

Article

CAF-Net: A Unified Framework for Resolving Spatial–Frequency Representation Conflicts in Multimodal Remote Sensing Segmentation

Ming Ma ¹ , Shuaijie Wu ¹ , Xiang Li ¹  and Hong Wei ^{2,*} 

¹ College of Computer Science, Inner Mongolia University, No. 235 West College Road, Saihan District, Hohhot 010021, China; csmaming@imu.edu.cn (M.M.)

² Department of Computer Science, University of Reading, Reading RG6 6DH, UK

* Correspondence: h.wei@reading.ac.uk

Highlights

What are the main findings?

- A previously underexplored spatial–frequency representation conflict is identified and quantitatively analyzed in multimodal remote sensing segmentation.
- CAF-Net introduces a unified conflict-resolution framework that combines cross-modal structure guidance, adaptive frequency alignment, and reliability-aware fusion.
- Extensive experiments demonstrate superior segmentation accuracy, robustness, and representation consistency on the ISPRS Vaihingen and Potsdam benchmarks.

What are the implications of the main findings?

- The study suggests that resolving representation conflicts is an effective alternative to simply increasing model complexity or feature extraction capacity.
- Explicit modeling of spatial and frequency interactions improves multimodal feature integration and robustness under degraded sensing conditions.
- The proposed conflict-aware learning paradigm can be extended to other heterogeneous remote sensing modalities, such as SAR, LiDAR, and hyperspectral imagery.

Abstract

Multimodal remote sensing segmentation commonly integrates optical imagery with Digital Surface Models (DSMs) to improve land-cover understanding. However, existing methods often overlook a critical issue, namely the inconsistency between spatial-domain structural details and frequency-domain semantics, which can lead to misaligned features and degraded performance in complex scenes. To address this problem, we formulate spatial–frequency representation conflict as a unified multimodal learning problem and propose the Conflict-Aware Fusion Network (CAF-Net). Specifically, Cross-Modal Structure Guidance (CMSG) extracts DSM-derived high-pass structural cues and conditionally modulates optical features to improve boundary consistency. The Adaptive Cross-Frequency Module (ACFM) separates DCT coefficients using a fixed radial mask, adaptively reweights low- and high-frequency components, and performs cross-modal alignment at the highest encoder stage. Uncertainty-Aware Fusion (UAF) predicts pixel-wise relative reliability scores and normalizes them across modalities to suppress low-confidence responses. This coordinated design links spatial refinement, frequency alignment, and reliability-guided fusion instead of treating them as independent feature-enhancement operations. Experiments on the ISPRS Vaihingen and Potsdam datasets yield mIoU scores of 84.35% and 86.86%, respectively. The results indicate that explicitly modeling spatial–frequency discrepancies can improve multimodal segmentation accuracy and representation consistency.



Academic Editors: Lei Wang, Xiyu Qi, Yongqiang Mao, Xiaoxuan Liu, Zicong Zhu and Zhuohong Li

Received: 3 May 2026

Revised: 24 June 2026

Accepted: 29 June 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: multimodal; uncertainty modeling; conflict-aware fusion; remote sensing segmentation; spatial–frequency representation

1. Introduction

Semantic segmentation of remote sensing imagery plays a fundamental role in a wide range of applications, including urban planning [1], environmental monitoring, and disaster management [2]. With the increasing availability of high-resolution and multi-source data, multimodal learning has become a dominant paradigm for improving scene understanding. In particular, the combination of optical imagery and Digital Surface Models (DSM) has been widely adopted, as these modalities provide complementary information. Optical imagery captures rich spectral and texture patterns, while DSM data encode geometric structure and elevation cues, enabling more precise delineation of complex objects [3].

Despite these advantages, multimodal remote sensing segmentation remains highly challenging in complex scenarios. High-resolution images often exhibit large intra-class variation, irregular object boundaries, and significant scale differences. Moreover, multimodal data introduce additional difficulties, including modality-specific noise, misalignment, and inconsistent feature distributions. These factors make it difficult for existing models to achieve stable and reliable fusion, especially in ambiguous regions.

Recent advances in deep learning have explored different strategies to address these challenges. Spatial-based approaches focus on enhancing structural consistency by leveraging attention mechanisms or geometric guidance [4], which improve boundary delineation and local continuity. In parallel, frequency-based methods have been introduced to capture global semantic information through spectral decomposition [5], enabling more robust modeling of long-range dependencies.

However, existing methods generally assume that spatial and frequency representations can be integrated consistently throughout the network. In practice, these two representations often emphasize different aspects of the scene. Spatial representations favor local continuity, boundary preservation, and geometric alignment, whereas frequency representations emphasize global distributions, periodic patterns, and semantic abstraction. As a result, the two representations may generate inconsistent responses at the same spatial location, particularly in regions with complex boundaries, heterogeneous land-cover distributions, or ambiguous semantic structures. Such inconsistencies can propagate through multimodal fusion and negatively affect feature alignment and segmentation accuracy. Despite the growing interest in frequency-aware segmentation, the inconsistency between spatial-domain structural representations and frequency-domain semantic representations has not been explicitly modeled in existing multimodal remote sensing segmentation frameworks. In this work, we refer to this phenomenon as spatial–frequency representation conflict and investigate its impact on multimodal feature fusion.

As illustrated in Figure 1, we construct two complementary feature views from optical and DSM inputs. The spatial representation is obtained using a Sobel-based gradient operator applied to intermediate feature maps, which emphasizes local intensity variations and boundary structures. The frequency representation is derived from a Discrete Cosine Transform (DCT) applied to the same feature maps, capturing global structure and high-frequency components such as edges and fine details. Both representations are extracted from the intermediate encoder features (before the fusion stage) of the baseline feature extractor, ensuring consistent spatial resolution and channel alignment. Due to their different sensitivity characteristics, these two views may exhibit discrepancies at corresponding spatial locations. In this work, we quantify such discrepancies using a

feature-level cosine similarity metric and refer to it as representation conflict. We observe that higher discrepancy is often associated with less stable predictions in challenging scenarios. Existing methods typically do not explicitly model this cross-view discrepancy, which may limit their ability to handle complex multimodal interactions.

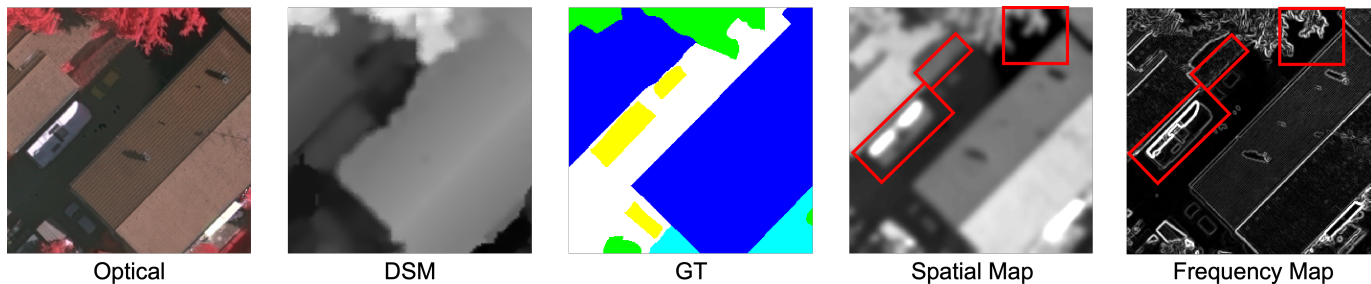


Figure 1. An illustration of spatial–frequency representation discrepancy in multimodal remote sensing segmentation. The spatial representation is obtained by applying a Sobel gradient operator to intermediate encoder features, emphasizing structural continuity and boundary-aware responses. The frequency representation is generated by applying a Discrete Cosine Transform (DCT) to the same intermediate features, capturing high-frequency components such as edges and texture variations. Both representations are extracted from the intermediate feature maps of the baseline encoder before multimodal fusion, and are used solely for visualization analysis rather than final prediction. As highlighted by the red boxes, the two representations may respond differently to the same scene structures in complex regions, leading to feature-level inconsistency and motivating explicit modeling of spatial–frequency interactions.

