# FSECrossNet: Frequency–Spatial Enhancement with Cross-Modal Attention for Remote Sensing Semantic Segmentation

Yingqi Liang
*Guangxi University*
Nanning, China
2413301020@st.gxu.edu.cn

***Abstract*—Multi-modal remote sensing semantic segmentation benefits from complementary cues such as RGB appearance and DSM elevation, yet effectively integrating fine boundary details with global context remains challenging. Moreover, many fusion pipelines rely solely on spatial-domain processing and employ heavy cross-modal attention, leading to limited efficiency. We propose FSECrossNet, a Transformer-style encoder–decoder that performs staged fusion by coupling frequency–spatial enhancement with lightweight cross-modal reasoning. Specifically, Gated Feature Fusion (GFF) builds balanced multi-scale skip features by adaptively weighting RGB and DSM cues. At the deepest stage, Dual Frequency–Spatial Fusion (DFSF) strengthens discriminative textures via DCT-based frequency enhancement and improves local structures using multi-scale depthwise convolutions. Finally, Cross-Modal Attention Fusion (CMAF) reuses shared projections across self- and cross-attention to enable efficient intra-/inter-modal interaction with reduced parameters. Experiments on ISPRS Vaihingen and Potsdam demonstrate consistent improvements over strong baselines, yielding higher segmentation accuracy and clearer boundary delineation in qualitative comparisons under a favorable accuracy–efficiency trade-off.**

***Index Terms*—Remote sensing, multi-modal fusion, frequency-spatial fusion, attention mechanism, image segmentation**

## I. Introduction

Multi-modal remote sensing semantic segmentation has become a fundamental task in Earth observation applications. In particular, RGB imagery provides rich appearance, texture, and color cues, while Digital Surface Model (DSM) data complements it with elevation structures that help disambiguate visually similar classes (e.g., buildings vs. roads, or low vegetation vs. impervious surfaces). However, effectively leveraging both global context and fine boundary details in a unified RGB–DSM segmentation pipeline remains challenging.

Recent progress in remote sensing semantic segmentation mainly falls into three directions. First, Transformer-style segmentation backbones have shown strong capability in modeling long-range context for urban-scene imagery [1], [2]. Second, RGB–DSM/LiDAR fusion networks explore cross-modal interaction to exchange complementary appearance and elevation cues [3], [4]. Third, frequency-aware designs emphasize textures and edges via frequency transforms such as DCT [5]. Despite these advances, jointly integrating frequency cues, spatial details, and efficient cross-modal interaction under a lightweight design remains non-trivial.

However, three issues still limit current RGB–DSM segmentation methods. First, frequency cues are often injected as an auxiliary branch and merged with spatial features through simple concatenation or addition [5], overlooking their complementary roles: frequency components are highly discriminative for textures and edges, whereas spatial processing is crucial for precise localization, small objects, and boundary delineation. Second, many fusion pipelines either fuse modalities too early (before modality-specific representations are sufficiently strengthened) or rely on heavy cross-modal attention stacks [3], [6], which increases computation and may yield unstable gains. Third, adaptive control for frequency–spatial fusion and multi-modal integration is often insufficient, making it difficult to balance global semantics and local details and potentially causing over-smoothed boundaries or under-utilized DSM cues in challenging regions.

To address these challenges, we propose FSECrossNet, a Transformer-style encoder–decoder for RGB–DSM semantic segmentation. Rather than treating frequency enhancement, spatial refinement, and cross-modal interaction as components, FSECrossNet follows a staged fusion strategy: it first strengthens modality-specific representations in frequency and spatial domains, then performs lightweight cross-modal reasoning. This design delays cross-modal interaction until discriminative structures and boundary-sensitive details are refined, improving fusion quality while keeping computation affordable. Our contributions are summarized as follows:

1) We present a staged RGB–DSM fusion strategy that couples frequency–spatial enhancement with cross-modal reasoning, enabling interaction after discriminative and boundary-sensitive cues are strengthened.
2) We propose a Dual Frequency–Spatial Fusion (DFSF) module that couples DCT-based frequency enhancement and multi-scale spatial refinement to improve texture discrimination and fine-structure preservation; extensive ablations (including DCT vs. FFT) support this design.
3) We develop a parameter-efficient Cross-Modal Attention Fusion (CMAF) module that reuses projection matrices across self-attention and cross-attention, reducing atten-

This work was supported in part by the National Natural Science Foundation of China, No. 62062007, and in part by the Guangxi Natural Science Foundation, No. 2025GXNSFAA069236.

tion overhead while maintaining effective intra-/inter-modal feature interaction.

