remote sensing

MDPI

Article

A Frequency Attention-Enhanced Network for Semantic Segmentation of High-Resolution Remote Sensing Images

Jianyi Zhong [1,†], Tao Zeng [1,†], Zhennan Xu [1], Caifeng Wu [1], Shangtuo Qian [2], Nan Xu [3], Ziqi Chen [4], Xin Lyu [1,5] and Xin Li [1,5,*]

[1] College of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China; zhongjianyi@hhu.edu.cn (J.Z.); tzeng.nj@hhu.edu.cn (T.Z.); zhennanxu@hhu.edu.cn (Z.X.); caifengwu@hhu.edu.cn (C.W.); lvxin@hhu.edu.cn (X.L.)

[2] College of Agricultural Science and Engineering, Hohai University, Nanjing 211100, China; stqian@hhu.edu.cn

[3] College of Geography and Remote Sensing, Hohai University, Nanjing 211100, China; hhuxunan@hhu.edu.cn

[4] Department of Earth System Science, Tsinghua University, Beijing 100084, China; chenzq21@mails.tsinghua.edu.cn

[5] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 211100, China

[*] Correspondence: li-xin@hhu.edu.cn; Tel.: +86-187-6186-0051

[†] These authors contributed equally to this work.

Abstract: Semantic segmentation of high-resolution remote sensing images (HRRSIs) presents unique challenges due to the intricate spatial and spectral characteristics of these images. Traditional methods often prioritize spatial information while underutilizing the rich spectral context, leading to limited feature discrimination capabilities. To address these issues, we propose a novel frequency attention-enhanced network (FAENet), which incorporates a frequency attention model (FreqA) to jointly model spectral and spatial contexts. FreqA leverages discrete wavelet transformation (DWT) to decompose input images into distinct frequency components, followed by a two-stage attention mechanism comprising inner-component channel attention (ICCA) and cross-component channel attention (CCCA). These mechanisms enhance spectral representation, which is further refined through a self-attention (SA) module to capture long-range dependencies before transforming back into the spatial domain. FAENet's encoder–decoder architecture facilitates multiscale feature refinement, enabling effective segmentation. Extensive experiments on the ISPRS Potsdam and LoveDA benchmarks demonstrate that FAENet outperforms state-of-the-art models, achieving superior segmentation accuracy. Ablation studies further validate the contributions of ICCA and CCCA. Moreover, efficiency comparisons confirm the superiority of the proposed FAENet over other models.

Keywords: high-resolution remote sensing images; semantic segmentation; attention mechanism; convolutional neural network

check for updates

Academic Editor: Shuying Li

Received: 25 November 2024

Revised: 16 January 2025

Accepted: 21 January 2025

Published: 24 January 2025

Citation: Zhong, J.; Zeng, T.; Xu, Z.; Wu, C.; Qian, S.; Xu, N.; Chen, Z.; Lyu, X.; Li, X. A Frequency Attention-Enhanced Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sens.* **2025**, *17*, 402. https://doi.org/10.3390/rs17030402

1. Introduction

Semantic segmentation entails partitioning an image into regions belonging to the same class, assigning a label to each pixel [1–6]. Unlike image classification, which categorizes an entire image holistically, semantic segmentation provides granular class-level delineations. Stemming from image segmentation techniques, it focuses on embedding semantic context, thereby rendering segmented regions meaningful. This capability is essential for numerous computer vision tasks and finds application in diverse fields, including medical and natural image parsing [7–10], autonomous driving [11–14], road network extraction [15–18], and land-use classification [19–23].

Early approaches to semantic segmentation relied on manually designed low-level features, such as color, shape, and texture [24–26]. These features were processed using clustering or classification algorithms in high-dimensional spaces to perform segmentation. While these methods laid the foundation for early advancements, they exhibited clear limitations, including low flexibility and inadequate abstraction to handle complex segmentation tasks.

To overcome these constraints, convolutional neural networks (CNNs) revolutionized segmentation tasks by enabling hierarchical feature learning [27–29]. Among them, fully convolutional networks (FCNs) [30–35] established a new standard for semantic segmentation with their end-to-end trainable frameworks. However, due to their simplistic decoder structures, FCNs often produce coarse predictions, which led to the development of more advanced encoder–decoder architectures, such as UNet [31], SegNet [32], and UNet++ [33]. These models leverage hierarchical features to capture both local and global contexts, substantially improving segmentation precision, especially in medical imaging and natural scene analysis. For instance, UNet [31] adopts a symmetric encoder–decoder structure with skip connections to fuse low-level spatial features with high-level semantic information, becoming a cornerstone model in segmentation tasks. SegNet [32] enhances computational efficiency by reusing max-pooling indices, making it particularly appealing for real-time applications. UNet++ [33] introduces nested and dense skip connections to refine feature fusion, further boosting segmentation accuracy. Despite these advancements, the inherent challenges of remote sensing images (RSIs) remain unresolved due to their unique spectral and spatial complexities.

RSIs are distinct from natural images due to their intricate spatial structures and diverse spectral information. These images often exhibit high intraclass variability and low interclass separability, such as urban areas with overlapping spectral characteristics, which significantly complicates semantic segmentation. A key limitation of existing methods is their inability to consistently handle small objects and preserve fine boundaries in high-resolution remote sensing images (RSIs). For example, in complex urban environments, densely packed cars and intricate building edges are often misclassified or blurred by conventional models. This occurs because most CNN-based methods rely heavily on localized convolutional operations, which struggle to capture the high-frequency details necessary for accurate boundary delineation. Similarly, attention mechanisms and transformer-based models, while improving context modeling, still face challenges in maintaining a balance between local detail preservation and long-range dependencies. This inadequacy in handling small, detailed objects and accurately segmenting complex boundaries motivates the need for an approach that can effectively integrate spectral and spatial information.

Attention mechanisms (AMs) [36,37] have emerged as a transformative solution for capturing spatial and contextual dependencies in RSIs. By dynamically focusing on salient regions and features within an image, AMs enhance interpretability and segmentation accuracy, particularly in heterogeneous and complex environments. For instance, MACU-Net [38] integrates multiscale features and hierarchical connections, outperforming UNet in satellite image segmentation tasks. Similarly, A2FPN [39] incorporates an attention aggregation module to simultaneously enhance spatial and contextual understanding, demonstrating its effectiveness across diverse datasets. SAPNet [40], a more recent contribution, combines spatial and channel attention, achieving finer segmentation granularity. These advances highlight the potential of attention-driven frameworks to tackle RSIs' multiscale challenges.

Transformers have recently garnered significant attention in computer vision due to their ability to capture both local and global dependencies [41–43]. SETR [44] reformulates semantic segmentation as a sequence-to-sequence prediction problem by encoding images as patch sequences, achieving strong results on natural image benchmarks. Similarly, Segmenter [45], based on a vision transformer, employs a pre-trained image classification model and a simple linear decoder to produce segmentation masks. In the context of remote sensing images (RSIs), Blaga and Nedevschi [46] proposed a transformer-based U-Net with Guided Focal-Axial Attention, combining global and localized attention to improve the segmentation of high-resolution RSIs. Wang et al. [47] introduced UNetFormer, a hybrid transformer model that integrates local spatial details and global dependencies for efficient urban scene segmentation. Lin et al. [48] presented the Swin Transformer Segmentor, which integrates Swin Transformers with CNNs to refine boundary details in RSIs and improve the segmentation accuracy. Li et al. [49] proposed GPINet, which fuses CNN and transformer features with geometric priors to enhance segmentation performance. Similarly, He et al. [50] developed ST-UNet by combining Swin transformers and CNNs to leverage global context and spatial detail for better RSI segmentation. Lastly, Long et al. [51] introduced CLCFormer, a hybrid network that combines fine-grained spatial features and long-range global contexts using CNNs and transformers, achieving state-of-the-art results on high-resolution RSI datasets. However, there are two main issues to be addressed:

1. Existing methods often rely heavily on spatial features, while spectral richness in RSIs remains underexplored. This limitation reduces their capacity to distinguish subtle interclass differences, particularly in complex scenarios involving overlapping spectral features.
2. Capturing both local spatial details and long-range global dependencies is crucial for high-resolution RSIs. However, many existing models struggle to effectively balance these two aspects, limiting their segmentation accuracy in heterogeneous landscapes.

In recent years, frequency domain learning has gained traction for its ability to complement spatial domain approaches in image processing tasks [52,53]. Research has revealed that neural networks exhibit a spectral bias, naturally favoring low-frequency representations while struggling to capture high-frequency details critical for tasks such as edge preservation and fine-grained segmentation [54]. Inspired by these advancements, this paper proposes a novel frequency attention-enhanced network (FAENet) for semantic segmentation of high-resolution RSIs. The key contributions are summarized as follows:

1. We propose a frequency attention model (FreqA) that explicitly incorporates spectral and spatial contexts. Using discrete wavelet transformation (DWT), FreqA decomposes feature maps into frequency components. Inner-component channel attention (ICCA) and cross-component channel attention (CCCA) are designed to selectively emphasize informative spectral bands. These enhanced features are processed by a self-attention (SA) module, enabling joint modeling of spectral and spatial dependencies.
2. We design FAENet, an encoder–decoder architecture equipped with FreqA modules. This design enables hierarchical learning and multiscale feature refinement, allowing the model to handle the complexity and variability in RSIs effectively. FAENet balances local spatial detail capture with long-range dependency modeling.
3. Extensive experiments on the ISPRS Potsdam [55] and LoveDA [56] benchmarks demonstrate that FAENet demonstrates state-of-the-art segmentation accuracy, achieving improvements across key metrics such as AF, OA, and mIoU. Ablation studies further confirm the critical roles of ICCA and CCCA in spectral–spatial modeling, validating the robustness and generalizability of the proposed approach for complex remote sensing tasks. Moreover, efficiency comparisons confirm the superiority of the proposed FAENet.

This paper is organized as follows: Section 2 provides an overview of related works in the semantic segmentation of RSIs and the advanced methods based on frequency analysis. Section 3 introduces FAENet. Section 4 gives the results. Section 5 draws the conclusion and points out future directions.

2. Related Works

2.1. CNN-Based Methods for Semantic Segmentation of Remote Sensing Images

CNNs have become a cornerstone for the semantic segmentation of RSIs, demonstrating their ability to extract hierarchical features effectively. RSIs, characterized by high spatial resolution and complex spectral variations, has driven the adaptation of CNN architectures to address unique challenges in this domain. A key advancement is the ResUNet-a model [57], which incorporates residual connections and atrous convolutions to enhance multiscale feature learning while maintaining efficient feature propagation. This architecture has shown significant improvements in handling RSIs with intricate spatial structures and complex boundaries. Similarly, D-LinkNet [58] extends traditional CNN-based architectures by introducing dense connections within an encoder–decoder framework, achieving enhanced feature reuse and better segmentation accuracy for road extraction tasks in RSIs. PSPNet [59] has been adapted for RSIs by leveraging pyramid pooling to capture global context, addressing the challenge of large-scale scene variability in RSIs. For vegetation segmentation, methods like DeepLabV3+ [60] have been customized to incorporate multiscale feature extraction and boundary refinement, enabling accurate mapping of vegetation classes in high-resolution aerial imagery.