To address the aforementioned issue, we propose the Conflict-Aware Fusion Network (CAF-Net), a coordinated architecture designed to explicitly address spatial–frequency conflicts in multimodal fusion. The proposed framework integrates three complementary operations: (1) Cross-Modal Structure Guidance (CMSG), which injects DSM-derived geometric priors to enhance spatial consistency; (2) Adaptive Cross-Frequency Module (ACFM), which performs frequency decomposition and alignment to improve semantic coherence; and (3) Uncertainty-Aware Fusion (UAF), which estimates pixel-wise reliability to dynamically balance multimodal features and suppress conflicting responses. Rather than independently stacking these operations, CAF-Net couples them through consistency objectives and stage-wise feature interaction: CMSG supplies structure-enhanced spatial features, ACFM aligns their high-level frequency responses, and UAF regulates the resulting cross-modal interaction according to relative reliability.

The main contributions of this work are summarized as follows:

- (1) We formulate spatial–frequency representation conflict as a feature-level discrepancy in optical–DSM segmentation and introduce global and boundary-focused diagnostic measures for its quantitative analysis.
- (2) We develop CMSG and ACFM as complementary conflict-processing components: CMSG extracts DSM-derived high-pass structure for conditional spatial modulation, whereas ACFM performs explicitly masked frequency decomposition, adaptive band reweighting, and cross-modal alignment.
- (3) We introduce UAF as a pixel-wise relative-reliability mechanism that normalizes modality confidence scores and regulates the interaction between structure-guided and frequency-aligned representations.
- (4) Experiments on the ISPRS Vaihingen and Potsdam datasets evaluate segmentation accuracy, component contributions, computational complexity, feature discrepancy, and robustness to modality degradation.

2. Related Work

2.1. Spatial-Domain Fusion

Spatial-domain fusion methods aim to improve multimodal integration by enhancing spatial consistency and structural alignment across heterogeneous inputs [6]. These approaches typically rely on attention-based interaction, geometric guidance, or feature alignment to model cross-modal relationships. Early methods mainly adopt symmetric feature concatenation, treating different modalities equally but often neglecting modality-specific characteristics. To address this limitation, recent studies have increasingly explored asymmetric modeling and modality-aware interaction.

For instance, ASANet [7] introduces asymmetric semantic alignment to explicitly capture modality discrepancies, while RS³Mamba [8] leverages state space models to efficiently model long-range spatial dependencies with linear complexity. More recent studies further explore multimodal fusion strategies in remote sensing scenarios. For example, a multimodal cross-fusion Mamba framework is proposed in [9], which enhances cross-modal interaction through complementary masked self-supervision, while projection-driven fusion methods integrate LiDAR and image data by explicitly modeling geometric consistency [10]. In addition, CMGFNet [11] employs cascaded fusion strategies to progressively refine feature selection, and RecLMIS [12] enhances cross-modal alignment through conditional reconstruction mechanisms.

However, these methods primarily focus on spatial consistency and local structural refinement, while largely overlooking discrepancies that arise from complementary frequency-domain representations. This limitation may result in incomplete modeling of global contextual dependencies, especially in complex scenes where multi-scale information is essential.

2.2. Frequency-Domain Fusion

To address the limitations of purely spatial modeling, frequency-domain approaches have been introduced to capture global context and long-range dependencies. By decomposing features into different frequency components, these methods enhance global semantic modeling while preserving fine-grained structural details.

For example, GFNet [13] replaces self-attention with global frequency filtering, achieving efficient long-range modeling. In multimodal remote sensing segmentation, FDMFNet [14] introduces frequency decoupling to reduce modality redundancy and improve semantic consistency. Similarly, Wave-ViT [15] employs wavelet transforms to reduce token redundancy while preserving high-frequency information. Beyond remote sensing, FreqFormer [16] and EW-ViT [17] further demonstrate the effectiveness of frequency-domain representations in capturing global dependencies.

Despite their effectiveness, these methods mainly emphasize global spectral modeling while often neglecting local spatial structure consistency. As a result, directly integrating frequency-domain features with spatial representations may introduce inconsistencies, particularly in boundary-sensitive regions where precise localization is required.

2.3. Spatial–Frequency Joint Modeling

Recently, several methods have explored joint modeling of spatial and frequency information to leverage their complementary strengths. For instance, FAENet [18] integrates wavelet transforms with attention mechanisms to capture spectral–spatial context, while AFNO [19] combines Fourier-based global filtering with convolutional operations to balance global and local representations.

Although these methods improve representation capacity through feature integration, they mainly focus on enhancing fused representations rather than explicitly addressing

inconsistencies between spatial and frequency domains. As pointed out in UFCNet [20], combining spatial and spectral operators improves performance but does not explicitly resolve conflicting responses arising from their fundamentally different characteristics.

In essence, spatial representations emphasize local continuity and structural alignment, whereas frequency representations capture global distributions and periodic patterns. These inherently different properties may lead to inconsistent responses during multimodal fusion, especially in complex scenes with heterogeneous structures. Such inconsistencies remain insufficiently explored in existing methods.

Therefore, explicitly modeling and resolving spatial–frequency representation conflict remains an open and important problem in multimodal remote sensing segmentation, which motivates the design of our CAF-Net.

3. Methodology

3.1. Overall Architecture

The overall architecture of CAF-Net is illustrated in Figure 2a. The proposed framework follows a dual-branch encoder design based on SegFormer-B4 [21] to process optical imagery and DSM modalities in parallel. Given an optical image I_{opt} and a DSM input I_{DSM} , two modality-specific branches extract hierarchical representations across four stages, where the spatial resolution is progressively reduced from $\frac{1}{4}$ to $\frac{1}{32}$ of the input. This hierarchical design enables the model to capture both low-level spatial details and high-level semantic information.

Although optical imagery and DSM data provide complementary cues for scene understanding, directly fusing these features is non-trivial. Optical imagery primarily encodes appearance information such as texture and spectral patterns, while DSM data describe geometric structures and elevation variations. These intrinsic differences often lead to representation conflicts, especially in challenging regions affected by noise, occlusion, or complex object boundaries.

Moreover, such conflicts may be further amplified when spatial representations interact with frequency-domain characteristics, resulting in inconsistent responses during feature fusion.

To address these challenges, CAF-Net is designed to explicitly model and resolve spatial–frequency representation conflicts in a unified framework. Instead of relying on a single fusion strategy, we decompose the fusion process into three complementary components, each handled by a dedicated module. First, Cross-Modal Structure Guidance (CMSG) is introduced to enhance spatial consistency. By injecting DSM-derived geometric priors into optical features, CMSG strengthens boundary awareness and improves structural alignment between modalities. Second, the Adaptive Cross-Frequency Module (ACFM) focuses on modeling complementary frequency representations. By decomposing features into different frequency components and adaptively reweighting them, ACFM improves global semantic consistency and mitigates discrepancies caused by heterogeneous data distributions. Third, the Uncertainty-Aware Fusion (UAF) serves as a dynamic fusion mechanism. Instead of directly aggregating features, UAF estimates pixel-wise reliability for each modality and adaptively balances their contributions, thereby suppressing unreliable responses and preserving complementary information.

These three components are progressively integrated across different stages of the encoder. CMSG and UAF operate throughout the network to refine spatial representations and stabilize multimodal fusion, while ACFM is applied only at the final stage to enhance high-level semantic alignment. This design is motivated by two considerations. First, frequency-domain modeling is more effective at higher semantic levels, where features encode global context rather than low-level textures. Applying frequency alignment at early

stages may disrupt local spatial details that are critical for boundary preservation. Second, the Discrete Cosine Transform (DCT) operations in ACFM involve global transformations, and applying them at lower-resolution feature maps significantly reduces computational overhead. Through this coordinated design, CAF-Net jointly models spatial and frequency representations, enabling robust and consistent multimodal feature learning in complex remote sensing scenarios.

