and unmanned devices," in *Computer-Aided Civil and Infrastructure Engineering*, vol. 40, no. 29, pp. 5672–5697, 2025.

[25] D. Guo, Z. Zhang, J. Liu, J. Zhang, and Y. Lin, "Multi-horizon flight trajectory prediction enabled by time-frequency wavelet transform," Nature Communications, 2025.

[26] H. Zhu, X. Xu, X. Zou, M. Tan, W. Song and L. Han, "Method for Estimating the Key Parameters of Phase-Coded Signals With Low Sampling Rate Based on Wavelet Analysis and Periodic Annihilating Filtering," in *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1-18, no. 6511118, 2025.

[27] N. Wang, Q. Wu, Y. Gui, Q. Hu, and W. Li, "Cross-modal segmentation network for winter wheat mapping in complex terrain using remote-sensing multi-temporal images and DEM data," *Remote Sensing*, vol. 16, no. 10, p. 1775, 2024.

[28] D. Yang, H. Sheng, S. Wang, S. Wang, Z. Xiong and W. Ke, "Boosting Light Field Spatial Super-Resolution via Masked Light Field Modeling," in *IEEE Transactions on Computational Imaging*, vol. 10, pp. 1317-1330, 2024.

[29] B. Tu, T. Zhou, B. Liu, Y. He, J. Li and A. Plaza, "Multi-Scale Autoencoder Suppression Strategy for Hyperspectral Image Anomaly Detection," in *IEEE Transactions on Image Processing*, vol. 34, pp. 5115-5130, 2025.

[30] Z. Shen, Y. He, X. Du, J. Yu, H. Wang and Y. Wang, "YCANet: Target Detection for Complex Traffic Scenes Based on Camera-LiDAR Fusion," in *IEEE Sensors Journal*, vol. 24, no. 6, pp. 8379-8389, 15 March15, 2024.

[31] Y. Yang, G. Yuan, and J. Li, "SFFNet: A Wavelet-Based Spatial and Frequency Domain Fusion Network for Remote Sensing Segmentation," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[32] W. Luo, H. Qiu, Y. Wei, W. Huangfu, and D. Yang, "A proposed method for landslide detection based on transfer learning and graph neural network," in *Geoscience Frontiers*, pp. 102171, 2025.

[33] G. Zhou et al., "Adaptive high-speed echo data acquisition method for bathymetric LiDAR," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, no. 5703017, 2024.

[34] G. Zhou et al., "Internal stray light suppression for single-band bathymetric LiDAR optical system," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1–12, no. 7007112, 2024.

[35] R. Zhang, Y. Wang, Z. Li, F. Ding, C. Wei and M. Wu, "Online Adaptive Keypoint Extraction for Visual Odometry Across Different Scenes," in *IEEE Robotics and Automation Letters*, vol. 10, no. 7, pp. 7539-7546, July 2025.

[36] J. Fan, J. Li, and Y. Liu, "Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 129, 2024, no. 107638.

[37] H. Cao, D. Chen, Y. Zhang, H. Zhou, D. Wen, and C. Cao, "MFINet: A multi-scale feature interaction network for point cloud registration," in *The Visual Computer*, vol. 41, no. 6, pp. 4067–4079, 2025.

[38] Z. Zhou, A. He, and Y. Wu, "Spatial-Frequency Dual Domain Attention Network for Medical Image Segmentation," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2024, pp. 4076–4081.

[39] B. Tu, Q. Ren, J. Li, Z. Cao, Y. Chen, and A. Plaza, "NCGLF2: Network combining global and local features for fusion of multisource remote sensing data," *Information Fusion*, vol. 104, p. 102192, 2024.

[40] J. Song, X. Zhou, C. Xu, S. Xu, and J. Shu, "Training-free automatic instance segmentation of girder bridge point cloud via large model fusion with reverse entity modelling verification," *Automation in Construction*, vol. 179, p. 106484, 2025.

[41] H. Sun, Z. Zhang, Z. Li, and J. Liu, "Large-capacity and robust video watermarking via DWT coefficient separation/reconstruction and multi-scale spatiotemporal fusion," *Neurocomputing*, vol. 654, p. 131275, 2025.

[42] K. Cheng, S. Liu, and M. Teng, "Towards the equalization of basic public services in Chinese cities: Investigating the role of urban polycentric structure," *Applied Spatial Analysis and Policy*, vol. 18, no. 4, p. 143, 2025.

[43] L. Yu, Y. Liu, T. Liu, and F. Yan, "Impact of recent vegetation greening on temperature and precipitation over China," in *Agricultural and Forest Meteorology*, vol. 295, p. 108197, 2020.

[44] H. Wu, Y. Li, A. Lin, H. Fan, K. Fan, J. Xie, and W. Luo, "A review of crowdsourced geographic information for land-use and land-cover mapping: Current progress and challenges," *International Journal of Geographical Information Science*, vol. 38, no. 11, pp. 2183–2215, 2024.