Despite these advancements, CNN-based methods exhibit notable limitations when applied to RSIs. Specifically, they struggle to capture long-range dependencies due to their reliance on localized convolutional operations. Additionally, their primary focus on spatial features leads to an underutilization of the rich spectral information inherent in RSIs, which is critical for differentiating spectrally similar classes. These shortcomings highlight the necessity for advanced architectures that can integrate spectral and spatial contexts effectively, motivating the development of approaches like FAENet proposed in this work.

2.2. Attention-Based Methods for Semantic Segmentation of Remote Sensing Images

Attention mechanisms (AMs) have proven to be a powerful augmentation to CNNs, addressing their limitations in capturing long-range dependencies and improving contextual representation for semantic segmentation tasks. By dynamically emphasizing salient regions and focusing on relevant features, attention-based methods have significantly enhanced segmentation accuracy in RSIs, which are characterized by complex spatial and spectral patterns. CAS-Net [61] enhances small object segmentation by integrating coordinate attention (CA) and SPD-Conv layers to better capture orientation-sensitive and positional information. MTCNet [62] combines CBAMs with multiscale transformers for improved spatial–contextual modeling. Similarly, the AD-HRNet model [63] leverages high-resolution attention modules to refine feature representations at multiple scales. LANet [64] introduced a patch-wise attention module to preserve local details during multi-level feature fusion, resulting in improved segmentation accuracy. Similarly, Li et al. [65] proposed a hybrid attention mechanism, applying spatial attention in shallow layers to capture local features and channel attention in deeper layers to enhance hierarchical feature learning for satellite image segmentation. Following this trend, hybrid designs such as SCAttNet [66], HMANet [67], and HCANet [37] effectively combined spatial and channel attention mechanisms to enrich feature representations before final inference. MDANet [68] introduces a deformable attention module (DAM) to enhance locality awareness and structural adaptability for high-resolution remote sensing (HRRS) images, achieving significant performance gains with a multiscale strategy. In contrast, CCAFFMNet [69] focuses on dual-spectral (RGB-thermal) segmentation, utilizing channel-coordinate attention feature-fusion modules (CCAFFMs) to refine infrared and RGB feature integration. More recent advancements have focused on improving long-range dependency modeling and global contextual representation. For instance, WiCoNet [70] employed a dual-branch structure with two CNNs to independently model local and global features, effectively capturing long-range dependencies for improved segmentation performance. Li et al. [40] introduced a synergistic attention model (SAM) to simultaneously capture spatial and channel dependencies, mitigating attention bias often present in traditional methods. By integrating SAM, the SAPNet framework achieved state-of-the-art performance by enhancing feature representations with comprehensive contextual information.

Despite these advancements, attention-based methods still face notable limitations. First, they often struggle to effectively decouple spatial and spectral features, leading to suboptimal fusion and reduced segmentation accuracy in spectrally complex RSIs. Second, these methods face challenges in balancing the preservation of local details with the integration of long-range dependencies, both of which are crucial for accurate segmentation in high-resolution remote sensing tasks. These limitations highlight the need for advanced architectures, such as FAENet, that address these issues by integrating frequency attention mechanisms and synergizing spatial and spectral feature learning.

2.3. Transformer-Based Methods for Semantic Segmentation of Remote Sensing Images

Transformer-based architectures have emerged as a transformative approach in semantic segmentation, particularly for RSIs, due to their ability to model long-range dependencies and capture global context effectively. These capabilities make transformers well-suited to address the challenges posed by high-resolution RSIs, such as complex spatial structures and spectral variability. In the context of RSIs, transformer-based models have shown great potential in improving segmentation performance. Blaga and Nedevschi [46] proposed a transformer-based U-Net model with Guided Focal-Axial Attention, which combines the global attention mechanism of transformers with localized attention to enhance feature representation in high-resolution RSIs. Wang et al. [47] introduced UNetFormer, a hybrid architecture that incorporates transformer layers into a U-Net structure, enabling the model to effectively capture both local spatial details and long-range global dependencies, making it particularly suited for urban scene segmentation tasks. Lin et al. [48] presented the Swin Transformer Segmentor, which integrates Swin Transformers with CNN-based architectures to refine boundary information and achieve significant improvements in segmentation accuracy.

More advanced transformer-based frameworks have further refined these ideas. GLOTS [71] introduces a unified transformer encoder–decoder structure, leveraging a masked image modeling pre-training strategy and a global–local attention mechanism to capture multiscale contexts effectively. Another study [72] employs the Swin Transformer backbone with a densely connected feature aggregation module (DCFAM) to restore resolutions, demonstrating the strength of transformers in producing fine-grained segmentation maps. EMRT [73] combines CNNs and deformable self-attention mechanisms within a transformer-based architecture, enabling efficient multiscale representation learning by fusing local and global features. These works highlight the growing focus on overcoming the limitations of standalone CNNs or transformers through hybrid models and advanced attention mechanisms. Li et al. [49] proposed GPINet, a hybrid network that combines CNN and transformer features with geometric priors to improve the semantic segmentation of high-resolution RSIs. Similarly, He et al. [50] developed ST-UNet, which integrates a

Swin Transformer with a CNN-based U-Net to leverage both global context and detailed spatial features. Hybrid models such as CLCFormer [51] balance fine-grained spatial details with global contextual information through a cross-learning framework, achieving state-of-the-art performance on several very high-resolution (VHR) remote sensing datasets. Meanwhile, LETFormer [74] addresses structural limitations in tokenization by integrating intra-window self-attention with cross-window context interactions, enhancing global feature modeling and spatial representation. Both CLCFormer [51] and LETFormer [74] exemplify the advancements in transformer-based methods for RSI segmentation, tackling key challenges such as balancing local spatial details and global contextual understanding.

However, these methods still face limitations in decoupling spatial and spectral features effectively and integrating multiscale contextual information. The proposed FAENet aims to address these issues by incorporating frequency attention mechanisms, enabling a more robust and balanced approach to spectral and spatial feature learning, which is critical for accurate semantic segmentation of high-resolution RSIs.

2.4. Learning in Frequency Domain

In recent years, frequency domain learning has gained significant attention for its ability to complement spatial domain approaches in image processing tasks [52,53]. Frequency-based methods enable the separation of high-frequency components, such as textures and edges, from low-frequency components, such as smooth regions, facilitating nuanced feature extraction. Xu et al. [54] conducted a theoretical analysis of neural networks' spectral bias using Fourier transformations, revealing a natural inclination toward low-frequency representations and challenges in capturing high-frequency details. These findings catalyzed further research into frequency domain techniques aimed at improving high-frequency representation. For instance, Azad et al. [75] redesigned the self-attention mechanism to operate in the frequency domain, enhancing contextual cues and uncovering finer details for improved feature representation. Similarly, Zhang et al. [76] explored the transformation of spatial-domain CNNs into frequency-domain equivalents to harness their unique properties, enabling better utilization of spectral information. Additionally, the development of frequency channel attention (FCA)-based networks enabled explicit spectral feature processing without complex frequency transformations, showing promising results in various applications. Zhang et al. [77] further advanced this concept with FsaNet, a framework leveraging frequency self-attention to improve edge preservation and computational efficiency. FsaNet demonstrated state-of-the-art performance on several benchmarks, achieving significant improvements in segmentation accuracy and efficiency.

In the domain of remote sensing, frequency domain methods have shown substantial potential for addressing the challenges of high-resolution remote sensing imagery (HRRSI). Techniques such as discrete cosine transformation (DCT) and Fourier analysis enable the effective separation of spectral components, aiding in feature extraction for tasks like semantic segmentation. Su et al. [78] proposed the Complete Frequency Channel Attention Network (CFCANet), which integrates DCT frequency components into feature maps by assigning the most significant eigenvalues to each channel. This approach significantly enhances noise resistance, particularly in noisy remote sensing imagery. For semantic segmentation, Li et al. [79] introduced the Spectrum-Space Collaborative Network (SSCNet), which employs a joint spectral–spatial attention (JSSA) module to simultaneously model spectral (SpeA) and spatial (SpaA) dependencies, leading to improved segmentation quality in HRRSIs. Hybrid architectures that integrate spatial and frequency domain features have also emerged as a promising direction for remote sensing tasks. Hong et al. [80] presented the Spatial-Frequency Information Integration Network (SFINet), which leverages invertible neural operators in the spatial domain and deep Fourier transformation in the frequency domain. SFINet excels in multimodal image fusion tasks, such as pan-sharpening and depth super-resolution, by effectively integrating local and global information to capture high-frequency details. Similarly, the Fourier Frequency Domain Convolutional Neural Network (FFDCNet) [81] introduces a dynamic frequency filtering mechanism that decomposes feature maps into low- and high-frequency components, improving classification accuracy and segmentation robustness. FFDCNet has shown particular effectiveness in addressing fragmented boundaries and incomplete extractions in crop segmentation tasks, making it highly relevant for remote sensing applications.

Despite these advancements, existing frequency domain methods still face notable limitations. Many approaches struggle to jointly integrate spectral and spatial features effectively, resulting in suboptimal segmentation performance, particularly in spectrally complex and high-resolution RSIs. Furthermore, current methods often lack a unified framework to balance local detail preservation and global contextual understanding across spatial and frequency domains. These limitations motivate the development of FAENet, which explicitly combines spectral and spatial learning by leveraging frequency attention mechanisms, providing a more robust and efficient solution for semantic segmentation of HRRSIs.

3. Method

3.1. Overview of FAENet

FAENet, as illustrated in Figure 1, is a frequency attention-enhanced network designed for the semantic segmentation of RSIs. It adopts a symmetric encoder–decoder architecture wherein the encoder progressively down-samples the input to extract multiscale features, while the decoder up-samples the encoded features to produce a pixel-wise segmentation mask. The core innovation of FAENet lies in its integration of the proposed Frequency Attention Model (FreqA) within the encoder–decoder framework, enabling joint spectral and spatial feature learning. Each stage of the encoder and decoder consists of multiple Conv Blocks interleaved with FreqA modules, ensuring both local detail preservation and global context modeling.

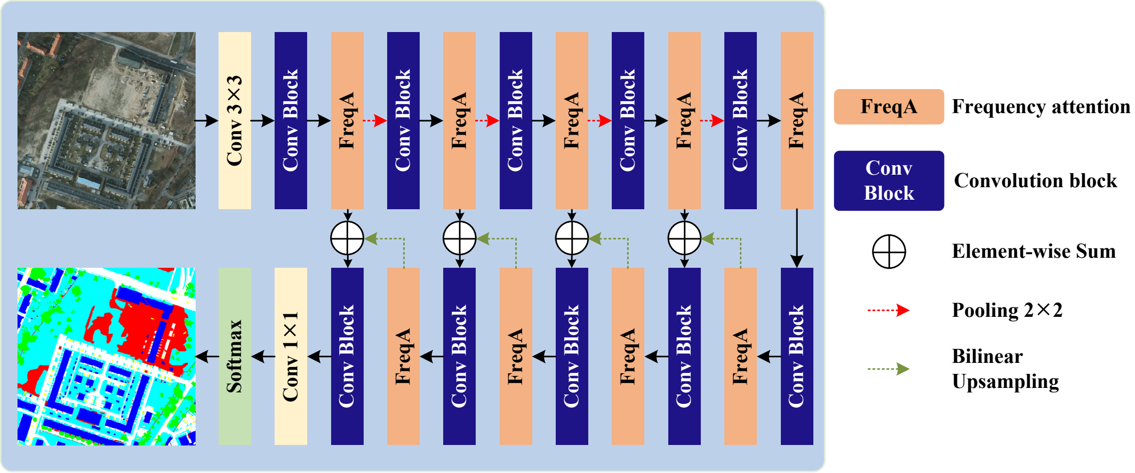

Figure 1. Overview of the proposed FAENet.