## II. Methodology

Most existing multimodal architectures combine spatial convolutions, frequency transforms, and cross-modal attention without a unified staged design, often applying heavy Transformer blocks directly on raw encoder features. In contrast, FSECrossNet is built on the hypothesis that frequency and spatial cues should be enhanced *before* intensive cross-modal reasoning, so that each modality provides a cleaner and more discriminative basis for interaction. Concretely, DFSF sharpens class-discriminative patterns in the frequency domain and restores fine-grained details in the spatial domain at the deepest encoder stage, and CMAF then performs lightweight self- and cross-attention on these refined representations. This staged design yields a more stable and efficient fusion process than directly stacking deep Transformer layers on unrefined multi-modal features.

As shown in Fig. 1, FSECrossNet consists of a dual-branch encoder, GFF, DFSF, and CMAF modules, and a decoder. Given the RGB image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ and DSM data $\mathbf{Y} \in \mathbb{R}^{H\times W\times 1}$ as inputs, the overall pipeline proceeds as follows. We adopt ResNet-50 [7] as the encoders. The encoder features are adaptively fused through GFF at multiple scales for skip connections, which balances the contributions from RGB and DSM modalities. At the deepest scale, the RGB feature and the DSM feature are separately enhanced by DFSF modules to improve detail feature representation and multi-scale feature extraction capabilities. The enhanced features are then processed through CMAF modules to enable efficient intra-modal and inter-modal feature learning. The final segmentation map is generated by a U-Net style decoder [8] with skip connections.

### A. Gated Feature Fusion

The Gated Feature Fusion (GFF) mechanism adaptively combines encoder features from RGB and DSM modalities by aggregating global contextual information to generate learnable gating weights.

Specifically, the GFF module first concatenates the RGB and DSM features $\mathbf{F}_x$ and $\mathbf{F}_y$, then applies global pooling to summarize modality-wise responses and generate gating weights through lightweight transformations and Sigmoid activation. The final gating mechanism can be formulated as:

$$\mathbf{F}_{gff} = \alpha \odot \mathbf{F}_x + (1-\alpha) \odot \mathbf{F}_y \tag{1}$$

where $\alpha$ is the learnable gating weight generated by the dual-branch architecture, and $\odot$ denotes element-wise multiplication. This design allows the network to adaptively balance the contributions from RGB and DSM modalities based on their distinct characteristics.

### B. Dual Frequency-Spatial Fusion Module

The DFSF module implements an architecture following the principle of frequency domain enhancement for discrimination and spatial enhancement for details.

**Frequency Enhancement**: The frequency domain branch applies Discrete Cosine Transform (DCT) [5] to project features into the frequency domain, selectively enhances high-frequency components, and transforms back to the spatial domain. At the deepest scale, DFSF takes as inputs the fourth-layer DSM encoder feature $\mathbf{F}_y^{(4)}$ and the fourth-layer GFF output $\mathbf{F}_{gff}^{(4)}$, which we denote generically as $\mathbf{F}$. Each input is processed through a frequency-domain projection, high-frequency enhancement, and inverse transformation:

$$\mathbf{F}_{freq} = \mathcal{T}_{\text{DCT}}^{-1}(\mathcal{H}(\mathcal{T}_{\text{DCT}}(\mathbf{F}))) \tag{2}$$

where $\mathcal{T}_{\text{DCT}}(\cdot)$ denotes the forward DCT that projects spatial features into the frequency domain, $\mathcal{H}(\cdot)$ is a high-frequency enhancement operator that re-weights informative frequency components to emphasize textures and edges, and $\mathcal{T}_{\text{DCT}}^{-1}(\cdot)$ denotes the inverse DCT that maps the enhanced frequency responses back to the spatial domain, producing a feature map with the same size as the input.

**Spatial Enhancement**: The spatial branch employs a Multi-Scale Convolution (MSConv) module to enhance multi-scale feature extraction capability, which is crucial for capturing local details at different spatial scales. For the same $\mathbf{F}$, the module preprocesses input features through layer normalization and convolutions, then applies three parallel depthwise separable convolutions [9] with kernel sizes of $3\times 3$, $5\times 5$, and $7\times 7$ to extract features at multiple scales. The multi-scale outputs are fused and combined with the input through a residual connection:

$$\mathbf{F}_{spat} = \sum_{k\in\{3,5,7\}} \text{DWConv}_{k\times k}(\mathbf{F}_{pre}) + \mathbf{F} \tag{3}$$

where $\mathbf{F}_{pre} = \text{Conv}_{3\times 3}(\text{Conv}_{1\times 1}(\text{LayerNorm}(\mathbf{F})))$ denotes the preprocessed features, $\text{DWConv}_{k\times k}$ denotes depthwise separable convolution with kernel size $k\times k$, and the summation represents element-wise addition of multi-scale features.