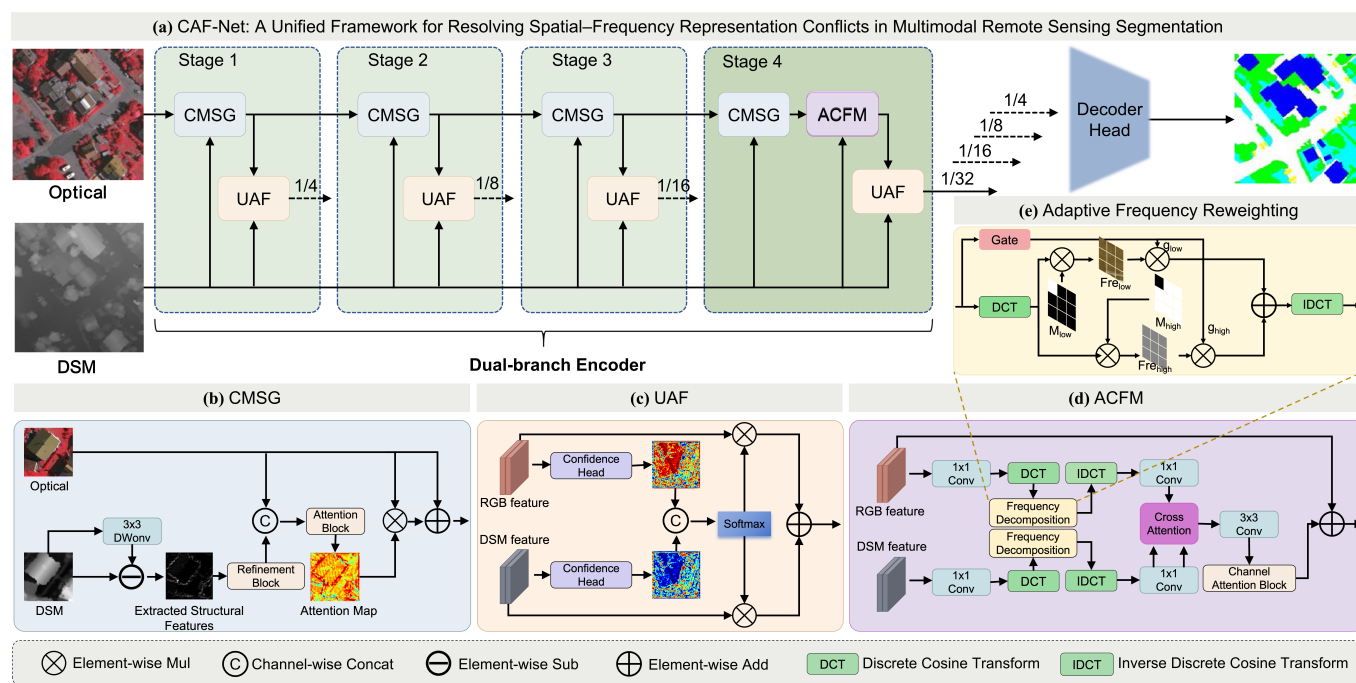


Figure 2. Overall architecture of the proposed CAF-Net. The framework consists of three key components: Cross-Modal Structure Guidance (CMSG) for spatial refinement, Adaptive Cross-Frequency Module (ACFM) for frequency-domain alignment, and Uncertainty-Aware Fusion (UAF) for adaptive multimodal fusion.

3.2. Cross-Modal Structure Guidance

In complex remote sensing scenes, accurate boundary delineation remains a challenging problem due to texture ambiguity, illumination variations, and background noise in optical imagery. Although optical features provide rich appearance information, they are often insufficient for capturing precise object structures. In contrast, DSM data encode stable geometric cues, such as elevation discontinuities and object contours, which are less sensitive to appearance variations.

As illustrated in Figure 2b, motivated by this observation, we introduce CMSG to explicitly leverage DSM-derived structural information for spatial refinement within the encoder. The key idea is to enhance structure-consistent regions in optical features while suppressing ambiguous responses caused by texture noise.

At stage l , structural representations are first extracted from DSM features by suppressing locally smooth components:

$$S^l = F_{\text{DSM}}^l - \mathcal{D}(F_{\text{DSM}}^l) \quad (1)$$

where $\mathcal{D}(\cdot)$ denotes a learnable depth-wise convolution with a kernel size of 3×3 , stride 1, padding 1, and no bias. It performs local smoothing independently on each channel. A separate CMSG instance is used at each encoder stage; therefore, its parameters are not shared across stages. All convolutional layers in CMSG are initialized from a zero-mean normal distribution with standard deviation $\sqrt{2/\text{fan_out}}$, where $\text{fan_out} = k_h k_w C_{\text{out}} / \text{groups}$; con-

volutional biases, when present, are initialized to zero. Batch-normalization scale and bias parameters are initialized to 1 and 0, respectively. This residual-like formulation highlights high-frequency information corresponding to edges and structural discontinuities.

The extracted structural representations are further refined through a lightweight transformation module, referred to as the Refinement Block, which consists of Batch Normalization; ReLU activation; a bias-free 1×1 convolution that preserves the channel dimension, followed by another Batch Normalization step; and ReLU activation. This design suppresses noise and enhances discriminative geometric patterns while introducing limited computational overhead.

The refined structural representations are then integrated with optical features through an attention-based interaction mechanism. Specifically, the Attention Block concatenates the optical and structural features and applies two 1×1 convolutions with an intermediate ReLU activation and a final sigmoid normalization. For an input with C_l channels, the intermediate width is $\max(8, C_l/4)$ and the output contains C_l spatial attention maps. This conditional attention mechanism allows the model to emphasize responses that are consistent with DSM-derived structures while retaining the original optical representation through the residual update in Equation (2).

Finally, the optical representation is modulated in a residual manner:

$$\hat{F}_{\text{opt}}^l = F_{\text{opt}}^l + F_{\text{opt}}^l \otimes A^l \quad (2)$$

where A^l denotes the spatial attention map generated from the fused representations using a lightweight convolutional attention head followed by a sigmoid activation. This formulation enhances boundary-related responses while preserving the original semantic information.

Compared with direct feature fusion, CMSG explicitly injects geometric priors into the spatial representation, enabling more accurate boundary delineation and improved alignment between modalities. CMSG outputs a structure-enhanced feature map with the same spatial resolution and channel dimension as the input features, which is used to refine the optical representation for subsequent fusion. This design is particularly effective in challenging regions where texture cues are unreliable but structural information remains consistent.

3.3. Adaptive Cross-Frequency Module

While CMSG improves spatial consistency, discrepancies between modalities also arise in the frequency domain. In multimodal remote sensing data, different modalities often exhibit complementary characteristics across frequency bands. In particular, low-frequency components mainly capture coarse semantic structures and global contextual information, whereas high-frequency components encode fine-grained details such as edges, textures, and local variations. However, directly combining these representations may introduce inconsistency, since different modalities often emphasize different frequency ranges and respond differently to scene variations.

As illustrated in Figure 2d, we introduce ACFM at the final stage of the encoder, where feature maps contain rich semantic information and reduced spatial resolution. Given the high-level features F_{opt}^4 and F_{DSM}^4 , we first project them into a compact embedding space and transform them into the frequency domain using a Discrete Cosine Transform:

$$\mathcal{F}_m = \mathcal{D}_{\text{DCT}}(F_m^4), \quad m \in \{\text{opt}, \text{DSM}\} \quad (3)$$

where $\mathcal{D}_{\text{DCT}}(\cdot)$ denotes the DCT operator. This transformation decomposes the feature map into different frequency components, making global semantic structures and local details explicitly separable.