The encoder is based on a modified ResNet-50 backbone, where each Conv Block corresponds to a residual block from ResNet-50. A Conv Block consists of three convolutional layers with kernel size 3 × 3, followed by batch normalization and ReLU activation. Skip connections are employed within each Conv Block to enable efficient gradient propagation and prevent vanishing gradients during training.

Let the input image be denoted as $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels, respectively. The encoder progressively down-samples the input by a factor of 2 using pooling layers, resulting in feature maps $\mathbf{F}_i$ at different scales, where i denotes the level in the encoder hierarchy.

$$\mathbf{F}_i = \mathrm{ConvBlock}(\mathbf{F}_{i-1}), \quad i \in \{1, 2, \ldots, N\} \tag{1}$$

Here, $\mathbf{F}_0 = \mathbf{I}$, and N is the total number of Conv Blocks in the encoder. After each Conv Block, a FreqA module is applied to enhance the extracted features by incorporating spectral and spatial contexts, but its details will be elaborated in Section 3.2.

The decoder follows a symmetric structure, where each level up-samples the corresponding encoder feature map using bilinear interpolation. Let $\mathbf{U}_i$ denote the up-sampled feature map at level i in the decoder, which is computed as follows:

$$\mathbf{U}_i = \mathrm{BilinearUpsample}(\mathbf{F}_{N-i}), \quad i \in \{1, 2, \ldots, N\} \tag{2}$$

To improve feature refinement, the up-sampled feature map $\mathbf{U}_i$ is concatenated with the corresponding feature map from the encoder at the same level:

$$\mathbf{G}_i = \mathrm{Concat}(\mathbf{U}_i, \mathbf{F}_i) \tag{3}$$

The concatenated feature $\mathbf{G}_i$ is then processed by a Conv Block to refine the combined representation before being passed to the next level in the decoder. The final output is obtained by applying a 1×1 convolution followed by a softmax activation to produce a pixel-wise class probability map $\mathbf{O} \in \mathbb{R}^{H \times W \times K}$, where K is the number of classes:

$$\mathbf{O} = \mathrm{Softmax}(\mathrm{Conv}_{1 \times 1}(\mathbf{G}_0)) \tag{4}$$

In summary, FAENet employs a carefully designed encoder–decoder architecture with integrated frequency attention to address the complex spectral–spatial characteristics of HRRSIs. By combining hierarchical feature extraction, multiscale feature fusion, and frequency attention mechanisms, FAENet is well-suited for high-resolution semantic segmentation tasks.

3.2. Frequency Attention Model

The Frequency Attention Model (FreqA) is designed to enhance feature representations by jointly modeling spectral and spatial contexts, which are crucial for accurate semantic segmentation in RSIs. As shown in Figure 2, FreqA operates by first transforming the input feature maps into the frequency domain using Discrete Wavelet Transformation (DWT) and applying attention mechanisms to emphasize the most informative spectral components. Finally, the refined features are transformed back into the spatial domain using Inverse Discrete Wavelet Transformation (iDWT).

Given an input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels, FreqA begins by applying DWT to decompose $\mathbf{X}$ into four frequency components: $\mathbf{F}_{LL}$: Low-frequency component, representing the coarse approximation of the input feature map. $\mathbf{F}_{LH}$: High-frequency component in the horizontal direction. $\mathbf{F}_{HL}$: High-frequency component in the vertical direction. $\mathbf{F}_{HH}$: High-frequency component in the diagonal direction.

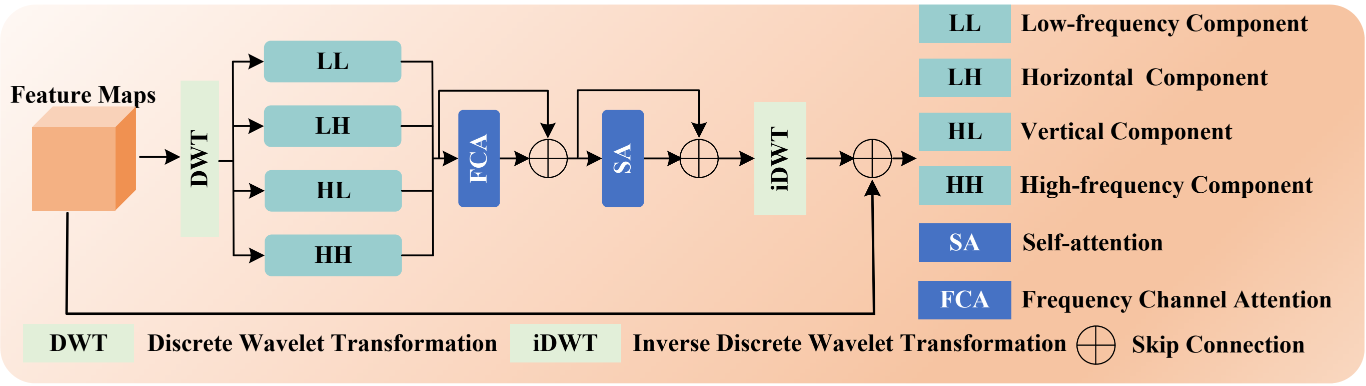
Feature Maps
DWT
LL
LH
HL
HH
FCA
SA
iDWT
LL Low-frequency Component
LH Horizontal Component
HL Vertical Component
HH High-frequency Component
SA Self-attention
FCA Frequency Channel Attention
DWT Discrete Wavelet Transformation
iDWT Inverse Discrete Wavelet Transformation
Skip Connection

Figure 2. Details of the proposed FreqA.

The result is a set of four sub-band feature maps, each of size $\frac{H}{2} \times \frac{W}{2} \times C$:

$$\{\mathbf{F}_{LL}, \mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}\} = \mathrm{DWT}(\mathbf{X}), \tag{5}$$

where DWT facilitates the separation of low- and high-frequency components, enabling nuanced feature extraction for tasks such as edge preservation and texture representation.

Once the feature maps are transformed into the frequency domain, FreqA applies the Frequency Channel Attention (FCA) mechanism to selectively emphasize important spectral information. FCA consists of two key sub-modules: (1) Inner-Component Channel Attention (ICCA): Enhances the discriminative power of each frequency component by modeling channel-wise dependencies within the same component. (2) Cross-Component Channel Attention (CCCA): Captures correlations across different frequency components, improving spectral coherence across the frequency domain.

For each frequency component $\mathbf{F}_k \in \{\mathbf{F}_{LL}, \mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}\}$, the FCA module produces refined feature maps $\mathbf{F}_k^{\mathrm{FCA}}$ by sequentially applying ICCA and CCCA:

$$\mathbf{F}_k^{\mathrm{FCA}} = \mathrm{CCCA}(\mathrm{ICCA}(\mathbf{F}_k)). \tag{6}$$

The outputs from all frequency components are then concatenated along the channel dimension to form an aggregated feature map:

$$\mathbf{F}_{\mathrm{agg}} = \mathrm{Concat}(\mathbf{F}_{LL}^{\mathrm{FCA}}, \mathbf{F}_{LH}^{\mathrm{FCA}}, \mathbf{F}_{HL}^{\mathrm{FCA}}, \mathbf{F}_{HH}^{\mathrm{FCA}}). \tag{7}$$

To further capture long-range dependencies and enhance global contextual representation, FreqA applies a self-attention (SA) module to the aggregated feature map $\mathbf{F}_{\mathrm{agg}}$. The self-attention mechanism computes query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) representations of the feature map:

$$\mathbf{Q} = \mathbf{W}_q \mathbf{F}_{\mathrm{agg}}, \quad \mathbf{K} = \mathbf{W}_k \mathbf{F}_{\mathrm{agg}}, \quad \mathbf{V} = \mathbf{W}_v \mathbf{F}_{\mathrm{agg}}, \tag{8}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable projection matrices. The attention map is computed using the scaled dot-product operation:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{9}$$

where d_k is the dimensionality of the key. The output of the self-attention module, denoted as $\mathbf{F}_{\mathrm{SA}}$, is computed by applying the attention map to the value representation:

$$\mathbf{F}_{\mathrm{SA}} = \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \tag{10}$$

Finally, the refined feature map $\mathbf{F}_{SA}$ is transformed back into the spatial domain using Inverse Discrete Wavelet Transformation (iDWT), producing the output feature map $\mathbf{Y} \in \mathbb{R}^{H\times W\times C}$:

$$\mathbf{Y} = \mathrm{iDWT}(\mathbf{F}_{SA}). \tag{11}$$

This process enables FAENet to jointly model spectral and spatial information, addressing the key challenges in high-resolution remote sensing segmentation tasks.

3.3. Frequency Channel Attention

The Frequency Channel Attention (FCA) module is a key component of the proposed Frequency Attention Model (FreqA), designed to enhance feature representations by capturing both intra-component and cross-component spectral dependencies. As shown in Figure 3, ICCA operates on individual frequency components to refine channel-wise features, while CCCA models interactions across different frequency components to improve spectral coherence. The final output of FCA is a concatenated feature map that integrates refined information from all frequency components.