**Adaptive Fusion**: For each modality, the outputs from both branches are adaptively combined through a gating mechanism:

$$\mathbf{F}_{dfsf} = \lambda_f \cdot \mathbf{F}_{freq} + \lambda_s \cdot \mathbf{F}_{spat}. \tag{4}$$

where $\lambda_f$ and $\lambda_s$ are learnable weights generated by an adaptive gate, and $\mathbf{F}_{dfsf}$ denotes the DFSF-enhanced feature for each modality. In our implementation, $\lambda_f, \lambda_s \in \mathbb{R}^{B\times C\times 1\times 1}$ are channel-wise gates predicted from globally pooled features.

### C. Cross-Modal Attention Fusion

The Cross-Modal Attention Fusion (CMAF) module processes enhanced features from the DFSF modules. Unlike conventional approaches that stack multiple deep Transformer layers, our CMAF module combines self-attention and cross-attention in a single parameter-friendly module, where attention operations share projection matrices. This design significantly reduces computational overhead while maintaining effective multi-modal feature learning capabilities.

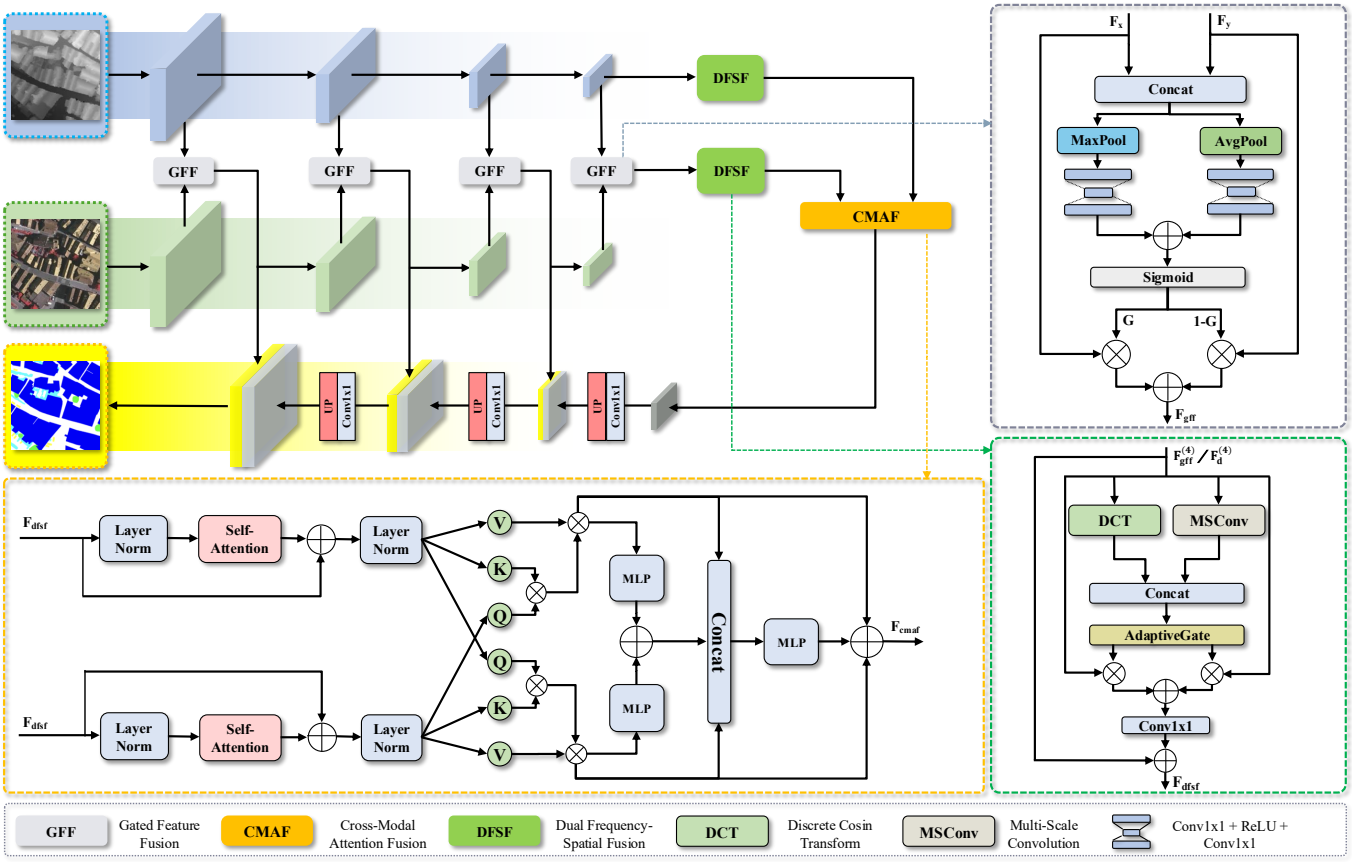


Fig. 1. The overall architecture of the proposed FSECrossNet. The network follows an encoder-decoder structure with dual-branch encoder, Gated Feature Fusion (GFF), Dual Frequency–Spatial Fusion (DFSF), and Cross-Modal Attention Fusion (CMAF) modules.

**Self- and Cross-Attention**: CMAF receives the DFSF-enhanced features from the previous stage. Specifically, the final-layer DSM encoder feature $\mathbf{F}_y^{(4)}$ and the last-layer GFF output $\mathbf{F}_{gff}^{(4)}$ are first processed by DFSF modules to obtain frequency–spatial enhanced features:

$$\widehat{\mathbf{F}}_r = \mathrm{DFSF}(\mathbf{F}_{gff}^{(4)}), \quad \widehat{\mathbf{F}}_d = \mathrm{DFSF}(\mathbf{F}_y^{(4)}). \tag{5}$$

Here, the subscripts $r$ and $d$ denote the RGB-side (from $\mathbf{F}_{gff}^{(4)}$) and DSM-side (from $\mathbf{F}_y^{(4)}$) features, respectively. We reshape each feature map into a token sequence $\mathbf{Z} \in \mathbb{R}^{B\times N\times d}$ with $N = HW$ (and $d = C$) before attention, and fold it back after attention.