[45] X. Li, F. Xu, F. Liu, X. Lyu, H. Gao, J. Zhou, and A. Kaup, "A Euclidean Affinity-Augmented Hyperbolic Neural Network for Semantic Segmentation of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–18, 2025.

[46] D. Su, S. Li, Y. Zhou, L. Gao, M. Jiang, X. Sun, H. Li, and E. Hou, "Dilated Transformation-Guided Unsupervised Multimodal Learning for Hyperspectral and Multispectral Image Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[47] J. Li, K. Zheng, Z. Li, L. Gao, and X. Jia, "X-Shaped Interactive Autoencoders With Cross-Modality Mutual Learning for Unsupervised Hyperspectral Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[48] R. Li, Y. Wang, S. Sun, Y. Zhang, F. Ding and H. Gao, "UE-Extractor: A Grid-to-Point Ground Extraction Framework for Unstructured Environments Using Adaptive Grid Projection," in *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5991-5998, June 2025.

[49] D. Zhou, M. Sheng, C. Bao, Y. Wang, J. Li and Z. Han, "Mission-Driven Resource Scheduling in Satellite-Terrestrial Networks: From Perspective of Collaboration and Reconfiguration," in *IEEE Transactions on Communications*, vol. 73, no. 8, pp. 6705-6719, Aug. 2025.

[50] S. Peng, N. Bao, N. Gu, H. Qian, Z. Han, B. Zhou, and L. Chang, "Enhanced assessment of chlorophyll-a and total nitrogen dynamics using unmanned aerial vehicle-based model and hyperspectral imagery in coastal wetland water," Environmental Technology and Innovation, pp. 104521, 2025.

[51] R. Li, S. Zheng, and C. Duan, "Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[52] L. Wang, R. Li, and C. Duan, "A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, 2022, pp. 1–1.

[53] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[54] L. Wang, D. Li, S. Dong, X. Meng, X. Zhang, and D. Hong, "PyramidMamba: Rethinking Pyramid Feature Fusion with Selective Space State Model for Semantic Segmentation of Remote Sensing Imagery," *arXiv preprint arXiv:2406.10828*, 2024.

[55] X. Li, F. Xu, A. Yu, X. Lyu, H. Gao, and J. Zhou, "A Frequency Decoupling Network for Semantic Segmentation of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–21, 2025.

[56] X. Li, F. Xu, J. Zhang, A. Yu, X. Lyu, H. Gao, and J. Zhou, "Dual-domain decoupled fusion network for semantic segmentation of remote sensing images," *Information Fusion*, vol. 124, p. 103359, 2025.

[57] X. Li, F. Xu, J. Zhang, H. Zhang, X. Lyu, F. Liu, H. Gao, and A. Kaup, "Frequency-Guided Denoising Network for Semantic Segmentation of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, p. 1, 2025.

[58] X. Li, F. Xu, F. Liu, Y. Tong, X. Lyu, and J. Zhou, "Semantic Segmentation of Remote Sensing Images by Interactive Representation Refinement and Geometric Prior-Guided Inference," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.

[59] Z. Yang, J. Xia, S. Zhang, L. Liu, Y. Fu and Y. Liu, "A Learning-Aided Unsupervised Method for Sparse Aperture ISAR Imaging and Autofocusing," in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 62, pp. 350-367, 2026.

[60] S. Qian, Y. Chen, W. Wang, G. Zhang, L. Li, Z. Hao, and Y. Wang, "Physics-guided deep neural networks for bathymetric mapping using Sentinel-2 multi-spectral imagery," in *Frontiers in Marine Science*, vol. 12, pp. 1636124, 2025.

[61] J. He, Y. Wang, L. Deng, Y. Li, and Z. L. Li, "Phase congruency image mosaicking approach for aerial mid-wave infrared low-overlap array scanning images," in *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 227, pp. 185–204, 2025.

[62] M. Zhu, K. Han, and C. Yu, "Dynamic Feature Pyramid Networks for Object Detection," *arXiv preprint arXiv:2012.00779*, 2021.

[63] Y. Dai, F. Gieseke, and S. Oehmcke, "Attentional Feature Fusion," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2021, pp. 3559–3568.

[64] Z. Guo, L. Bian, and H. Wei, "DSNet: A Novel Way to Use Atrous Convolutions in Semantic Segmentation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 35, pp. 3679–3692, 2024.

[65] Y. Rao, W. Zhao, and Z. Zhu, "Global filter networks for image classification," in *Proc. 35th Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2021, pp. 76:1–76:14.

[66] J. Zhang, Z. Zeng, and P. K. Sharma, "A dual encoder crack segmentation network with Haar wavelet-based high–low frequency attention," *Expert Systems with Applications*, vol. 256, 2024, no. 124950.