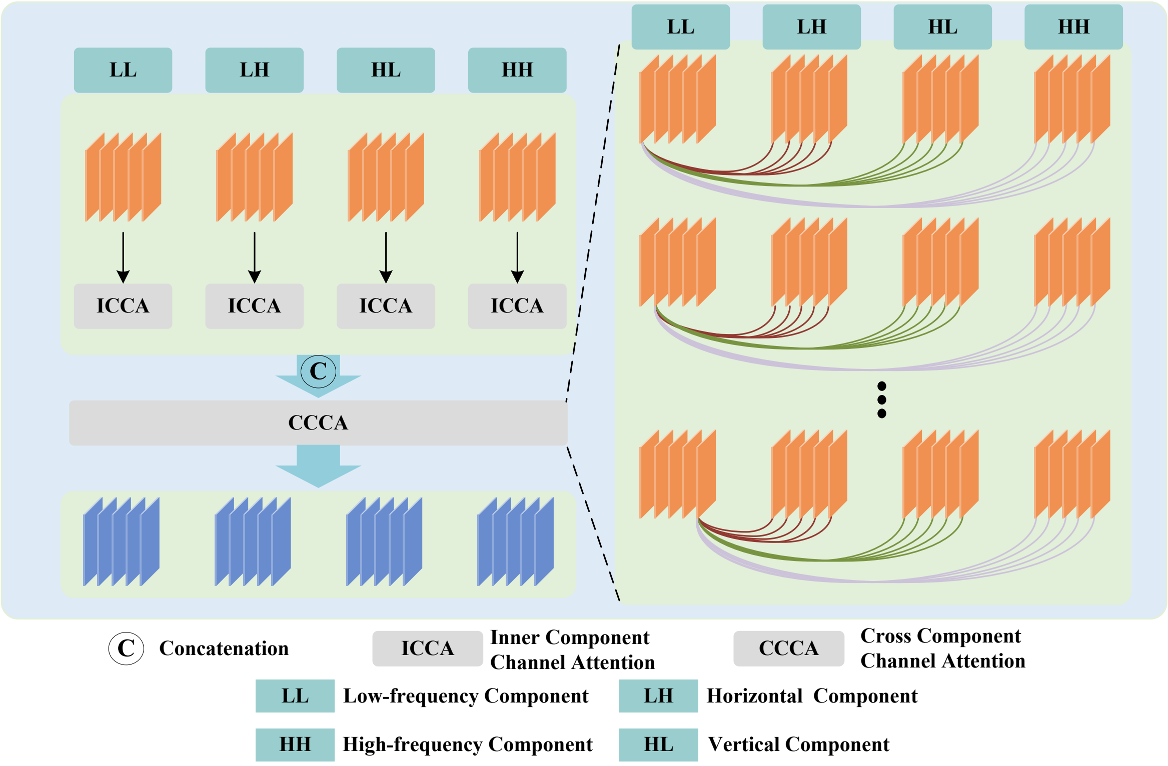
LL
LH
HL
HH
ICCA
ICCA
ICCA
ICCA
C
CCCA
LL
LH
HL
HH
C Concatenation
ICCA Inner Component Channel Attention
CCCA Cross Component Channel Attention
LL Low-frequency Component
LH Horizontal Component
HH High-frequency Component
HL Vertical Component

Figure 3. Details of the proposed FCA.

ICCA aims to enhance the discriminative power of each frequency component by modeling channel-wise dependencies within the same component. Given a frequency component $\mathbf{F}_k \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$, where H, W, and C are the height, width, and number of channels, respectively, ICCA computes a refined feature map $\mathbf{F}_k^{\mathrm{ICCA}}$ by applying a channel attention mechanism.

First, a channel descriptor $\mathbf{z}_k \in \mathbb{R}^C$ is obtained by applying global average pooling (GAP) across spatial dimensions:

$$\mathbf{z}_k(c) = \frac{1}{\frac{H}{2}\cdot\frac{W}{2}} \sum_{i=1}^{\frac{H}{2}} \sum_{j=1}^{\frac{W}{2}} \mathbf{F}_k(i,j,c), \quad c \in \{1,2,\ldots,C\}, \tag{12}$$

where $\mathbf{F}_k(i,j,c)$ denotes the value at position (i,j) in channel c of the frequency component $\mathbf{F}_k$.

Next, a fully connected (FC) layer followed by a ReLU activation is applied to $\mathbf{z}_k$ to obtain a transformed channel descriptor $\mathbf{z}_k^{\mathrm{FC}}$:

$$\mathbf{z}_k^{\mathrm{FC}} = \mathrm{ReLU}(\mathbf{W}_1\mathbf{z}_k + \mathbf{b}_1), \tag{13}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\times\frac{C}{r}}$ and $\mathbf{b}_1 \in \mathbb{R}^{\frac{C}{r}}$ are learnable parameters, and r is the reduction ratio.

To obtain the final attention weights, another FC layer with a sigmoid activation is applied:

$$\mathbf{a}_k = \mathrm{Sigmoid}(\mathbf{W}_2\mathbf{z}_k^{\mathrm{FC}} + \mathbf{b}_2), \tag{14}$$

where $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{b}_2 \in \mathbb{R}^{C}$ are learnable parameters. The attention weights $\mathbf{a}_k \in \mathbb{R}^{C}$ are then used to reweight the original feature map $\mathbf{F}_k$ channel-wise:

$$\mathbf{F}_k^{\mathrm{ICCA}}(i,j,c) = \mathbf{a}_k(c)\cdot\mathbf{F}_k(i,j,c), \tag{15}$$

where $\mathbf{F}_k^{\mathrm{ICCA}}$ is the refined output of ICCA for the frequency component $\mathbf{F}_k$.

After applying ICCA to each frequency component independently, CCCA is employed to capture cross-frequency dependencies by modeling interactions across different frequency components. The goal of CCCA is to improve spectral coherence by leveraging correlations between the low-frequency and high-frequency components.

Let $\mathbf{F}_{LL}^{\mathrm{ICCA}}, \mathbf{F}_{LH}^{\mathrm{ICCA}}, \mathbf{F}_{HL}^{\mathrm{ICCA}}, \mathbf{F}_{HH}^{\mathrm{ICCA}}$ denote the refined frequency components after ICCA. For a given channel c, CCCA computes the correlation between the low-frequency component $\mathbf{F}_{LL}^{\mathrm{ICCA}}$ and the high-frequency components $\mathbf{F}_{LH}^{\mathrm{ICCA}}, \mathbf{F}_{HL}^{\mathrm{ICCA}}, \mathbf{F}_{HH}^{\mathrm{ICCA}}$:

$$\mathbf{a}_{LL}(c) = \mathbf{F}_{LL}^{\mathrm{ICCA}}(c) + \sum_{k\in\{LH,HL,HH\}} \alpha_k\cdot\mathrm{Corr}(\mathbf{F}_{LL}^{\mathrm{ICCA}}(c), \mathbf{F}_k^{\mathrm{ICCA}}(c)), \tag{16}$$

where α_k are learnable weights, and $\mathrm{Corr}(\cdot,\cdot)$ denotes a correlation function that employs cosine similarity to compute across corresponding channels in different frequency components. This operation ensures that each channel in the low-frequency component is enriched with cross-component information from the high-frequency components.

The same process is applied to all channels in all frequency components, resulting in refined feature maps $\mathbf{F}_{LL}^{\mathrm{CCCA}}, \mathbf{F}_{LH}^{\mathrm{CCCA}}, \mathbf{F}_{HL}^{\mathrm{CCCA}}, \mathbf{F}_{HH}^{\mathrm{CCCA}}$.

Finally, the refined frequency components after applying CCCA are concatenated along the channel dimension to form the output feature map of FCA:

$$\mathbf{F}_{\mathrm{FCA}} = \mathrm{Concat}(\mathbf{F}_{LL}^{\mathrm{CCCA}}, \mathbf{F}_{LH}^{\mathrm{CCCA}}, \mathbf{F}_{HL}^{\mathrm{CCCA}}, \mathbf{F}_{HH}^{\mathrm{CCCA}}), \tag{17}$$

where $\mathbf{F}_{\mathrm{FCA}} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 4C}$ is the final output of the FCA module, containing enriched spectral and spatial information from all frequency components.

To sum up, FCA ensures that both intra-frequency and inter-frequency dependencies are effectively captured, enabling FAENet to jointly model spectral and spatial contexts for enhanced segmentation performance in HRRSIs.

4. Experiments

4.1. Settings

In experiments, we implemented our FAENet and all comparative models under the same settings, using PyTorch on a Linux OS with an NVIDIA A40 GPU. Data augmentations, such as random flipping and cropping operations, were applied to all datasets and networks. The initial learning rate and maximum epoch were fixed at 0.02 and 500, respectively.

We adopt stochastic gradient descent (SGD) as the optimizer, with a momentum of 0.9. The learning rate was adjusted using a polynomial decay strategy, defined as follows:

$$\eta_t = \eta_0 \left(1 - \frac{t}{T}\right)^p, \tag{18}$$

where η_t is the learning rate at iteration t, η_0 is the initial learning rate, T is the total number of iterations, and p is a decay exponent set to 0.9 in our experiments. This strategy ensures that the learning rate smoothly decreases as the number of iterations increases, helping the model achieve better convergence. The model parameter file with the lowest validation loss was saved for final evaluation.

We employ four metrics to evaluate the performance of the predicted results on the test set: class-wise $F1$-score ($F1$), average $F1$-score across all classes (AF), overall accuracy (OA), and mean intersection over union (mIoU). The equations for $F1$, OA, and IoU are provided in Equation (19), Equation (20), and Equation (21), respectively.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}, \tag{19}$$

$$\mathrm{OA} = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{20}$$

$$\mathrm{IoU} = \frac{TP}{(TP + FP + FN)}, \tag{21}$$

Here, *precision* and *recall* are defined as follows:

$$precision = \frac{TP}{TP + FP}, \tag{22}$$

$$recall = \frac{TP}{TP + FN}, \tag{23}$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative samples, respectively.

As for comparative methods, we selected several well-established baselines and state-of-the-art (SOTA) methods tailored for RSI segmentation. FCN-8s [30] and DANet [36] are pioneering fully convolutional and attention-based models that have achieved notable success in general computer vision tasks. For RSIs, ResUNet-a [57] was designed specifically for RSI segmentation by incorporating atrous convolution, multiscale feature fusion, and residual connections, effectively addressing the challenges posed by complex spatial structures in RSIs.

More recent advancements in attention-based methods for RSI segmentation include MACU-Net [38], HCANet [37], SCAttNet [66], and A2FPN [39], which leverage various attention mechanisms to improve feature representation and segmentation accuracy. These models, published after 2021, have demonstrated state-of-the-art performance across multiple RSI datasets by dynamically capturing long-range dependencies and enhancing spatial context understanding.

Furthermore, we included the latest transformer-based architectures tailored for RSI segmentation: ICTNet [82], CLCFormer [51], and LETFormer [74]. These models exploit the global attention capabilities of transformers, providing a more holistic understanding of complex scenes in high-resolution RSIs. Specifically, CLCFormer employs a cross-learning mechanism to balance fine-grained spatial details with global context, while LETFormer introduces a novel intra-window self-attention mechanism for improved structural modeling in RSIs. These methods serve as strong baselines for benchmarking the performance of our proposed FAENet.

4.2. Datasets

The ISPRS Potsdam dataset [55] comprises 38 high-resolution orthophotos with a ground sampling distance (GSD) of 5 cm, each measuring 6000 × 6000 pixels. The dataset includes four spectral bands: near-infrared (NIR), red (R), green (G), and blue (B), along with corresponding digital surface model (DSM) and normalized digital surface model (NDSM) data. These additional data sources provide valuable elevation information, aiding in distinguishing objects with similar spectral characteristics. The dataset covers diverse urban scenes, such as buildings, roads, trees, and cars, making it a benchmark for evaluating semantic segmentation performance in high-resolution urban imagery. For this study, we focused on the RGB image data for training and testing, as they are widely used in standard semantic segmentation tasks.

The LoveDA dataset [56] presents a unique challenge in semantic segmentation by incorporating large-scale satellite images with a spatial resolution of 0.3 m. The dataset spans over 536 square kilometers and includes both rural and urban regions from three cities: Nanjing, Changzhou, and Wuhan. Each image has a spatial size of 1024 × 1024 pixels and exhibits substantial variability in object scale, size, and surface type. LoveDA is designed to evaluate the robustness of segmentation models in handling imbalanced class distributions and challenging environmental conditions, such as varying lighting and atmospheric effects. For our experiments, we utilized 2522 images for training, 834 for validation, and 835 for testing.