Based on this representation, we perform frequency decomposition using a fixed radial binary mask. For a DCT coefficient at index (u, v) in an $H \times W$ feature map, the normalized radial coordinate and low-frequency mask are defined as

$$r(u, v) = \frac{\sqrt{(u/(H-1))^2 + (v/(W-1))^2}}{\max_{i,j} \sqrt{(i/(H-1))^2 + (j/(W-1))^2}}, \quad M_{\text{low}}(u, v) = \mathbb{I}[r(u, v) \leq \rho], \quad (4)$$

where a denominator is replaced by 1 if the corresponding spatial dimension equals 1. The high-frequency mask is $M_{\text{high}} = 1 - M_{\text{low}}$, and $\rho = 0.35$ in all experiments. Thus, $\mathcal{F}_{\text{low},m} = \mathcal{F}_m \odot M_{\text{low}}$ and $\mathcal{F}_{\text{high},m} = \mathcal{F}_m \odot M_{\text{high}}$. The low-frequency component mainly preserves smooth and coarse semantic information, while the high-frequency component retains local structural variations and detailed responses.

Specifically, a lightweight gating branch predicts the importance of low-frequency and high-frequency responses, and the recomposed frequency feature is formulated as

$$\mathcal{F}_{\text{mix},m} = g_{\text{low}} \mathcal{F}_{\text{low},m} + g_{\text{high}} \mathcal{F}_{\text{high},m} \quad (5)$$

where $\mathcal{F}_{\text{low},m}$ and $\mathcal{F}_{\text{high},m}$ denote the low-frequency and high-frequency components of modality $m \in \{\text{opt}, \text{DSM}\}$, respectively. The coefficients g_{low} and g_{high} are sample-wise scalars of shape $B \times 1 \times 1 \times 1$. They are predicted from each projected modality feature by global average pooling, a 1×1 convolution that reduces the width to $\max(8, C/4)$, ReLU, a second 1×1 convolution with two outputs, and element-wise sigmoid activation. The two coefficients are independently learned and are not constrained to sum to one, allowing the network to enhance or suppress the two frequency bands independently.

After adaptive reweighting, the fused frequency representation is mapped back to the spatial domain through the inverse discrete cosine transform. The Inverse Discrete Cosine Transform (IDCT) operation is necessary because subsequent multimodal interaction and feature refinement are performed in the spatial domain:

$$\tilde{\mathcal{F}}_m^4 = \mathcal{D}_{\text{IDCT}}(\mathcal{F}_{\text{mix},m}), \quad m \in \{\text{opt}, \text{DSM}\} \quad (6)$$

where $\mathcal{D}_{\text{IDCT}}(\cdot)$ denotes the inverse DCT operator. This step converts frequency-enhanced features into spatial feature maps while preserving the information selected during frequency decomposition and reweighting.

To further improve cross-modal consistency, the reconstructed optical and DSM features are then aligned through a cross-attention mechanism. In this process, optical imagery provides the query while DSM provides the key and value, enabling the model to explicitly capture complementary responses across modalities and reduce semantic discrepancy between them. Through this operation, ACFM encourages the frequency-enhanced features of the two modalities to focus on semantically corresponding regions rather than remaining independently distributed.

Following cross-attention, a Channel Attention Block is employed to refine the aligned feature responses. The purpose of this block is to adaptively recalibrate channel-wise importance after cross-modal interaction. Since different channels may carry different levels of semantic relevance after frequency alignment, channel attention helps emphasize informative responses while suppressing redundant or noisy channels, thereby improving representation compactness and discriminability.

Finally, the refined feature is fused with the original optical feature through a residual connection:

$$F_{ACFM} = F_{opt}^4 + \tilde{F}_{align} \quad (7)$$

where $\tilde{F}_{align}$ denotes the aligned and channel-refined frequency feature in the spatial domain. This residual design preserves the original semantic information while injecting complementary frequency-domain cues.

Compared with direct spatial fusion, ACFM explicitly models frequency-domain interactions and progressively completes four steps: frequency decomposition, adaptive reweighting, inverse reconstruction, and cross-modal alignment. Through this design, the module resolves discrepancies between global semantic patterns and local structural details, improves cross-modal consistency, and provides more robust high-level representations in challenging remote sensing scenarios.

3.4. Uncertainty-Aware Fusion

While CMSG and ACFM respectively address spatial and frequency discrepancies, inconsistencies between modalities may still persist during feature integration. In real-world remote sensing scenarios, the reliability of optical and DSM representations often varies significantly across spatial locations. For example, optical representations may become unreliable in shadowed regions or areas with severe texture ambiguity, whereas DSM representations may be affected by noise, missing data, or sensor artifacts. Under such conditions, directly aggregating representations from different modalities can propagate unreliable information and degrade segmentation performance.

To address this issue, we introduce UAF, which formulates multimodal fusion as a reliability-aware decision process. The confidence maps are learned jointly through the segmentation objective and are used as relative reliability indicators rather than calibrated probabilistic uncertainty estimates. Accordingly, the term “uncertainty-aware” refers to adaptive suppression of comparatively low-confidence modality responses, not Bayesian uncertainty estimation. As illustrated in Figure 2c, UAF estimates the relative confidence of each modality at each spatial location and dynamically adjusts their contributions.

At stage l , given the structure-enhanced optical representation $\hat{F}_{opt}^l$ and the DSM representation F_{DSM}^l , two modality-specific confidence maps are first predicted:

$$C_{opt}^l = \Psi_{opt}(\hat{F}_{opt}^l), \quad C_{DSM}^l = \Psi_{DSM}(F_{DSM}^l) \quad (8)$$

where $\Psi_{opt}(\cdot)$ and $\Psi_{DSM}(\cdot)$ are two modality-specific confidence estimation modules with independent parameters. Each module consists of a bias-free depth-wise 3×3 convolution with stride 1 and padding 1, Batch Normalization, ReLU, a bias-free 1×1 projection to $\max(8, C_l/4)$ channels, another ReLU, and a final 1×1 convolution that produces a single-channel score map. The outputs C_{opt}^l and C_{DSM}^l are unnormalized relative-confidence logits; they are not probabilities before the cross-modal normalization in Equation (9):

$$w_{opt}^l, w_{DSM}^l = \text{Softmax}(C_{opt}^l/T, C_{DSM}^l/T) \quad (9)$$

The confidence maps are concatenated along the modality dimension and normalized using a temperature-scaled Softmax operation, producing two spatially adaptive weighting maps whose values sum to one at each pixel location. Here, $T = 1.0$ is used in the reported experiments. A smaller T leads to a more peaked distribution, emphasizing the dominant modality, while a larger T produces smoother weighting across modalities.

These weights are finally used to compute the fused representation:

$$F_{UAF}^l = w_{opt}^l \odot \hat{F}_{opt}^l + w_{DSM}^l \odot F_{DSM}^l \quad (10)$$

This formulation introduces an explicit competition mechanism between modalities, allowing the model to prioritize more reliable representations while suppressing low-confidence responses. By dynamically adjusting fusion weights across spatial locations, UAF effectively mitigates the impact of modality-specific noise and reduces over-reliance on any single modality. This design complements the spatial refinement in CMSG and the frequency alignment in ACFM, enabling more robust and conflict-aware multimodal representation learning across both spatial and frequency domains.

3.5. Multi-Criteria Loss Function

To effectively optimize multimodal representations and alleviate spatial–frequency inconsistencies, we adopt a multi-criteria training objective that jointly considers segmentation accuracy and cross-domain consistency. The overall loss function is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \alpha \mathcal{L}_{\text{cons}} + \beta \mathcal{L}_{\text{low}} \quad (11)$$

where $\mathcal{L}_{\text{seg}}$ denotes the segmentation loss, and α and β are balancing coefficients.

Segmentation Loss ($\mathcal{L}_{\text{seg}}$). The segmentation loss supervises pixel-wise classification and ensures accurate semantic prediction. It is implemented as a combination of cross-entropy and Dice loss to balance class distribution and boundary accuracy.