Using shared projection matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$, we form $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K} = \mathbf{Z}\mathbf{W}_k$, and $\mathbf{V} = \mathbf{Z}\mathbf{W}_v$ for each modality, and apply scaled dot-product attention:

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{6}$$

The self-attention outputs are computed as:

$$\mathrm{SA}_r = \mathcal{A}(\mathbf{Q}_r,\mathbf{K}_r,\mathbf{V}_r), \quad \mathrm{SA}_d = \mathcal{A}(\mathbf{Q}_d,\mathbf{K}_d,\mathbf{V}_d). \tag{7}$$

The self-attention features $\mathrm{SA}_r$ and $\mathrm{SA}_d$ then act as the inputs of the cross-attention stage. Reusing the same projection matrices yields:

$$\begin{aligned} \mathrm{CA}_r &= \mathcal{A}(\mathbf{Q}_r^{\mathrm{ca}},\mathbf{K}_d^{\mathrm{ca}},\mathbf{V}_d^{\mathrm{ca}}), \\ \mathrm{CA}_d &= \mathcal{A}(\mathbf{Q}_d^{\mathrm{ca}},\mathbf{K}_r^{\mathrm{ca}},\mathbf{V}_r^{\mathrm{ca}}), \end{aligned} \tag{8}$$

where $\mathbf{Q}_r^{\mathrm{ca}} = \mathrm{SA}_r\mathbf{W}_q$, $\mathbf{K}_d^{\mathrm{ca}} = \mathrm{SA}_d\mathbf{W}_k$, $\mathbf{V}_d^{\mathrm{ca}} = \mathrm{SA}_d\mathbf{W}_v$ (and symmetrically for $\mathrm{CA}_d$).

The modality-specific features are then updated through feature enhancement, fusion layers, and residual connections to obtain $\mathbf{F}'_r$ and $\mathbf{F}'_d$. The CMAF block produces the final fused representation through residual aggregation:

$$\mathbf{F}_{cmaf} = \mathbf{F}'_r + \mathbf{F}'_d. \tag{9}$$

### *D. Loss Function*

We use a weighted cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c}\log(\hat{y}_{i,c}) \tag{10}$$

where $N$ is the number of pixels, $C$ is the number of classes, $w_c$ is the class weight computed based on the inverse frequency of each class in the training set, $y_{i,c}$ is the ground truth label, and $\hat{y}_{i,c}$ is the predicted probability.