For both datasets, the images were cropped into subpatches of size 256 × 256 to ensure uniform input dimensions for the model. These subpatches were randomly divided into training, validation, and testing sets in a 1:1:1 ratio, providing a balanced and comprehensive evaluation framework.

4.3. Results on the ISPRS Potsdam Dataset

4.3.1. Numerical Evaluations

The numerical results in Table 1 demonstrate the superior performance of FAENet compared to existing methods across key evaluation metrics, highlighting its effectiveness in semantic segmentation of the ISPRS Potsdam dataset. FAENet achieves the highest OA of 92.31%, surpassing both LETFormer (91.17%) and CLCFormer (89.97%), which are state-of-the-art transformer-based methods. This improvement underscores the ability of FAENet to generalize effectively across diverse scenes, benefiting from the proposed frequency attention mechanism that enhances both spectral and spatial feature extraction.

In terms of mIoU, FAENet achieves a score of 83.58%, outperforming LETFormer (82.67%) by nearly 1% and CLCFormer (81.68%) by almost 2%. The consistent improvement in mIoU indicates that FAENet excels in capturing inter-class separability and handling challenging scenarios with overlapping class boundaries. Despite the improvement being under 1%, statistical significance tests (t-tests) were conducted over five repeated runs, confirming that the observed improvements are statistically significant ($p < 0.05$).

When examining class-wise F1-scores, FAENet demonstrates outstanding performance in the "Low vegetation" and "Car" categories, achieving scores of 88.21 and 94.75, respectively. These results are particularly noteworthy because "Car" is a small and intricate class, often challenging for segmentation models due to its limited spatial representation. Similarly, FAENet's ability to achieve the highest F1-score in "Low vegetation" reflects its effectiveness in managing fine-grained spectral details, which are critical for distinguishing between similar classes in HRRSIs. Compared to LETFormer, FAENet improves the F1-score for "Low vegetation" by nearly 1% (88.21 vs. 87.25) and for "Car" by 0.42% (94.75 vs. 94.33), indicating consistent performance across both large and small objects.

FAENet also achieves an AF of 92.71, reflecting its balanced segmentation performance across all classes. This improvement over LETFormer (92.47) and CLCFormer (91.61) demonstrates the efficacy of the frequency channel attention and self-attention modules in harmonizing spectral and spatial features. These mechanisms allow FAENet to mitigate class imbalances and achieve more precise segmentation results, especially in complex urban environments.

The overall trends in the results demonstrate that FAENet consistently outperforms existing state-of-the-art methods across various evaluation metrics. Specifically, FAENet achieves superior class-wise segmentation precision, particularly in challenging categories such as "Low vegetation" and "Car", which require precise boundary delineation and fine-grained feature discrimination. The improvements in these categories indicate that the proposed frequency attention mechanism effectively captures and integrates both spectral and spatial information, leading to better overall segmentation accuracy. Moreover, FAENet's performance in terms of OA, mIoU, and AF highlights its robustness in handling diverse and complex urban scenes in RSIs. The combination of frequency-based feature decomposition and attention mechanisms allows FAENet to generalize well across varying scene types and object scales, making it a strong candidate for real-world RSI segmentation applications.

Table 1. Results on the ISPRS Potsdam dataset. Class-wise F1-score, AF, OA, and mIoU are listed, where the bold text indicates the best results.

Methods	Impervious Surfaces	Building	Low Vegetation	Tree	Car	AF	OA	mIoU
FCN-8s [30]	85.08	74.26	65.86	80.68	39.11	69.00	68.24	63.84
DANet [36]	86.37	91.15	79.60	79.17	88.49	84.96	83.17	76.38
ResUNet-a [57]	89.79	94.77	86.61	81.01	76.85	85.80	84.38	77.19
MACU-Net [38]	88.56	91.86	86.21	82.22	78.48	85.46	85.07	77.23
HCANet [37]	92.25	95.96	86.66	87.30	92.97	91.03	89.79	81.12
SCAttNet [66]	91.25	95.96	84.66	86.20	91.88	89.99	88.19	80.05
A2FPN [39]	90.05	94.76	84.95	84.61	90.66	89.01	87.40	79.89
ICTNet [82]	91.78	95.44	86.23	86.92	92.53	90.58	89.38	80.40
CLCFormer [51]	92.66	96.64	87.05	88.00	93.69	91.61	89.97	81.68
LETFormer [74]	**94.19**	97.49	87.25	89.11	94.33	92.47	91.17	82.67
Ours (FAENet)	93.84	**97.57**	**88.21**	**89.18**	**94.75**	**92.71**	**92.31**	**83.58**

4.3.2. Statistical Significance Analysis

To ensure the robustness of our results and address potential variability caused by random initialization, we conducted repeated experiments with fixed random seeds. Specifically, we ran the experiments five times and computed the mean and standard deviation of all overall evaluation metrics. A two-tailed paired t-test was performed to assess the statistical significance of the improvements achieved by FAENet compared to LETFormer, the most competitive baseline. Table 2 summarizes the mean, standard deviation, and p-values obtained from the t-test. The p-values for all metrics are below 0.05, indicating that the improvements achieved by FAENet are statistically significant.

4.3.3. Visual Comparisons

The visual comparisons presented in Figure 4 showcase the segmentation outputs of various state-of-the-art methods on the ISPRS Potsdam dataset. FAENet consistently produces clearer segmentation maps, particularly in regions with intricate boundaries and small objects. In complex scenes with transitions between "Building" and "Impervious surfaces", FAENet demonstrates superior boundary delineation compared to earlier methods such as FCN-8s (c) and DANet (d), which tend to blur boundaries and introduce artifacts. The proposed frequency attention mechanism enables FAENet to preserve fine-grained details, resulting in sharper edges and fewer misclassified pixels.

Table 2. Mean, standard deviation, and *p*-values from *t*-test analysis over five repeated runs on the ISPRS Potsdam dataset.

Metric	Mean ± Std (FAENet)	Mean ± Std (LETFormer)	*p*-Value
OA (%)	92.31 ± 0.12	91.17 ± 0.14	0.012
mIoU (%)	83.58 ± 0.18	82.67 ± 0.21	0.015
AF (%)	92.71 ± 0.10	92.47 ± 0.13	0.020

FAENet also excels in segmenting small and detailed objects like "Car", where methods such as ResUNet-a (e) and MACU-Net (f) struggle with fragmentation. As shown in the marked regions of the figure, FAENet achieves more cohesive and accurate car segments, underscoring its capability to enhance high-frequency feature representation through the frequency attention mechanism.

When comparing FAENet with advanced transformer-based models like CLCFormer (k) and LETFormer (l), FAENet demonstrates better spatial consistency and reduced boundary misalignment. Although LETFormer produces reasonably accurate results, minor inconsistencies are observed in densely vegetated areas ("Low vegetation" and "Tree"), where FAENet delivers smoother transitions and better-defined class boundaries. The highlighted regions in Figure 4 indicate FAENet's ability to maintain structural integrity and spatial coherence in complex scenes.

Overall, the marked improvements in the figure emphasize FAENet's robustness in capturing both local and global contexts, leading to segmentation outputs that closely resemble ground truth labels. These results validate FAENet as a highly effective solution for RSI segmentation.

4.4. Results on the LoveDA Dataset

4.4.1. Numerical Evaluations

As shown in Table 3, FAENet achieves superior performance across key metrics on the LoveDA dataset, which poses unique challenges due to its mixed urban and rural landscapes, varying spatial resolutions, and imbalanced class distributions. The dataset's diversity in land cover scenarios requires models to generalize well across both densely built-up urban areas and sparsely populated rural regions.

Table 3. Results on the LoveDA dataset. Class-wise F1-score, AF, OA, and mIoU are listed, where the bold text indicates the best results.

Methods	Background	Building	Road	Water	Barren	Forest	Agriculture	AF	OA	mIoU
FCN-8s [30]	50.09	52.79	51.48	75.28	12.75	44.82	61.44	49.81	49.24	46.07
DANet [36]	53.92	60.40	62.73	78.37	26.36	51.75	69.31	57.55	54.09	49.67
ResUNet-a [57]	54.46	61.56	64.91	80.21	28.95	53.56	73.32	59.57	58.35	53.38
MACUNet [38]	58.56	63.43	66.05	80.19	31.90	55.24	75.02	61.49	59.05	53.61
HCANet [37]	65.72	70.04	74.35	87.39	50.62	63.27	80.25	70.23	68.77	62.13
SCAttNet [66]	65.28	71.15	76.26	85.73	50.27	60.57	81.17	70.06	66.63	60.47
A2FPN [39]	64.51	72.58	74.43	87.12	48.32	59.35	78.90	69.32	66.21	60.52
ICTNet [82]	67.63	74.84	78.22	88.30	52.69	65.35	81.50	72.65	69.71	62.71
CLCFormer [51]	67.17	74.34	77.69	87.71	52.34	64.91	80.96	72.16	70.45	62.55
LETFormer [74]	70.90	76.47	82.03	91.24	**56.75**	**70.05**	85.48	76.13	72.12	66.01
Ours (FAENet)	**70.93**	**81.74**	**82.82**	**92.50**	53.53	68.33	**85.57**	**76.49**	**72.93**	**66.91**