Collaborative Consistency Loss ($\mathcal{L}_{\text{cons}}$). To explicitly reduce inconsistencies between different feature representations, we introduce a collaborative consistency constraint across multiple branches. Specifically, at Stage 4, we enforce alignment between the outputs of CMSG, ACFM, and UAF:

$$\mathcal{L}_{\text{cons}} = \|F_{\text{ACFM}} - F_{\text{UAF}}^4\|_1 + \|\hat{F}_{\text{opt}}^4 - F_{\text{UAF}}^4\|_1 \quad (12)$$

This loss encourages the spatially refined features (from CMSG) and frequency-aligned features (from ACFM) to be consistent with the fused representation. By explicitly constraining multiple representations to converge, this term effectively mitigates conflicts between spatial and frequency domains and promotes coherent feature learning.

Low-frequency Consistency Loss ($\mathcal{L}_{\text{low}}$). To further stabilize global semantic structures, we introduce a low-frequency consistency constraint. Based on the frequency decomposition in ACFM, we focus on aligning low-frequency components that encode global contextual information:

$$\mathcal{L}_{\text{low}} = \text{MSE}(\mathcal{D}_{\text{DCT}}(F_{\text{opt}}^4) \odot M_{\text{low}}, \mathcal{D}_{\text{DCT}}(F_{\text{DSM}}^4) \odot M_{\text{low}}) \quad (13)$$

Here, the low-frequency mask is applied to the DCT coefficients rather than directly to the spatial-domain feature maps. Specifically, $\mathcal{D}_{\text{DCT}}(F_{\text{opt}}^4)$ and $\mathcal{D}_{\text{DCT}}(F_{\text{DSM}}^4)$ denote the frequency-domain representations of the Stage-4 optical and DSM features, respectively, and M_{low} denotes the same fixed radius-based low-frequency mask used in ACFM, with the normalized radius threshold $\rho = 0.35$. The loss therefore measures the discrepancy between the masked low-frequency coefficients of the two modalities. This term enforces consistency in coarse semantic representations across modalities, reducing discrepancies in global structure and improving overall semantic coherence.

Overall, the proposed loss function complements the architectural design by jointly constraining spatial refinement, frequency alignment, and multimodal fusion. Through this coordinated optimization, the model achieves more consistent and robust representations under complex multimodal conditions.

The balancing coefficients α and β control the relative importance of consistency constraints. In practice, α is set to 0.5 to moderately enforce cross-branch alignment without

overwhelming the segmentation objective, while β is set to 0.1 to provide a mild regularization on low-frequency components. We empirically observe that this configuration achieves a good trade-off between optimization stability and performance, and the model is not highly sensitive to small variations of these parameters.

3.6. Conflict Measurement

To analyze feature discrepancies between the spatial and frequency processing paths, we introduce a cosine-distance-based diagnostic metric. For every model variant, the two compared tensors are taken before UAF at Stage 4: $F_1 = F_{\text{MSG}}^4$ is the output of CMSG and $F_2 = F_{\text{ACFM}}^4$ is the output of ACFM. Both tensors are generated in parallel from the same Stage-4 optical and DSM encoder features. They have identical dimensions $C_4 \times H_4 \times W_4$ ($C_4 = 512$ and H_4, W_4 are $1/32$ of the input dimensions), so no channel projection or spatial interpolation is required before comparison. When a module is disabled in an ablation variant, its corresponding path returns the unmodified Stage-4 optical feature, which preserves a consistent comparison interface across variants.

Given two feature maps F_1 and F_2 , the conflict at each spatial location (h, w) is defined as

$$\mathcal{C}(h, w) = 1 - \frac{F_1(:, h, w) \cdot F_2(:, h, w)}{\|F_1(:, h, w)\|_2 \cdot \|F_2(:, h, w)\|_2} \quad (14)$$

where $F(:, h, w)$ denotes the feature vector at spatial position (h, w) . This formulation measures the angular difference between feature embeddings, where larger values indicate higher feature discrepancy. Based on the resulting conflict map $\mathcal{C}$, we compute two complementary statistics.

Global Conflict. The overall discrepancy across the feature map is computed as

$$\mathcal{C}_{\text{global}} = \frac{1}{HW} \sum_{h,w} \mathcal{C}(h, w) \quad (15)$$

Boundary Conflict. To focus on structurally sensitive regions, we further evaluate conflict around object boundaries:

$$\mathcal{C}_{\text{boundary}} = \frac{\sum_{h,w} B(h, w) \cdot \mathcal{C}(h, w)}{\sum_{h,w} B(h, w)} \quad (16)$$

where $B(h, w)$ denotes a binary boundary mask derived independently for each ground-truth patch. Specifically, the class-index label map is converted to an 8-bit image and processed using the Canny operator with lower and upper thresholds of 0 and 1. The resulting one-pixel-wide boundary map is binarized and resized to the Stage-4 feature resolution using nearest-neighbor interpolation. No additional morphological dilation is applied. Global Conflict is averaged over all Stage-4 spatial positions and evaluated patches, whereas Boundary Conflict is averaged only over positions selected by the resized boundary masks.

The two metrics are computed after training and are used only as diagnostic indicators of global and boundary-focused feature discrepancy. They do not form part of the optimization objective and should not be interpreted as causal evidence that a particular module resolves the observed discrepancies.

4. Results

4.1. Experimental Setup

Datasets. Experiments are conducted on two widely used benchmarks for high-resolution remote sensing semantic segmentation, namely ISPRS Vaihingen and Potsdam [22].

ISPRS Vaihingen. The Vaihingen dataset consists of 33 high-resolution aerial orthophotos acquired over Vaihingen, Germany, with an average ground sampling distance (GSD) of 9 cm. Each image contains three spectral bands (near-infrared, red, and green), referred to as optical imagery, along with corresponding normalized digital surface models (nDSMs), which provide elevation information. The dataset is annotated with six land-cover classes, including impervious surfaces, buildings, low vegetation, trees, cars, and background. Following the standard protocol, 12 tiles are used for training and 4 tiles for testing. Due to the large image size, a sliding-window strategy with a patch size of 256×256 is adopted during both training and inference.

ISPRS Potsdam. The Potsdam dataset provides higher-resolution urban scenes with a GSD of 5 cm. It contains 24 orthophotos of size approximately 6000×6000 pixels, captured over Potsdam, Germany. Each image includes four spectral bands (red, green, blue, and near-infrared) along with corresponding nDSM data. Similar to Vaihingen, the dataset is annotated with the same set of semantic classes. We follow the official split, using 18 images for training and 6 for testing. A sliding-window approach with a patch size of 256×256 is also applied to handle large-scale inputs.

Implementation Details. All experiments were implemented in PyTorch 2.5.0 and trained using AdamW with an initial learning rate of 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. A cosine-annealing schedule was applied at each optimization step over the complete 100-epoch training process, without learning-rate warmup. The batch size was 10. No explicit class rebalancing was used. Following the adopted ISPRS split, no additional validation subset was constructed; models were trained on the predefined training tiles and evaluated on the stated test tiles. During training, 256×256 patches were randomly sampled. Data augmentation comprised random horizontal and vertical flips, each with probability 0.5. Optical values were scaled to $[0, 1]$ by division by 255, while each DSM tile was independently min-max-normalized. During inference, a 256×256 sliding window with a stride of 32 pixels was used, and logits from overlapping windows were accumulated before the pixel-wise argmax operation. For fair comparison, no official test-set results or official task-specific checkpoints were directly copied into the reported tables. All compared methods were retrained on the ISPRS Vaihingen and Potsdam datasets using their publicly available implementations. The dataset splits, input preprocessing, patch size, data augmentation, total number of epochs, sliding-window inference protocol, and evaluation metrics were unified across methods. At the same time, method-specific architectural settings were preserved, including the published backbone, fusion structure, channel configuration, and officially recommended module hyperparameters, so that the original designs were not altered. When a method officially adopted a pretrained backbone initialization, the same initialization strategy was retained; however, the segmentation heads and fusion modules were trained on the ISPRS training sets, and no official remote-sensing segmentation checkpoint was used for the reported test values. Unimodal methods received only the optical input, whereas multimodal methods received the co-registered optical and DSM pair. For multimodal methods originally developed for another auxiliary sensor, only the auxiliary-input interface was adapted to accept the single-channel DSM; the published backbone and fusion architecture were retained. Each ablation configuration was independently trained and evaluated using five fixed random seeds, $\{0, 1, 2, 3, 42\}$. The seed set was specified before aggregation; seed 42 retains the original experiment, while

the other four seeds provide distinct initialization and data-sampling states. The ablation experiments report the mean and standard deviation obtained over these five repeated runs. The source code and implementation details of the proposed CAF-Net are provided in the Supplementary Materials to facilitate reproducibility.