## III. Experiments

### A. Experimental Setup

Unless otherwise stated, all models are trained and evaluated using $256 \times 256$ patches cropped from the original large images, with batch size 10 for training. We use pure FP32 (no AMP). Data augmentation includes random horizontal and vertical flips applied synchronously to RGB, DSM, and the label map to preserve cross-modal alignment.

*1) Datasets:* We evaluate FSECrossNet on the ISPRS 2D semantic labeling benchmarks, **Vaihingen** and **Potsdam**, which provide paired RGB images and DSM. The GSD is approximately 9 cm for Vaihingen and 5 cm for Potsdam. Following the common 5-class protocol, we report results on impervious surfaces, buildings, low vegetation, trees, and cars, excluding clutter/background. We adopt the train/test split used in prior work. For Vaihingen, tiles {1, 3, 7, 11, 13, 17, 23, 26, 28, 32, 34, 37} are used for training and {5, 15, 21, 30} for testing. For Potsdam, the training set contains orthophotos indexed by {6_10, 7_10, 2_12, 3_11, 2_10, 7_8, 5_10, 3_12, 5_12, 7_11, 7_9, 6_9, 7_7, 4_12, 6_8, 6_12, 6_7, 4_11}, whereas the test set contains {2_11, 3_10, 4_10, 5_11, 6_11, 7_12}.

*2) Baselines:* We compare FSECrossNet with representative recent baselines including both single-modal and multi-modal approaches. For single-modal methods, we include TransUNet [6], UNetFormer [1], and LoGCAN++ [2]. For multi-modal methods, we include CMGFNet [4], FTransUNet [10], MCFNet [11], and MGFNet [12]. All baselines are re-implemented and evaluated under the same split and metrics.

*3) Criteria:* We report mean Intersection over Union (mIoU), mean F1-score (mF1), and overall accuracy (OA) across the five major classes, as well as floating point operations (FLOPs) and parameters for efficiency evaluation.

### B. Implementation Details

Our FSECrossNet is implemented in PyTorch with ResNet-50 dual encoders (RGB/DSM) [7]. The DFSF module uses DCT and MSConv modules [5]. CMAF adopts shared projections and gating to reduce compute. The decoder follows a U-Net style with skip connections [8].

For data preprocessing, images are tiled into patches of size $256 \times 256$. During training, patches are randomly sampled from the full images; during inference, we use a sliding window with stride 32 and batch size 1. Overlapping regions are fused by accumulating logits from multiple predictions, then applying argmax to obtain the final segmentation. RGB channels are normalized to [0, 1] by dividing by 255; DSM is min–max normalized per tile to reduce scale variance.

Models are trained with SGD (momentum 0.9, weight decay $5e-4$), an initial learning rate of 0.001, batch size 10, and a MultiStepLR scheduler with milestones at epochs 50, 70, and 90 (gamma 0.1). The loss function is weighted cross-entropy loss with class weights computed based on inverse class frequency. Training runs for 100 epochs on a single NVIDIA RTX 3090 GPU.

### C. Result Analysis

Tables I and II show the quantitative comparison results on the Vaihingen and Potsdam datasets, respectively. FSECrossNet (RGB-D) surpasses the best RGB baseline on both datasets, verifying the benefit of DSM. Within multi-modal methods, FSECrossNet is consistently competitive or better than CMGFNet, MCFNet, MGFNet, and FTransUNet. On Vaihingen, our method achieves the highest mIoU of 83.34% and mF1 of 90.69% with 27.96M parameters compared to 179.42M for FTransUNet. On Potsdam, FSECrossNet attains higher OA/mIoU/mF1 (91.43%/86.22%/92.43%) than FTransUNet with markedly fewer parameters, and notably improves impervious surfaces, low vegetation, and cars, reflecting stronger boundary detail preservation.

Figure 2 presents visual comparisons of segmentation results on both test sets, demonstrating that our method achieves better segmentation quality in complex scenarios with multiple object classes and challenging boundaries. In the Vaihingen examples (rows 1–2), FSECrossNet more accurately delineates cars and tree canopies, reducing confusion between adjacent objects compared with other methods. In the Potsdam cases (rows 3–4), our model produces cleaner and more compact building regions, while competing approaches exhibit fragmented or noisy building predictions.