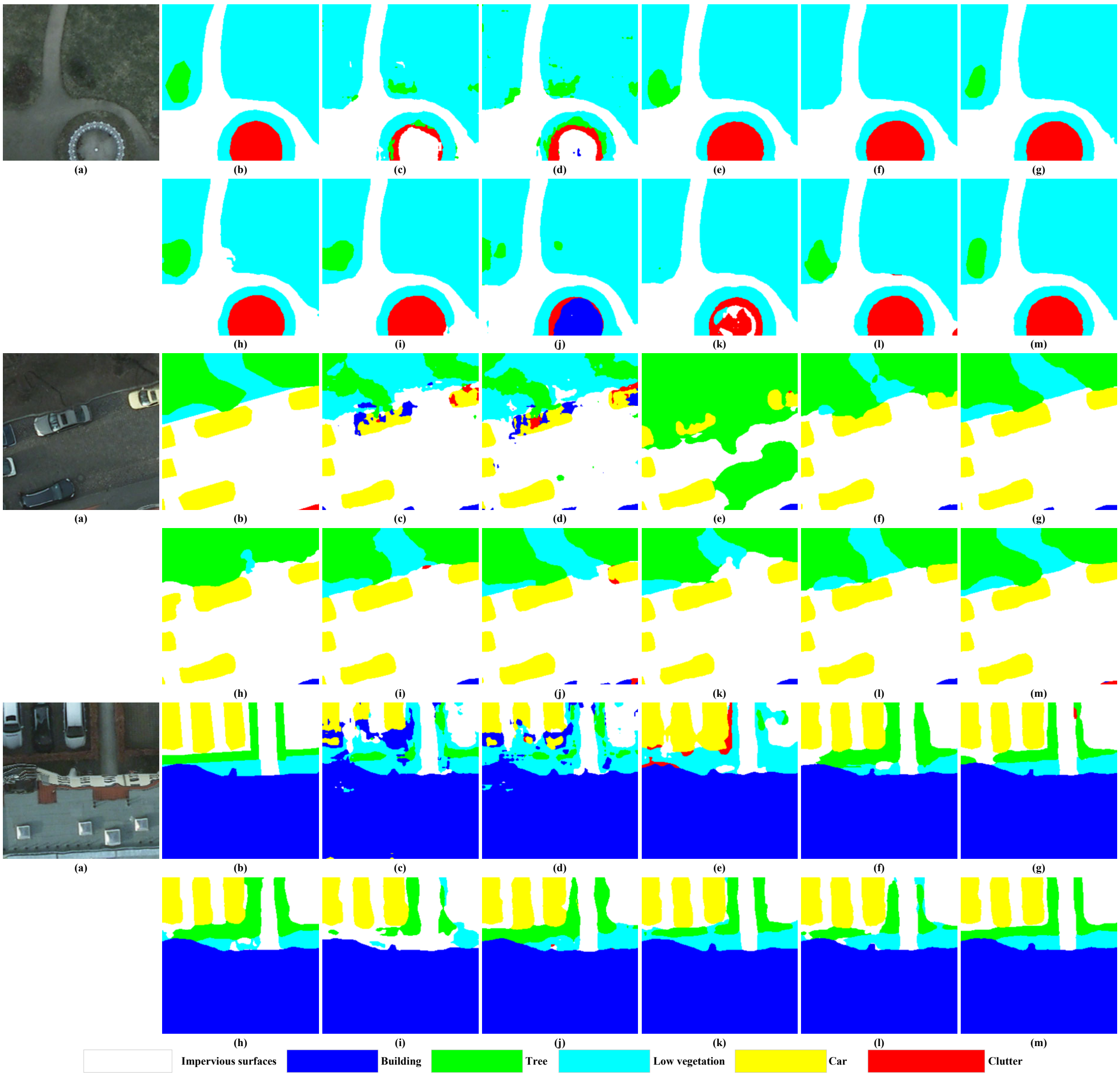

Figure 4. Visual comparisons of ISPRS Potsdam dataset. (**a**) Input, (**b**) ground truth, (**c**) FCN-8s, (**d**) DANet, (**e**) ResUNet-a, (**f**) MACU-Net, (**g**) HCANet, (**h**) SCAttNet, (**i**) A2FPN, (**j**) ICTNet, (**k**) CLCFormer, (**l**) LETFormer, (**m**) FAENet (ours).

FAENet attains the highest OA of 72.93%, outperforming LETFormer (72.12%) and CLCFormer (70.45%), demonstrating its improved generalization ability across complex landscape types. Additionally, FAENet achieves the best mIoU of 66.91%, surpassing LETFormer (66.01%) by 0.9%, highlighting its capacity to handle diverse land cover categories effectively.

In terms of class-wise F1-scores, FAENet excels in critical categories such as “Building” (81.74), “Road” (82.82), and “Water” (92.50), where accurate boundary delineation and fine-grained segmentation are crucial. Compared to LETFormer, FAENet records a 0.79% improvement in the “Road” category (82.82 vs. 82.03) and a notable 1.26% gain in the “Water” category (92.50 vs. 91.24). These improvements underscore FAENet’s ability to capture fine-grained details and segment linear and irregular features effectively.

Although LETFormer slightly outperforms FAENet in the “Barren” and “Forest” categories, with F1-scores of 56.75 and 70.05, respectively, FAENet remains competitive, achieving scores of 53.53 and 68.33 in these categories. Despite this trade-off, FAENet achieves the highest overall AF of 76.49, surpassing LETFormer (76.13) and other state-of-the-art models, indicating consistent performance across all classes.

The results validate FAENet's frequency attention mechanism, which enhances the integration of spectral and spatial features, enabling robust performance across diverse environments. Compared to the ISPRS Potsdam dataset, where the focus is on urban-specific segmentation, the LoveDA dataset presents additional challenges due to the significant class imbalance and landscape variability. FAENet's consistent improvements across both datasets highlight its versatility and robustness in handling different types of RSIs.

4.4.2. Visual Comparisons

The visual results in Figure 5 showcase segmentation outputs for the LoveDA dataset, enabling a comprehensive comparison of FAENet (m) with state-of-the-art models such as FCN-8s (c), DANet (d), ResUNet-a (e), and transformer-based architectures like CLCFormer (k) and LETFormer (l). Each subfigure highlights the ability of these models to delineate and classify diverse land cover types, including "Building", "Road", "Water", "Barren", "Forest", and "Agriculture".

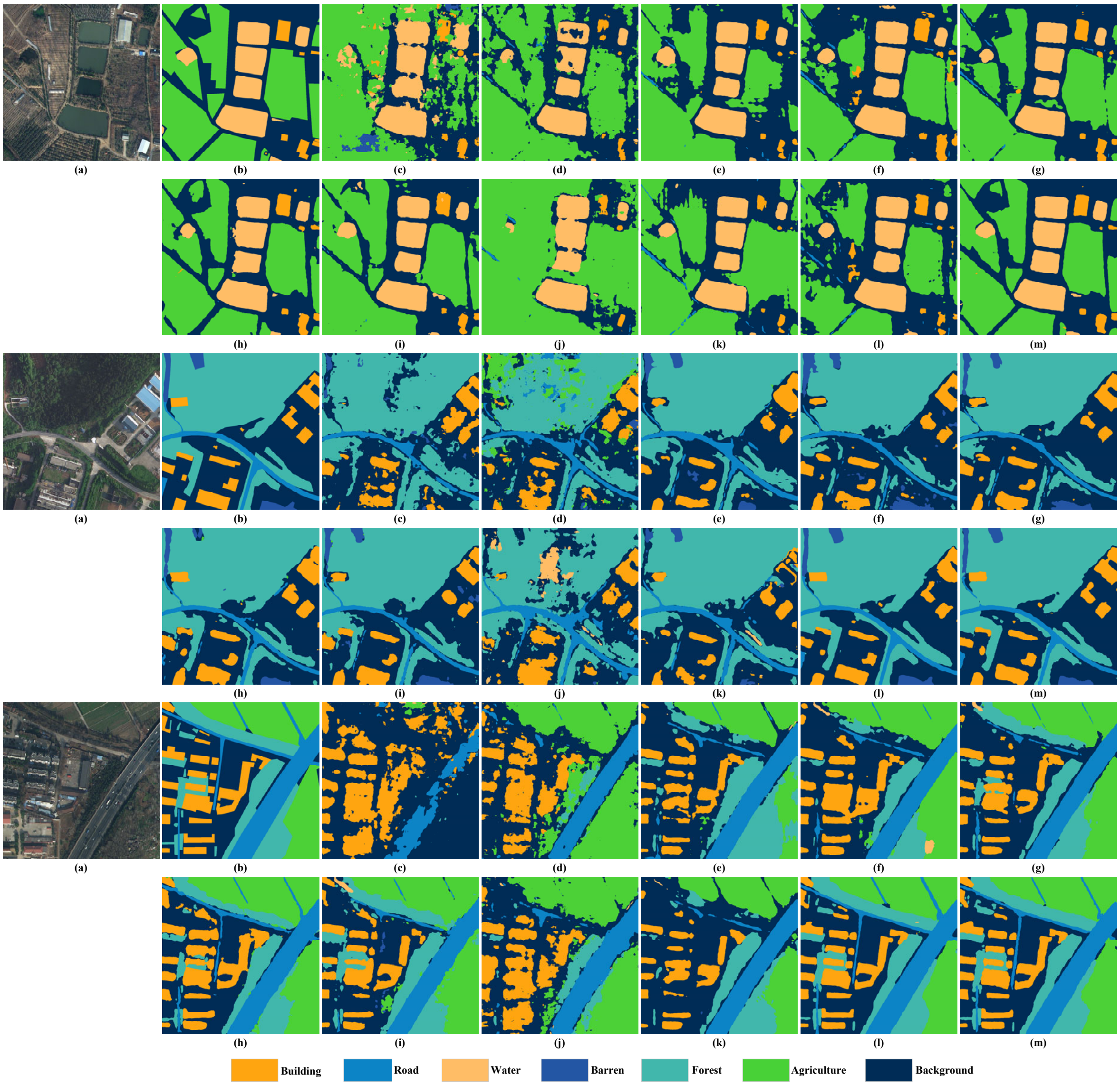

Figure 5. Visual comparisons of LoveDA dataset. (**a**) Input, (**b**) ground truth, (**c**) FCN-8s, (**d**) DANet, (**e**) ResUNet-a, (**f**) MACU-Net, (**g**) HCANet, (**h**) SCAttNet, (**i**) A2FPN, (**j**) ICTNet, (**k**) CLCFormer, (**l**) LETFormer, (**m**) FAENet (ours).

FAENet demonstrates superior segmentation quality, particularly in capturing fine-grained details and complex boundaries. For example, in areas dominated by “Building” and “Road”, FAENet generates outputs that closely match the ground truth (b), with sharper edges and fewer misclassifications compared to other models like FCN-8s (c) and DANet (d). These improvements underline the strength of FAENet’s frequency attention mechanism in refining both spectral and spatial representations.

Compared to ResUNet-a (e) and MACU-Net (f), FAENet significantly enhances the segmentation of challenging classes like “Water” and “Agriculture”. In “Water” regions, FAENet accurately captures smooth boundaries and mitigates over-segmentation issues prevalent in other models. Similarly, in agricultural areas, where fine texture details are crucial, FAENet outperforms other models by producing more uniform and accurate classifications.

Transformer-based models such as LETFormer (l) and CLCFormer (k) provide strong baseline results, particularly in handling large-scale features like “Forest.” However, FAENet surpasses these methods in maintaining spatial consistency and reducing noise in densely packed regions. For example, in mixed-class areas with overlapping “Barren” and “Forest” regions, FAENet demonstrates better discrimination and smoother transitions between classes.

Overall, the visual comparisons clearly illustrate FAENet’s ability to produce segmentation maps with superior boundary alignment, reduced artifacts, and enhanced class differentiation. These results validate the efficacy of FAENet’s spectral–spatial feature integration in handling the diverse and complex landscapes of the LoveDA dataset, further emphasizing its robustness and generalization capabilities.

4.5. Efficiency Analysis

The results in Table 4 demonstrate that FAENet achieves a strong balance between computational efficiency and segmentation accuracy, outperforming both CNN-based and transformer-based state-of-the-art methods in terms of inference speed and computational cost. FAENet’s inference time of 42.2 ms and 60.7 GFLOPs place it among the most efficient models evaluated.

FAENet exhibits a faster inference time compared to transformer-based architectures such as LETFormer (48.3 ms) and CLCFormer (49.6 ms). This reduction of approximately 6.1 ms and 7.4 ms, respectively, highlights FAENet’s streamlined architecture, which effectively integrates spectral and spatial attention mechanisms without imposing excessive computational demands. This efficiency makes FAENet suitable for real-time or large-scale remote sensing applications.