Evaluation Metrics. Model performance is evaluated using standard metrics for semantic segmentation, including Overall Accuracy (OA), mean Intersection over Union (mIoU), mean F1-score (mF1), and the Kappa coefficient (κ). These metrics provide complementary assessments of classification accuracy, region overlap, and agreement consistency.

4.2. Comparison with the State of the Art

As shown in Table 1, CAF-Net achieves the best mIoU on both the Vaihingen and Potsdam datasets and also obtains the best mF1 on both benchmarks, while remaining competitive in OA and Kappa. Compared with the unimodal baseline SegFormer-B4, CAF-Net achieves clear improvements in mIoU, indicating that incorporating DSM-derived geometric information significantly enhances structural representation. Furthermore, compared with recent multimodal methods such as FDMF-Net, CAF-Net achieves superior performance, particularly on the Potsdam dataset. This improvement demonstrates that explicitly modeling spatial–frequency interactions is more effective than relying on static frequency manipulation or direct feature fusion. Overall, these results suggest that resolving spatial–frequency representation conflicts is a key factor for achieving robust and consistent multimodal segmentation.

Table 1. Quantitative results (%) on the ISPRS Vaihingen and ISPRS Potsdam datasets. All methods are retrained under identical protocols (see Section 4.1); the best results are highlighted in bold.

Type	Method	Backbone	Vaihingen				Potsdam			
			OA	Kappa	mF1	mIoU	OA	Kappa	mF1	mIoU
UniModal	SegFormer [21]	SegFormer-B4	90.95	88.10	89.78	82.42	90.87	87.24	91.89	84.90
	TransUNet [23]	R50-ViT-B	91.75	88.82	90.69	83.26	90.96	88.03	92.16	85.71
	UNetFormer [24]	ResNet-18	91.00	87.92	89.54	81.60	90.10	86.89	91.23	84.25
	RS ³ Mamba [8]	R18-Mamba-T	91.48	88.57	90.22	82.70	90.00	86.88	91.22	84.32
	PyramidMamba [25]	ResNet-18	88.10	83.92	85.74	75.54	85.15	80.45	86.62	76.74
	SFFNet [5]	ConvNext-T	90.88	87.66	89.40	81.12	89.25	85.80	90.58	83.05
	AFENet [26]	ResNet-18	91.62	88.75	90.40	82.95	90.62	87.69	91.80	85.30
MultiModal	MCANet [27]	ResNet-101	89.84	86.27	87.42	78.11	88.58	84.94	89.43	81.14
	CMFNet [28]	VGG-16	91.14	88.02	89.67	81.60	89.79	86.52	90.87	83.52
	CMX [6]	SegFormer-B2	92.01	89.31	90.76	83.57	91.10	88.31	92.22	86.01
	MultiSenseSeg [29]	ResNet-34	91.80	88.91	89.79	81.86	89.19	85.74	90.60	83.18
	FTransUNet [30]	R50-ViT-B	92.25	89.52	90.78	83.45	90.76	87.81	91.73	84.95
	MANet [31]	ViT-L	92.62	90.06	90.66	83.34	90.42	87.40	91.61	84.84
	ASANet [7]	ConvNeXtV2-T	90.21	86.77	87.07	77.67	89.04	85.50	89.96	82.18
	AMDANet [32]	SegFormer-B2	88.52	83.52	85.54	76.08	87.65	83.85	88.12	78.76
	MRFS [33]	SegFormer-B4	91.94	89.10	89.95	83.83	91.26	88.43	92.03	85.82
	FDMF-Net [34]	SegFormer-B4	92.07	89.32	91.04	83.94	91.64	89.00	92.45	86.24
	CAF-Net (Ours)	SegFormer-B4	92.51	89.93	91.28	84.35	91.87	89.15	92.79	86.86

Table 2 presents the category-wise IoU comparison. CAF-Net achieves strong performance across most categories, especially for buildings and roads, where accurate structural modeling is critical. The improvement on buildings reflects enhanced boundary preservation, while the gain on roads indicates better sensitivity to fine-scale structures. For spectrally ambiguous categories such as low vegetation, CAF-Net maintains competitive performance, benefiting from uncertainty-aware fusion in challenging regions. Although the improvement on cars is relatively moderate, the overall results demonstrate that jointly

modeling spatial and frequency representations leads to more balanced and reliable predictions across diverse semantic categories.

Table 2. Category-wise comparison on the ISPRS Vaihingen and Potsdam datasets. The best results are highlighted in bold.

Method	Backbone	ISPRS Vaihingen (IoU)					ISPRS Potsdam (IoU)				
		Roads	Build.	Low veg.	Trees	Cars	Roads	Build.	Low veg.	Trees	Cars
MCANet [27]	ResNet-101	78.68	89.35	61.45	79.32	81.75	81.52	90.45	70.68	72.35	90.70
CMFNet [28]	VGG-16	83.18	92.82	64.65	82.38	84.97	83.55	91.72	72.68	76.45	93.20
CMX [6]	SegFormer-B2	86.85	94.39	67.78	83.72	85.11	87.62	95.15	75.68	77.65	93.95
FTransUNet [30]	R50-ViT-B	86.18	94.25	67.35	83.32	86.15	86.35	94.22	75.15	76.38	92.65
MANet [31]	ViT-L	86.35	93.82	67.72	83.58	85.23	86.68	93.95	75.35	76.78	91.44
ASANet [7]	ConvNeXtV2-T	78.82	89.92	61.95	80.28	77.38	81.95	90.88	71.35	73.18	93.54
AMDANet [32]	SegFormer-B2	76.45	88.92	60.18	77.65	77.20	80.45	89.55	68.42	75.73	79.65
MRFS [33]	SegFormer-B4	87.07	94.35	68.85	83.92	84.95	87.25	94.95	77.52	77.05	92.33
FDMF-Net [34]	SegFormer-B4	87.15	94.45	68.95	84.05	85.12	87.57	95.21	77.99	77.90	92.55
CAF-Net (Ours)	SegFormer-B4	88.33	95.43	69.81	84.49	83.72	88.45	95.90	77.88	77.95	94.13

4.3. Qualitative Results

Figure 3 presents qualitative comparisons on the ISPRS Vaihingen and Potsdam datasets. To make local differences easier to inspect, one representative region is enlarged for each scene, with identical crop coordinates used for the GT and every compared method. The enlarged crops show that several existing methods produce less sharp boundaries, locally fragmented regions, or isolated misclassified pixels in the selected challenging areas. Multimodal methods such as FDMF-Net alleviate some of these errors but may still exhibit incomplete structures or local confusion. In comparison, CAF-Net generally produces predictions that are more spatially coherent and closer to the displayed GT boundaries. The car-dominated Potsdam example also shows fewer spurious responses around the selected small-object region. These visual results provide qualitative support for improved boundary consistency and local prediction stability, while the quantitative comparisons provide the primary performance evidence.

These improvements can be attributed to the coordinated design of CAF-Net. CMSG enhances spatial consistency by reinforcing geometric structures, ACFM improves global semantic alignment in the frequency domain, and UAF dynamically resolves inconsistencies between modalities. As a result, CAF-Net achieves more consistent predictions across both spatial and frequency domains, leading to robust performance in challenging multimodal scenarios.