**Inference Speed Comparison**. Table III compares the inference speed of different methods on the Vaihingen dataset. Inference time is measured with batch size 1 on $256 \times 256$ patches, averaged over 50 runs after warmup. FSECrossNet achieves a favorable accuracy–efficiency trade-off compared to other multi-modal methods, with significantly fewer parameters and FLOPs while maintaining high segmentation accuracy.

### D. Ablation Studies

We conduct comprehensive ablation studies on the Vaihingen dataset to validate the effectiveness of each component in our FSECrossNet architecture. The full model achieves 83.34% mIoU; ablations report the performance drop when removing or modifying each component.

**Effects of Key Modules**. Table IV shows that removing GFF, DFSF, or CMAF decreases mIoU by 2.13%, 2.41%, and 1.96%, respectively, underscoring that their collaborative pipeline—GFF aligning modalities, DFSF enriching them with dual-branch cues, and CMAF exchanging complemented context—delivers the best synergy.

**Effects of DFSF Components**. Table V shows the internal ablation results for the DFSF module. The results demonstrate that removing the frequency branch or the spatial branch causes mIoU drops of 1.66% and 1.46%, respectively, and their combination achieves the best performance.

**DCT vs FFT Comparison**. To justify the choice of frequency transform in the DFSF module, we compare Discrete Cosine Transform (DCT) with Fast Fourier Transform (FFT) in the frequency enhancement branch. Together with the "w/o Freq Branch" setting in Table V, this forms a hierarchy from no frequency modeling, to FFT using magnitude spectra, and finally to our DCT-based design. As reported in Table VI,

TABLE I
QUANTITATIVE COMPARISON RESULTS ON THE VAIHINGEN DATASET.

| Method | Modal | FLOPs (G) | Params (M) | OA | Impervious surfaces | Buildings | Low vegetation | Trees | Cars | mF1 | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| UNetFormer | RGB | 5.97 | 11.69 | 89.90 | 91.26 | 92.59 | 77.83 | 90.94 | 74.76 | 86.59 | 76.86 |
| TransUNet | RGB | 64.55 | 93.23 | 89.40 | 92.44 | 91.45 | 79.02 | 90.60 | 74.14 | 86.82 | 77.15 |
| LoGCAN++ | RGB | 18.23 | 24.30 | 89.31 | 89.31 | 94.42 | 75.13 | 92.32 | 74.25 | 86.13 | 76.21 |
| CMGFNet | RGB-D | 77.71 | 85.22 | 91.36 | 91.89 | 97.64 | 78.36 | 91.23 | 86.19 | 89.56 | 81.27 |
| FTransUNet | RGB-D | 112.22 | 179.42 | 92.18 | 92.74 | 98.18 | 81.95 | 91.24 | 86.34 | 89.78 | 82.14 |
| MCFNet | RGB-D | 55.72 | 49.69 | 90.21 | 90.89 | 95.98 | 76.66 | 91.94 | 72.32 | 87.24 | 78.01 |
| MGFNet | RGB-D | 81.26 | 82.91 | 91.48 | 91.64 | 97.21 | 79.99 | 92.01 | 87.63 | 89.88 | 82.06 |
| FSECrossNet (Ours) | RGB-D | **33.32** | **27.96** | **92.30** | **92.88** | 97.77 | **82.82** | 91.44 | **88.29** | **90.69** | **83.34** |