In terms of FLOPs, FAENet achieves significant reductions compared to transformer-based models, such as CLCFormer (75.6 G) and HCANet (72.3 G). This reduction, amounting to over 20%, demonstrates FAENet’s ability to deliver high segmentation accuracy while minimizing resource usage. The integration of frequency attention mechanisms enables FAENet to focus computational efforts on the most relevant spectral and spatial features, leading to improved performance at a lower computational cost.

Compared to CNN-based models, FAENet demonstrates slightly higher computational costs than MACUNet (55.1 G) and ResUNet-a (58.2 G) while delivering substantially better segmentation results. The 42.2 ms inference time is competitive, demonstrating that the additional complexity introduced by spectral–spatial attention does not significantly impact processing speed. This balance underscores FAENet’s capability to combine the strengths of CNN and transformer designs.

FAENet’s frequency domain approach enhances computational efficiency by leveraging DWT to decompose features, allowing for the targeted refinement of spectral–spatial representations. This design reduces redundancy and focuses processing power where it is most impactful, resulting in both faster inference and superior accuracy compared to conventional methods.

In summary, FAENet's efficiency analysis reveals that it is both computationally economical and highly effective, setting a new standard for balancing speed, complexity, and accuracy in semantic segmentation of remote sensing images. Its scalability and efficiency make it an excellent choice for diverse applications, including real-time processing and large-scale geographic analysis.

Table 4. Efficiency evaluations. Inference time is calculated as an average time for test set.

Methods	Inference Time (ms)	FLOPs (G)
FCN-8s [30]	35.6	53.6
DANet [36]	47.3	67.4
ResUNet-a [57]	41.8	58.2
MACUNet [38]	39.4	55.1
HCANet [37]	50.1	72.3
SCAttNet [66]	45.7	69.8
A2FPN [39]	44.9	65.2
ICTNet [82]	46.4	70.1
CLCFormer [51]	49.6	75.6
LETFormer [74]	48.3	73.4
Ours (FAENet)	42.2	60.7

4.6. Effects of ICCA and CCCA

Table 5 presents the results of an ablation study assessing the contributions of the ICCA and CCCA modules to the performance of FAENet on the ISPRS Potsdam and LoveDA datasets. The study includes four configurations: FAENet without ICCA and CCCA, FAENet with only ICCA, FAENet with only CCCA, and FAENet with both modules combined. This comprehensive analysis reveals that the combined incorporation of ICCA and CCCA achieves the highest scores across all metrics, emphasizing their complementary roles in refining spectral–spatial representations.

Table 5. Effects of ICCA and CCCA on two benchmarks. Results are in the form of AF/OA/mIoU, where bold text indicates the best.

Networks	ICCA	CCCA	Potsdam	LoveDA
FAENet			82.01/81.34/73.15	66.22/63.45/57.93
FAENet	✓		85.95/84.59/76.97	69.78/67.11/61.07
FAENet		✓	86.67/85.40/77.72	70.46/67.76/61.66
FAENet	✓	✓	**92.71/92.31/83.58**	**76.49/72.93/66.91**

The baseline configuration (FAENet without ICCA and CCCA) achieves AF/OA/mIoU scores of 82.01/81.34/73.15 on the Potsdam dataset and 66.22/63.45/57.93 on the LoveDA dataset. While these results are competitive, they are significantly lower than those achieved by FAENet with either ICCA or CCCA individually, and even more so when both modules are combined.

When only ICCA is incorporated, FAENet achieves AF/OA/mIoU scores of 85.95/84.59/76.97 on the Potsdam dataset and 69.78/67.11/61.07 on LoveDA, highlighting ICCA's ability to capture spectral nuances within individual frequency components. Incorporating only CCCA results in AF/OA/mIoU scores of 86.67/85.40/77.72 on Potsdam and 70.46/67.76/61.66 on LoveDA, demonstrating that CCCA effectively enhances cross-frequency interactions, further improving segmentation performance.

The combined use of ICCA and CCCA achieves the highest scores, illustrating their synergistic effect in enhancing feature representation. ICCA focuses on refining channel-specific spectral information, while CCCA facilitates cross-component interaction, enabling comprehensive spectral–spatial context modeling. Together, these modules improve class-wise segmentation precision, ensure better boundary preservation, and enhance feature discrimination, particularly in complex remote sensing scenarios.

These findings validate the design of the frequency attention mechanism and confirm that the integration of ICCA and CCCA is critical to achieving state-of-the-art performance in semantic segmentation tasks. The consistent improvements across the ISPRS Potsdam and LoveDA datasets further underscore the generalizability of the proposed approach.

5. Conclusions

This study introduces FAENet, a novel frequency attention-enhanced network, specifically designed for the semantic segmentation of HRRSIs. By leveraging the FreqA, FAENet effectively integrates spectral and spatial contexts, addressing the limitations of traditional CNN and transformer-based approaches in capturing fine-grained spectral details. Experimental evaluations on the ISPRS Potsdam and LoveDA datasets demonstrate that FAENet outperforms state-of-the-art methods, achieving superior segmentation accuracy, particularly in complex and heterogeneous scenes. Ablation studies further validate the contributions of the ICCA and CCCA modules, underscoring their complementary roles in enhancing spectral–spatial feature representation.

An important aspect of FAENet is its potential transferability to other datasets or applications beyond the ISPRS Potsdam and LoveDA datasets. Given that FAENet effectively models both spectral and spatial information, it can be applied to other high-resolution remote sensing datasets with similar characteristics, such as urban mapping or land-use classification tasks. Moreover, the frequency attention mechanism is designed to handle diverse spectral variations, making it adaptable to datasets with varying spectral bands, including hyperspectral and multispectral imagery.

Additionally, FAENet's encoder–decoder architecture, combined with frequency attention, positions it as a candidate for broader remote sensing applications, such as object detection, instance segmentation, and even change detection. Future work could explore fine-tuning FAENet on such tasks, potentially leading to enhanced generalization across different remote sensing domains.

In conclusion, FAENet represents a significant advancement in remote sensing semantic segmentation, with its innovative frequency domain approach setting a new benchmark for feature refinement. Future research could extend this framework to incorporate additional modalities, such as hyperspectral and LiDAR data, which provide richer spectral and elevation information, respectively. By leveraging the fine spectral granularity of hyperspectral imagery and the precise elevation details from LiDAR data, FAENet has the potential to further enhance segmentation performance in applications requiring high spatial–spectral discrimination or detailed topographic analysis. Furthermore, future work could explore its application in other remote sensing tasks like object detection and change detection. The promising results of FAENet pave the way for more robust, generalizable, and efficient methods in remote sensing image analysis.

Author Contributions: Conceptualization, J.Z., T.Z., Z.X., C.W., Z.C., X.L. (Xin Li) and X.L. (Xin Lyu); methodology, J.Z., T.Z., Z.X., C.W., S.Q. and N.X.; software, J.Z., T.Z., Z.X., Z.C. and C.W.; validation, J.Z., T.Z., Z.X. and C.W.; formal analysis, J.Z., T.Z., Z.X. and C.W.; investigation, C.W., X.L. (Xin Li) and X.L. (Xin Lyu); resources, C.W., X.L. (Xin Li) and X.L. (Xin Lyu); data curation, C.W., X.L. (Xin Li) and X.L. (Xin Lyu); writing—original draft preparation, J.Z., T.Z., Z.X., C.W., X.L. (Xin Li), S.Q., N.X. and X.L. (Xin Lyu); writing—review and editing, X.L. (Xin Li), S.Q., N.X., Z.C. and X.L. (Xin Lyu);

visualization, J.Z., T.Z., Z.X. and C.W.; supervision, X.L. (Xin Li), S.Q., N.X. and X.L. (Xin Lyu); project administration, X.L. (Xin Li) and X.L. (Xin Lyu); funding acquisition, X.L. (Xin Li) and X.L. (Xin Lyu). All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the National Key Research and Development Program of China under Grant No. 2023YFC3209201 and 2023YFC3209301, National Natural Science Foundation of China under Grant No. 62401196, Natural Science Foundation of Jiangsu Province under Grant No. BK20241508, Fundamental Research Funds for the Central Universities under Grant No. B230201007, B230204009, and B220206006, and Excellent Post-doctoral Program of Jiangsu Province under Grant No. 2022ZB166.

Data Availability Statement: Public datasets were used in this paper. The download links are [https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx], accessed on 2 December 2022, and [https://github.com/Junjue-Wang/LoveDA], accessed on 2 December 2022.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. https://doi.org/10.1109/MGRS.2017.2762307.
2. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968.
3. Ahmad, M.N.; Shao, Z.; Xiao, X.; Fu, P.; Javed, A.; Ara, I. A novel ensemble learning approach to extract urban impervious surface based on machine learning algorithms using SAR and optical data. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *132*, 104013.
4. Li, Y.; Shi, T.; Zhang, Y.; Ma, J. SPGAN-DA: Semantic-Preserved Generative Adversarial Network for Domain Adaptive Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5406717. https://doi.org/10.1109/TGRS.2023.3313883.
5. Xu, N.; Gong, P. Significant coastline changes in China during 1991–2015 tracked by Landsat data. *Sci. Bull.* **2018**, *63*, 883–886.
6. Liu, H.; Shi, Y.; Chang, Q.; Guluzade, R.; Pan, X.; Xu, N.; Hu, P.; Kong, X.; Yang, Y. A New Extraction Method of Surface Water Based on Dense Time-Sequence Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3151–3166.
7. Xu, G.; Zhang, X.; He, X.; Wu, X. Levit-unet: Make faster encoders with transformer for medical image segmentation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Xiamen, China, 13–15 October 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 42–53.
8. Asgari Taghanaki, S.; Abhishek, K.; Cohen, J.P.; Cohen-Adad, J.; Hamarneh, G. Deep semantic segmentation of natural and medical images: A review. *Artif. Intell. Rev.* **2021**, *54*, 137–178.
9. Yuan, F.; Zhang, Z.; Fang, Z. An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognit.* **2023**, *136*, 109228.
10. Heidari, M.; Kazerouni, A.; Soltany, M.; Azad, R.; Aghdam, E.K.; Cohen-Adad, J.; Merhof, D. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 6202–6212.
11. Yao, S.; Guan, R.; Huang, X.; Li, Z.; Sha, X.; Yue, Y.; Lim, E.G.; Seo, H.; Man, K.L.; Zhu, X.; et al. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Trans. Intell. Veh.* **2023**, *9*, 2094–2128.
12. Xiao, X.; Zhao, Y.; Zhang, F.; Luo, B.; Yu, L.; Chen, B.; Yang, C. BASeg: Boundary aware semantic segmentation for autonomous driving. *Neural Netw.* **2023**, *157*, 460–470.
13. Rossolini, G.; Nesti, F.; D'Amico, G.; Nair, S.; Biondi, A.; Buttazzo, G. On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 18328–18342.
14. Yang, L.; Bai, Y.; Ren, F.; Bi, C.; Zhang, R. Lcfnets: Compensation strategy for real-time semantic segmentation of autonomous driving. *IEEE Trans. Intell. Veh.* **2024**, *9*, 4715–4729.
15. Qiu, L.; Yu, D.; Zhang, C.; Zhang, X. A semantics-geometry framework for road extraction from remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6004805.
16. Chen, X.; Yu, A.; Sun, Q.; Guo, W.; Xu, Q.; Wen, B. Updating Road Maps at City Scale With Remote Sensed Images and Existing Vector Maps. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5616521. https://doi.org/10.1109/TGRS.2024.3375807.
17. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78.
18. Rasti, B.; Hong, D.; Hang, R.; Ghamisi, P.; Kang, X.; Chanussot, J.; Benediktsson, J.A. Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 60–88.