4.4. Computational Complexity Analysis

Table 3 presents the comparison of model complexity and segmentation performance on the Potsdam dataset. CAF-Net achieves the best mIoU among all compared methods while maintaining a moderate computational cost, demonstrating a favorable balance between accuracy and efficiency. Compared with recent multimodal fusion methods using the same backbone, CAF-Net achieves higher segmentation accuracy with reduced computational overhead. In particular, CAF-Net reduces the parameter count by approximately 28M and lowers FLOPs by 15.9% compared with FDMF-Net, while still improving mIoU. This indicates that explicitly modeling spatial–frequency interactions enables more efficient feature utilization than direct feature fusion or static frequency modeling. Compared with CNN-based methods, CAF-Net also achieves competitive efficiency. Despite using a transformer-based backbone, the model maintains a relatively low computational cost, benefiting from lightweight operations such as depth-wise separable convolutions and adaptive feature selection. This design avoids unnecessary redundancy while preserving strong representation capability. Overall, these results demonstrate that CAF-Net achieves

a superior accuracy–efficiency trade-off. By resolving spatial–frequency inconsistencies in a structured manner, the model not only improves segmentation performance but also reduces redundant computation, making it more suitable for practical remote sensing applications.

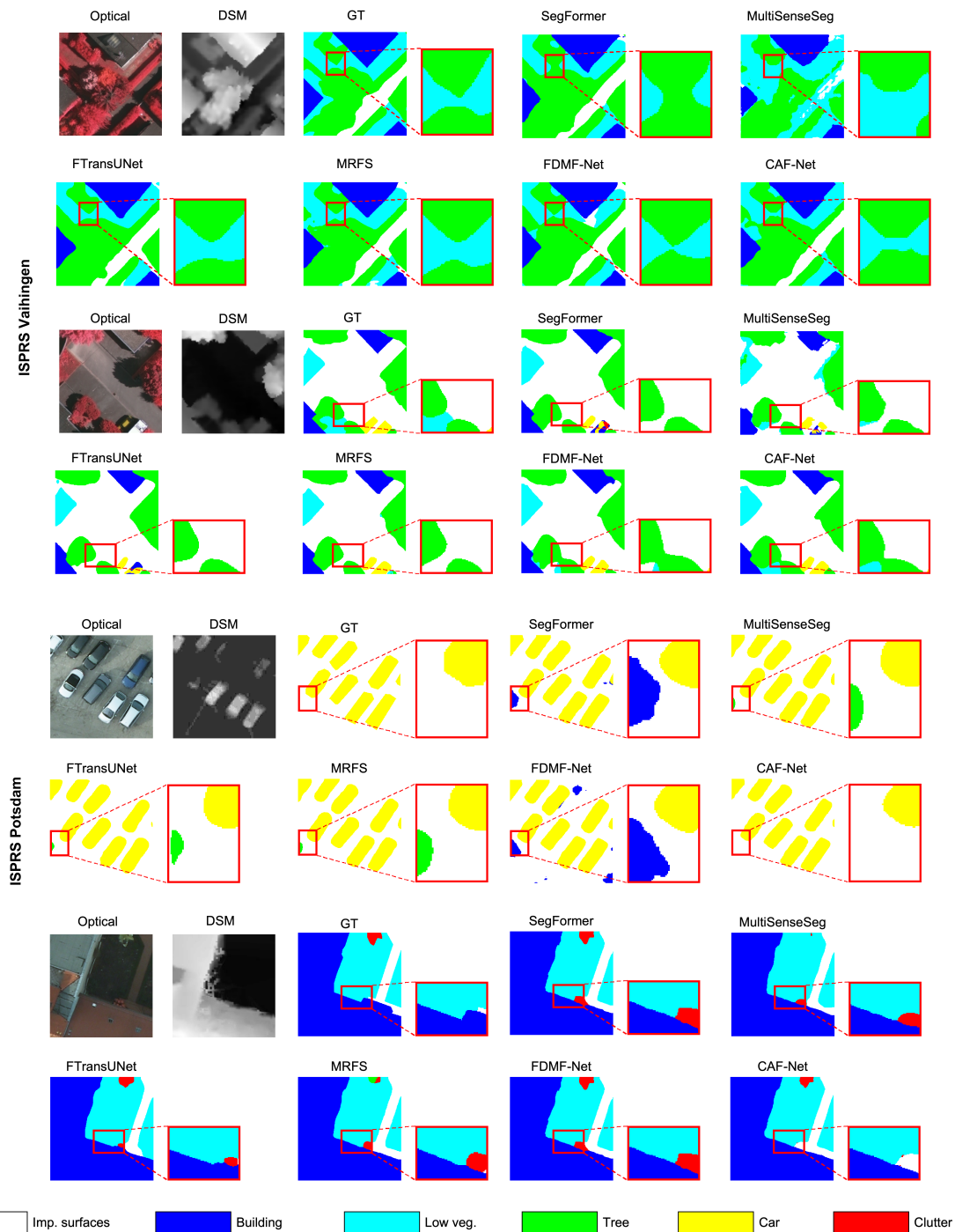


Figure 3. Qualitative comparison of semantic segmentation results on the ISPRS Vaihingen and Potsdam datasets. From left to right: optical image, DSM, ground truth (GT), and predictions from SegFormer, MultiSenseSeg, FTransUNet, MRFS, FDMF-Net, and the proposed CAF-Net. For each scene, the same representative region is marked across the GT and all prediction maps, and the corresponding enlarged crop is connected to the marked region by red dashed lines. These crops highlight representative errors involving boundary ambiguity, small-object confusion, and locally fragmented predictions.

Table 3. Model scale analysis with a 256×256 input on the Potsdam dataset. The best results are highlighted in bold.

Model	Backbone	mIoU (%)	Params (M)	FLOPs (G)
MCANet [27]	ResNet-101	81.14	70.62	141.26
CMFNet [28]	VGG-16	83.52	104.07	156.52
CMX [6]	SegFormer-B2	86.01	66.56	28.53
MultiSenseSeg [29]	ResNet-34	83.18	57.73	132.14
FTransUNet [30]	R50-ViT-B	84.95	203.40	40.61
MANet [31]	ViT-L	84.84	317.25	12.89
ASANet [7]	ConvNeXtV2-T	82.18	82.32	57.73
AMDANet [32]	SegFormer-B2	78.76	135.73	35.69
MRFS [33]	SegFormer-B4	85.82	134.96	59.35
FDMF-Net [34]	SegFormer-B4	86.24	156.08	66.15
CAF-Net (Ours)	SegFormer-B4	86.86	128.03	55.64

4.5. Ablation Study

To evaluate the contribution of each component, we conduct a comprehensive ablation study, as shown in Table 4. Starting from the baseline, introducing CMSG yields the largest individual improvement, increasing mIoU from 82.35% to 83.60% on Vaihingen and from 84.76% to 86.00% on Potsdam. This result highlights the effectiveness of DSM-derived structural guidance in enhancing boundary delineation and spatial consistency. ACFM also consistently improves segmentation performance, demonstrating that frequency-domain modeling provides complementary semantic information beyond conventional spatial representations.

When applied independently, UAF yields relatively modest gains compared with CMSG and ACFM. The combination of CMSG and ACFM produces a stronger result than either frequency or structure guidance alone, but residual inconsistencies may still exist between spatially guided and frequency-enhanced representations. By further incorporating UAF, the full model achieves the best overall performance in Table 4, reaching 84.35% mIoU on Vaihingen and 86.86% mIoU on Potsdam.

These results indicate that the primary role of UAF is not merely to improve segmentation accuracy in isolation, but to facilitate effective interaction between structure-guided and frequency-aligned features through reliability-aware fusion. Overall, the full model achieves the best performance, demonstrating that jointly modeling structural guidance, frequency alignment, and adaptive fusion is essential for robust multimodal remote sensing segmentation.

Table 4. Comprehensive ablation study on the ISPRS Vaihingen and Potsdam datasets (%). The results are reported as mean $\pm$ std. “✓” indicates that the module is enabled. **Base** denotes the SegFormer [21] modified into a dual-branch architecture. The best results are highlighted in bold.