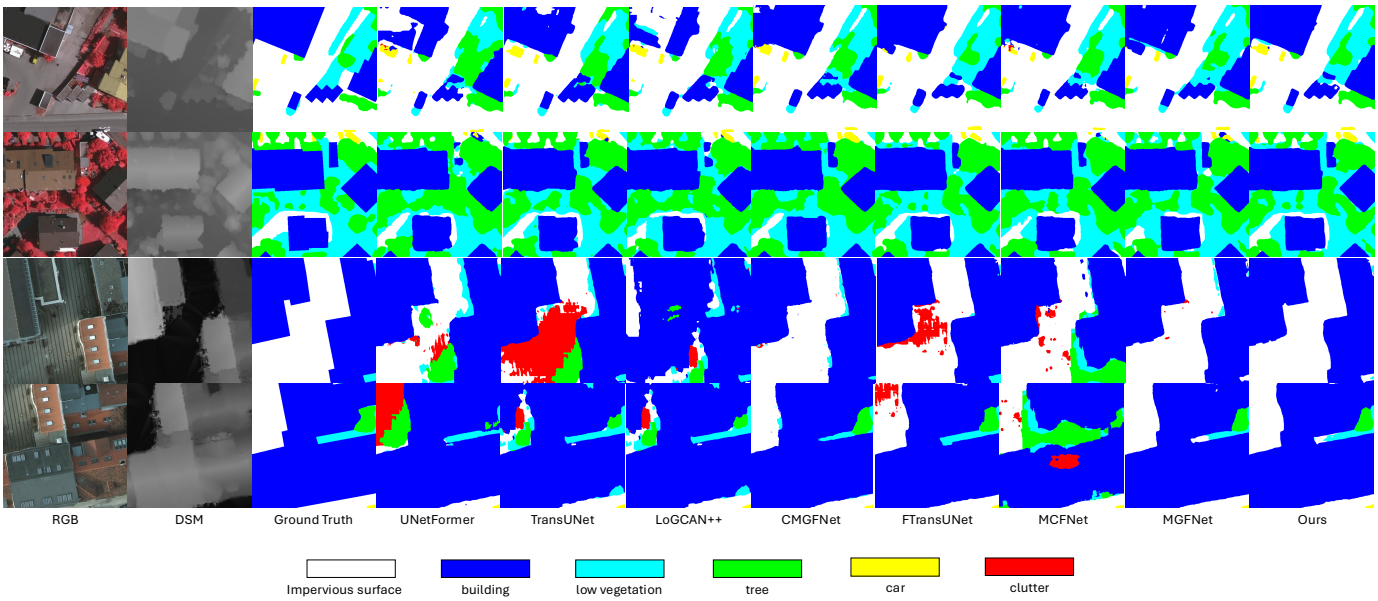


Fig. 2. Qualitative performance comparisons on the Vaihingen (rows 1-2) and Potsdam (rows 3-4) test sets. Each row shows RGB images, DSM data, ground truth, and segmentation results from different methods (UNetFormer, TransUNet, LoGCAN++, CMGFNet, FTransUNet, MCFNet, MGFNet, and our proposed FSECrossNet). All images are cropped to 512×512 pixels for better visualization of segmentation details.

replacing DCT with FFT consistently degrades both mIoU and mF1, indicating that DCT is a more suitable frequency representation for our framework. One plausible reason is that DCT tends to concentrate most of the important information into a small number of coefficients while still preserving edges, which makes it easier for the network to focus on class-discriminative structures in remote sensing imagery.

**Effects of CMAF Components**. Table VII shows the ablation results for the CMAF module. Removing cross-attention or self-attention causes mIoU drops of 1.30% and 1.97%, respectively, and their joint use yields the best accuracy, indicating that aligning intra-modal context before exchanging cross-modal cues helps CMAF deliver more reliable interactions.

**Effects of Kernel Size Combinations**. Table VIII compares different kernel patterns in MSConv. The proposed multi-scale square design achieves the best mIoU, consistently outperforming reduced-kernel and alternative patterns. This indicates that moderate multi-scale receptive fields better balance boundary detail preservation and contextual aggregation.

## IV. CONCLUSION

In this paper, we introduced FSECrossNet for multi-modal remote sensing semantic segmentation, whose main contribution lies in a staged fusion pipeline that couples frequency–spatial enhancement with cross-modal attention rather than in a single standalone block. GFF first builds balanced multi-scale skip features by adaptively weighting RGB and DSM cues. DFSF delivers frequency-aware discrimination and spatial detail restoration at the deepest encoder stage, while CMAF performs lightweight intra/inter-modal reasoning with shared projections on these refined representations. Extensive experiments on Vaihingen and Potsdam verify that this combination yields superior mIoU with markedly lower computation than recent CNN or Transformer counterparts. Future work will explore further improvements to enhance model performance and extend the framework to more diverse remote sensing scenarios.

TABLE II
QUANTITATIVE COMPARISON RESULTS ON THE POTSDAM DATASET.

| Method | Modal | OA | Impervious surfaces | Buildings | Low vegetation | Trees | Cars | mF1 | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| UNetFormer | RGB | 86.00 | 90.57 | 97.34 | 83.21 | 72.04 | 88.28 | 86.11 | 76.26 |
| TransUNet | RGB | 88.47 | 91.85 | 96.93 | 82.06 | 85.52 | 92.70 | 89.16 | 80.95 |
| LoGCAN++ | RGB | 82.89 | 86.89 | 91.98 | 84.97 | 68.38 | 91.59 | 77.79 | 66.22 |
| CMGFNet | RGB-D | 90.24 | 92.32 | 97.47 | 88.35 | 86.28 | 84.67 | 86.55 | 77.97 |
| FTransUNet | RGB-D | 90.16 | 92.02 | 97.48 | 84.95 | 88.10 | 93.90 | 90.95 | 83.74 |
| MCFNet | RGB-D | 83.79 | 89.20 | 84.62 | 81.05 | 83.63 | 92.37 | 85.60 | 75.42 |
| MGFNet | RGB-D | 90.85 | 93.04 | 96.83 | 86.09 | 89.00 | 94.52 | 88.35 | 80.16 |
| FSECrossNet (Ours) | RGB-D | **91.43** | **93.60** | **98.00** | 87.76 | 87.77 | **96.43** | **92.43** | **86.22** |