19. Jamali, A.; Roy, S.K.; Hong, D.; Atkinson, P.M.; Ghamisi, P. Spatial Gated Multi-Layer Perceptron for Land Use and Land Cover Mapping. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5502105.
20. Shao, Z.; Cheng, T.; Fu, H.; Li, D.; Huang, X. Emerging issues in mapping urban impervious surfaces using high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 2562.
21. Ahmad, M.N.; Shao, Z.; Javed, A. Modelling land use/land cover (LULC) change dynamics, future prospects, and its environmental impacts based on geospatial data models and remote sensing data. *Environ. Sci. Pollut. Res.* **2023**, *30*, 32985–33001.
22. Tong, X.Y.; Xia, G.S.; Zhu, X.X. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 178–196.
23. Zhou, X.; Xie, X.; Huang, H.; Shao, Z.; Huang, X. WodNet: Weak Object Discrimination Network for Cloud Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5627020. https://doi.org/10.1109/TGRS.2024.3406542.
24. Lizarazo, I. SVM-based segmentation and classification of remotely sensed data. *Int. J. Remote Sens.* **2008**, *29*, 7277–7283.
25. Zhou, C.; Liang, D.; Yang, X.; Yang, H.; Yue, J.; Yang, G. Wheat ears counting in field conditions based on multi-feature optimization and TWSVM. *Front. Plant Sci.* **2018**, *9*, 1024.
26. Wang, D.; Liu, S.; Zhang, C.; Xu, M.; Yang, J.; Yasir, M.; Wan, J. An improved semantic segmentation model based on SVM for marine oil spill detection using SAR image. *Mar. Pollut. Bull.* **2023**, *192*, 114981.
27. Hmimid, A.; Sayyouri, M.; Qjidaa, H. Image classification using a new set of separable two-dimensional discrete orthogonal invariant moments. *J. Electron. Imaging* **2014**, *23*, 013026.
28. Karmouni, H.; Jahid, T.; Hmimid, A.; Sayyouri, M.; Qjidaa, H. Fast computation of inverse Meixner moments transform using Clenshaw's formula. *Multimed. Tools Appl.* **2019**, *78*, 31245–31265.
29. Jahid, T.; Karmouni, H.; Hmimid, A.; Sayyouri, M.; Qjidaa, H. Image moments and reconstruction by Krawtchouk via Clenshaw's reccurence formula. In Proceedings of the 2017 International Conference on Electrical and Information Technologies (ICEIT), Rabat, Morocco, 15–18 November 2017; pp. 1–7.
30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.
31. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
32. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
33. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867.
34. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
35. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
36. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149. https://doi.org/10.1109/CVPR.2019.00326.
37. Li, X.; Xu, F.; Xia, R.; Lyu, X.; Gao, H.; Tong, Y. Hybridizing Cross-Level Contextual and Attentive Representations for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2021**, *13*, 2986.
38. Li, R.; Duan, C.; Zheng, S.; Zhang, C.; Atkinson, P.M. MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8007205. https://doi.org/10.1109/LGRS.2021.3052886.
39. Li, R.; Wang, L.; Zhang, C.; Duan, C.; Zheng, S. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. *Int. J. Remote Sens.* **2022**, *43*, 1131–1155. https://doi.org/10.1080/01431161.2022.2030071.
40. Li, X.; Xu, F.; Liu, F.; Lyu, X.; Tong, Y.; Xu, Z.; Zhou, J. A Synergistical Attention Model for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5400916. https://doi.org/10.1109/TGRS.2023.3243954.
41. Tahiri, M.A.; Karmouni, H.; Bencherqui, A.; Daoui, A.; Sayyouri, M.; Qjidaa, H.; Hosny, K.M. New color image encryption using hybrid optimization algorithm and Krawtchouk fractional transformations. *Vis. Comput.* **2023**, *39*, 6395–6420.
42. Jahid, T.; Hmimid, A.; Karmouni, H.; Sayyouri, M.; Qjidaa, H.; Rezzouk, A. Image analysis by Meixner moments and a digital filter. *Multimed. Tools Appl.* **2018**, *77*, 19811–19831.
43. Daoui, A.; Karmouni, H.; Sayyouri, M.; Qjidaa, H. Efficient methods for signal processing using Charlier moments and artificial bee Colony algorithm. *Circuits Syst. Signal Process.* **2022**, *41*, 166–195.

44. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; 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, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
45. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 7262–7272.
46. Blaga, B.C.Z.; Nedevschi, S. Semantic Segmentation of Remote Sensing Images With Transformer-Based U-Net and Guided Focal-Axial Attention. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 18303–18318. https://doi.org/10.1109/JSTARS.2024.3470316.
47. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214.
48. Lin, R.; Zhang, Y.; Zhu, X.; Chen, X. Local-Global Feature Capture and Boundary Information Refinement Swin Transformer Segmentor for Remote Sensing Images. *IEEE Access* **2024**, *12*, 6088–6099. https://doi.org/10.1109/ACCESS.2024.3350645.
49. Li, X.; Xu, F.; Liu, F.; Tong, Y.; Lyu, X.; Zhou, J. Semantic Segmentation of Remote Sensing Images by Interactive Representation Refinement and Geometric Prior-Guided Inference. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3339291.
50. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. https://doi.org/10.1109/TGRS.2022.3144165.
51. Long, J.; Li, M.; Wang, X. Integrating Spatial Details With Long-Range Contexts for Semantic Segmentation of Very High-Resolution Remote-Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2501605. https://doi.org/10.1109/LGRS.2023.3262586.
52. Ma, R.; Zhang, Y.; Zhang, B.; Fang, L.; Huang, D.; Qi, L. Learning Attention in the Frequency Domain for Flexible Real Photograph Denoising. *IEEE Trans. Image Process.* **2024**, *33*, 3707–3721.
53. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
54. Xu, Z.Q.J.; Zhang, Y.; Luo, T. Overview frequency principle/spectral bias in deep learning. *arXiv* **2022**, arXiv:2201.07395.
55. International Society for Photogrammetry and Remote Sensing. ISPRS 2D Semantic Labeling Contest—Potsdam. Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx (accessed on 20 October 2021).
56. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
57. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114.
58. Zhou, H.; Zhang, Y.; Wu, J.; Wang, C. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
59. Sun, Y.; Zheng, W. HRNet-and PSPNet-based multiband semantic segmentation of remote sensing images. *Neural Comput. Appl.* **2023**, *35*, 8667–8675.
60. Du, S.; Du, S.; Liu, B.; Zhang, X. Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *Int. J. Digit. Earth* **2021**, *14*, 357–378.
61. Yang, Z.; Wu, Q.; Zhang, F.; Zhang, X.; Chen, X.; Gao, Y. A New Semantic Segmentation Method for Remote Sensing Images Integrating Coordinate Attention and SPD-Conv. *Symmetry* **2023**, *15*, 1037.
62. Wang, W.; Tan, X.; Zhang, P.; Wang, X. A CBAM based multiscale transformer fusion approach for remote sensing image change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 6817–6825.
63. Li, X.; Xu, F.; Yu, A.; Gao, H.; Zhou, J. A Frequency Decoupling Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 2025, *early access*.
64. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 426–435.
65. Li, X.; Xu, F.; Lyu, X.; Gao, H.; Tong, Y.; Cai, S.; Li, S.; Liu, D. Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images. *Int. J. Remote Sens.* **2021**, *42*, 3583–3610.
66. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic Segmentation Network With Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 905–909. https://doi.org/10.1109/LGRS.2020.2988294.
67. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5603018. https://doi.org/10.1109/TGRS.2021.3065112.
68. Zuo, R.; Zhang, G.; Zhang, R.; Jia, X. A Deformable Attention Network for High-Resolution Remote Sensing Images Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4406314. https://doi.org/10.1109/TGRS.2021.3119537.

69. Yi, S.; Li, J.; Liu, X.; Yuan, X. CCAFFMNet: Dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module. *Neurocomputing* **2022**, *482*, 236–251.
70. Ding, L.; Lin, D.; Lin, S.; Zhang, J.; Cui, X.; Wang, Y.; Tang, H.; Bruzzone, L. Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13.
71. Liu, Y.; Zhang, Y.; Wang, Y.; Mei, S. Rethinking Transformers for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5617515. https://doi.org/10.1109/TGRS.2023.3302024.
72. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. https://doi.org/10.1109/LGRS.2022.3143368.
73. Xiao, T.; Liu, Y.; Huang, Y.; Li, M.; Yang, G. Enhancing Multiscale Representations With Transformer for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605116. https://doi.org/10.1109/TGRS.2023.3256064.
74. Li, X.; Xu, F.; Xia, R.; Xu, N.; Liu, F.; Yuan, C.; Huang, Q.; Lyu, X. Locality-Enhanced Transformer for Semantic Segmentation of High-Resolution Remote Sensing Images. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 2870–2874. https://doi.org/10.1109/ICASSP48485.2024.10446525.
75. Azad, R.; Kazerouni, A.; Sulaiman, A.; Bozorgpour, A.; Aghdam, E.K.; Jose, A.; Merhof, D. Unlocking Fine-Grained Details with Wavelet-Based High-Frequency Enhancement in Transformers. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Vancouver, BC, Canada, 8–12 October 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 207–216.
76. Ehrlich, M.; Davis, L.S. Deep residual learning in the jpeg transform domain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3484–3493.
77. Zhang, F.; Panahi, A.; Gao, G. FsaNet: Frequency self-attention for semantic segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 4757–4772.
78. Su, B.; Liu, J.; Su, X.; Luo, B.; Wang, Q. CFCANet: A complete frequency channel attention network for sar image scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11750–11763.
79. Li, X.; Xu, F.; Yong, X.; Chen, D.; Xia, R.; Ye, B.; Gao, H.; Chen, Z.; Lyu, X. SSCNet: A Spectrum-Space Collaborative Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5610.
80. Zhou, M.; Huang, J.; Yan, K.; Hong, D.; Jia, X.; Chanussot, J.; Li, C. A General Spatial-Frequency Learning Framework for Multimodal Image Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**.
81. Song, B.; Min, S.; Yang, H.; Wu, Y.; Wang, B. A Fourier frequency domain convolutional neural network for remote sensing crop classification considering global consistency and edge specificity. *Remote Sens.* **2023**, *15*, 4788.
82. Li, X.; Xu, F.; Xia, R.; Li, T.; Chen, Z.; Wang, X.; Xu, Z.; Lyu, X. Encoding Contextual Information by Interlacing Transformer and Convolution for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2022**, *14*, 4065.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.