[67] L. Chen, Y. Fu, and L. Gu, "Frequency-Aware Feature Fusion for Dense Image Prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 10763–10780, 2024.

[68] X. Huo, G. Sun, and S. Tian, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomedical Signal Processing and Control*, vol. 87, 2024, no. 105534.

[69] K. He, X. Zhang, and S. Ren, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[70] L. Wang, R. Li, and C. Zhang, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[71] W. Liu, X. Shi, W. Yan, S. Gou, D. J. Parker, and Z. Xu, "Spatiotemporal patterns and propagation characteristics of convective activity on the northeast slope of Tibetan Plateau: A high-resolution radar perspective," *Atmos. Res.*, pp. 108609, 2025.

[72] Q. An, Y. Dong, W. Dong, and S. Xiao, "Spatiotemporal dynamics and nonlinear landscape-driven mechanisms of urban heat islands in a winter city: A case study of Harbin, China," *Sustainable Cities and Society*, pp. 106842, 2025.

[73] X. Zhou, Z. Zhao, J. Shen, Z. Liu, Y. Liu, and B. Xue, "Sparse aperture ISAR autofocusing and imaging algorithm based on log-sum regularization," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, Art. no. 5103614, 2025.

[74] Z. Zhu, R. Lu, B. Yu, T. Li, and S. W. Yeh, "A moderator of tropical impacts on climate in Canadian Arctic Archipelago during boreal summer," in *Nature Communications*, vol. 15, no. 1, p. 8644, 2024.

[75] Y. Liu, Y. Tian, and Y. Zhao, "VMamba: Visual State Space Model," *arXiv preprint arXiv:2401.10166*, 2024. [Online]. Available: https://arxiv.org/abs/2401.10166

[76] ISPRS Vaihingen, "ISPRS Vaihingen Dataset," 2010. [Online]. Available: https://paperswithcode.com/dataset/isprs-vaihingen

[77] J. Wang, Z. Zheng, and A. Ma, "LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation," in *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1, 2021.

[78] ISPRS Potsdam, "ISPRS Potsdam Dataset," 2010. [Online]. Available: https://paperswithcode.com/dataset/isprs-potsdam

[79] R. Li, C. Duan, and S. Zheng, "MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[80] M. Hu, Y. Li, and L. Fang, "A2-FPN: Attention Aggregation based Feature Pyramid Network for Instance Segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15338–15347.

[81] R. Li, S. Zheng, and C. Zhang, "Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[82] H. Wu, P. Huang, and M. Zhang, "CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[83] R. Wu, X. Lu, and Z. Yao, "MFMSNet: A Multi-frequency and Multi-scale Interactive CNN-Transformer Hybrid Network for Breast Ultrasound Image Segmentation," *Computers in Biology and Medicine*, vol. 177, 2024, no. 108616.

[84] X. Jiang, Z. Zhou, and H. Wang, "Texlivernet: Leveraging Medical Knowledge and Spatial-Frequency Perception for Enhanced Liver Tumor Segmentation," in *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI)*, 2025, pp. 1–5.

[85] J. Ruan, J. Gao, and M. Xie, "Learning multi-axis representation in frequency domain for medical image segmentation," *Machine Learning*, vol. 114, no. 10, 2025.

[86] P. J. P, and A. Sethi, "WaveMix: Resource-efficient Token Mixing for Images," *arXiv preprint arXiv:2203.03689*, 2022. [Online]. Available: https://arxiv.org/abs/2203.03689

[87] F. Gao, M. Fu, J. Cao, J. Dong, and Q. Du, "Adaptive Frequency Enhancement Network for Remote Sensing Image Semantic Segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[88] A. Lin, B. Chen, and J. Xu, "DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[89] X. Zhao, H. Jia, and Y. Pang, "M^2SNet: Multi-scale in Multi-scale Subtraction Network for Medical Image Segmentation," *arXiv preprint arXiv:2303.10894*, 2023.

[90] S. Zhao, X. Zhang, and P. Xiao, "Exchanging Dual-Encoder–Decoder: A New Strategy for Change Detection With Semantic Guidance and Spatial Localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[91] M. Cui, K. Li, and J. Chen, "CM-Unet: A Novel Remote Sensing Image Segmentation Method Based on Improved U-Net," *IEEE Access*, vol. 11, pp. 56994–57005, 2023.

[92] Y. Zhang, S. Zhou, and H. Li, "Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 2846–2855.

[93] W. Lu, S.-B. Chen, and C. H. Q. Ding, "LWGANet: A Lightweight Group Attention Backbone for Remote Sensing Visual Tasks," *arXiv preprint arXiv:2501.10040*, 2025.

[94] X. Li, X. Qin, and C. Huang, "SUnet: A multi-organ segmentation network based on multiple attention," *Computers in Biology and Medicine*, vol. 167, 2023, no. 107596.