TABLE III
ACCURACY–EFFICIENCY COMPARISON ON THE VAIHINGEN DATASET.

| Method | Modal | Params (M) | FLOPs (G) | mIoU(%) | Inference Time (ms) |
|---|---|---|---|---|---|
| UNetFormer | RGB | 11.69 | 5.97 | 76.86 | 12.5 |
| TransUNet | RGB | 93.23 | 64.55 | 77.15 | 45.3 |
| LoGCAN++ | RGB | 24.30 | 18.23 | 76.21 | 18.7 |
| CMGFNet | RGB-D | 85.22 | 77.71 | 81.27 | 52.8 |
| FTransUNet | RGB-D | 179.42 | 112.22 | 82.14 | 78.5 |
| MCFNet | RGB-D | 49.69 | 55.72 | 78.01 | 38.2 |
| MGFNet | RGB-D | 82.91 | 81.26 | 82.06 | 48.6 |
| FSECrossNet | RGB-D | **27.96** | **33.32** | **83.34** | **22.3** |

TABLE IV
ABLATION STUDY RESULTS FOR KEY MODULES.

| Configuration | GFF | DFSF | CMAF | mIoU(%) |
|---|---|---|---|---|
| Full Model | ✓ | ✓ | ✓ | **83.34** |
| w/o GFF | × | ✓ | ✓ | 81.21 |
| w/o DFSF | ✓ | × | ✓ | 80.93 |
| w/o CMAF | ✓ | ✓ | × | 81.38 |

TABLE V
ABLATION STUDY RESULTS FOR DFSF MODULE COMPONENTS.

| Configuration | Freq Branch | Spatial Branch | mIoU(%) |
|---|---|---|---|
| Full Model | ✓ | ✓ | **83.34** |
| w/o Freq Branch | × | ✓ | 81.68 |
| w/o Spatial Branch | ✓ | × | 81.88 |

TABLE VI
COMPARISON OF DCT AND FFT IN FREQUENCY ENHANCEMENT BRANCH.

| Frequency Method | mIoU(%) | mF1(%) |
|---|---|---|
| DCT | **83.34** | **90.69** |
| FFT | 82.15 | 89.82 |

TABLE VII
ABLATION STUDY RESULTS FOR CMAF MODULE COMPONENTS.

| Configuration | Self-Attn | Cross-Attn | mIoU(%) |
|---|---|---|---|
| Full Model | ✓ | ✓ | **83.34** |
| w/o Cross-Attn | ✓ | × | 82.04 |
| w/o Self-Attn | × | ✓ | 81.37 |

TABLE VIII
ABLATION STUDY RESULTS FOR KERNEL PATTERNS IN MSCONV MODULE.

| Configuration | Pattern type | mIoU(%) |
|---|---|---|
| 3×3 + 5×5 + 7×7 | multi-scale square | **83.34** |
| 3×3 + 5×5 | two-scale square | 83.00 |
| 3×3 + 7×7 | two-scale square | 82.90 |
| 5×5 + 7×7 | two-scale square | 82.95 |
| 1×3 + 3×1 | asymmetric | 82.70 |
| 7×7 + 9×9 | large-kernel | 82.60 |

## REFERENCES


[1] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[2] X. Ma *et al.*, "LOGCAN++: Adaptive local-global class-aware network for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, 2025.

[3] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[4] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 96–115, 2022.

[5] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 783–792.

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

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[10] X. Ma et al., "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[11] H. Wu et al., "MCFNet: Multiscale cross-modal fusion network for remote sensing image semantic segmentation," *IEEE Signal Process. Lett.*, 2025.

[12] H. Wu, Z. Li, and Z. Wen, "MGFNet: A multiscale gated fusion network for multimodal semantic segmentation," *Vis. Comput.*, pp. 1–13, 2025.