Modules				Vaihingen		Class-Specific IoU (Vaihingen)					Potsdam		Class-Specific IoU (Potsdam)				
Base	CMSG	ACFM	UAF	mIoU	Roads	Build.	Low veg.	Trees	Cars		mIoU	Roads	Build.	Low veg.	Trees	Cars	
✓				82.35 $\pm$ 0.28	85.41	93.34	65.63	83.68	83.71		84.76 $\pm$ 0.23	86.64	93.69	74.03	76.12	93.31	
✓	✓			83.60 $\pm$ 0.15	87.01	94.14	68.15	84.48	84.22		86.00 $\pm$ 0.17	87.04	94.19	76.69	78.05	93.81	
✓		✓		83.43 $\pm$ 0.16	86.67	94.45	67.94	84.34	83.76		85.87 $\pm$ 0.14	87.03	94.02	77.02	77.57	93.70	
✓			✓	83.33 $\pm$ 0.18	86.48	93.93	68.68	84.08	83.45		85.35 $\pm$ 0.19	86.75	93.85	75.45	76.88	93.82	
✓	✓	✓		83.87 $\pm$ 0.24	87.77	94.91	69.04	83.24	84.39		86.35 $\pm$ 0.10	87.62	95.06	77.01	78.05	94.01	
✓	✓	✓	✓	84.35 $\pm$ 0.10	88.33	95.43	69.81	84.49	83.72		86.86 $\pm$ 0.08	88.45	95.90	77.88	77.95	94.13	

5. Discussion

5.1. Robustness Under Modality Degradation

As shown in Table 5, the Full model achieves the highest clean performance and remains the best-performing variant under all evaluated degradations. Optical noise produces the largest reduction for the Baseline model, whereas the w/o UAF variant shows much stronger sensitivity to DSM-related degradations. Under optical noise, the Full model retains 67.10% mIoU. Under local DSM missing regions and DSM elevation noise, the Full model reaches 82.58% and 83.50% mIoU, respectively, corresponding to relatively small decreases from its clean result. Under DSM–optical misalignment, the Baseline decreases from 82.35% to 81.09%, whereas the Full model retains 84.04% mIoU, indicating lower sensitivity to the simulated spatial displacement.

Table 5. Performance and robustness comparison on the ISPRS Vaihingen dataset under different degradation conditions (mIoU, %). Optical Noise denotes Gaussian noise added to optical images ($\sigma = 0.1$). DSM Noise denotes Gaussian perturbation of the normalized DSM ($\sigma = 0.05$). DSM Misalignment applies a random horizontal and vertical displacement of at most 4 pixels. DSM Local Missing replaces randomly sampled 32×32 DSM regions covering approximately 30% of each input with zeros, while DSM Missing replaces the complete DSM input with zeros. w/o UAF denotes the model without Uncertainty-Aware Fusion, and Full denotes CAF-Net. The best results are highlighted in bold. The symbol $\uparrow$ indicates higher performance is better.

Model	Clean $\uparrow$	DSM Local Missing $\uparrow$	DSM Misalignment $\uparrow$	DSM Missing $\uparrow$	DSM Noise $\uparrow$	Optical Noise $\uparrow$
Baseline	82.35	81.45	81.09	78.13	79.70	47.39
w/o UAF	83.87	22.99	23.39	23.83	23.47	51.30
Full	84.35	82.58	84.04	80.00	83.50	67.10

The w/o UAF variant is highly sensitive to every DSM degradation, falling to approximately 23% mIoU under local missing regions, misalignment, complete missingness, and elevation noise. The similar values under these perturbations do not indicate that the perturbations are identical; rather, they suggest that the model enters a similar failure regime once DSM-derived guidance becomes unreliable. In the w/o UAF variant, DSM features are still actively used by CMSG and ACFM, but no reliability-aware mechanism is available to reduce their contribution. Therefore, corrupted DSM responses are not simply ignored; instead, they are injected into the optical representation as misleading structural and frequency-domain guidance. Local missing regions introduce discontinuous zero-valued structures, misalignment destroys pixel-level optical–DSM correspondence, complete missingness produces an out-of-distribution constant DSM input, and elevation noise introduces spurious high-frequency responses. Although these perturbations differ in form, they all invalidate the DSM-guided modulation and frequency-alignment pathways, leading to a common low-performance regime with mIoU values ranging from 22.99% to 23.83%. To simulate complete modality failure, the DSM input is replaced with an all-zero tensor during inference, while all models are trained using complete optical–DSM pairs. For consistency, all degraded evaluations use the same trained checkpoint and sliding-window inference protocol as the clean evaluation, and the perturbations are applied only at inference.

Under complete DSM missingness, the Full model retains 80.00% mIoU, compared with 23.83% for w/o UAF. Taken together, these experiments provide broader evidence that relative-reliability weighting reduces sensitivity to several forms of modality degradation. Nevertheless, these tests represent controlled synthetic perturbations and do not cover every source of variability encountered in operational remote sensing.

5.2. Conflict Analysis

To further analyze the relationship between feature discrepancy and model behavior, we report the conflict metrics defined in Section 3.6. The results are summarized in Table 6.

The baseline model exhibits relatively high feature conflict, indicating a noticeable discrepancy between spatial and frequency representations. When UAF is removed, the measured conflict increases, with Global Conflict rising from 0.958 to 1.997. This indicates that without adaptive fusion, the interaction between spatial and frequency features becomes less aligned. In contrast, the Full model shows lower conflict values, achieving a Global Conflict of 0.328 and a Boundary Conflict of 0.200. This suggests reduced feature-level discrepancy under the proposed fusion strategy. We further observe that the trend of conflict values is consistent with the performance changes reported under noise and modality degradation settings. This consistency suggests an association between feature-level discrepancy and segmentation performance under challenging conditions. Overall, these results provide a diagnostic analysis of feature consistency across different model variants, rather than serving as causal evidence of model behavior.

Table 6. Feature conflict analysis. Lower values indicate better cross-domain consistency. The best results are highlighted in bold. The symbol ↓ denotes lower values correspond to better performance.

Model	Global Conflict ↓	Boundary Conflict ↓
Baseline	0.958	0.661
w/o UAF	1.997	1.379
Full	0.328	0.200

5.3. Mechanism Interpretability and Qualitative Analysis

To analyze the internal mechanics of the proposed framework, we visualize the intermediate feature states of the three core modules in Figure 4. This visualization tracks the progressive refinement of spatial structures, modality fusion weights, and frequency domain distributions.

The first row of Figure 4 illustrates the CMSG branch. The optical input contains urban textures with ambiguous transitions between the building roof and its surroundings. The DSM-derived structural map emphasizes elevation discontinuities near the object boundary. The learned attention response is also concentrated around parts of these discontinuities. Compared with the feature before CMSG, the refined representation shows more localized responses near the depicted boundary. This qualitative observation is consistent with the intended role of DSM-guided structural modulation, but does not by itself establish a causal explanation.

The second row illustrates the pixel-level weighting strategy of UAF for a scene containing a building shadow. The optical reliability map assigns a comparatively lower response to part of the shadowed region, whereas the DSM reliability map gives a higher response in the corresponding area. The fused representation retains a more coherent activation pattern over the displayed structure. This example suggests that relative-reliability weighting can reduce the contribution of locally degraded modality features; broader quantitative evidence is provided by the degradation experiments.

The third row visualizes frequency responses before and after ACFM. The displayed spectrum changes from a more dispersed response to a distribution with stronger central concentration after adaptive reweighting and alignment. The reconstructed spatial feature also shows a more spatially coherent response over the depicted object. These observations are consistent with frequency selection and cross-modal alignment, although the visualization should be interpreted together with the quantitative ablation and conflict analyses.

Together, these stage-wise visualizations illustrate how CAF-Net differs from direct concatenation by combining spatial structure modulation, relative-reliability weighting, and frequency alignment. They are presented as qualitative support rather than direct proof of the proposed mechanism.

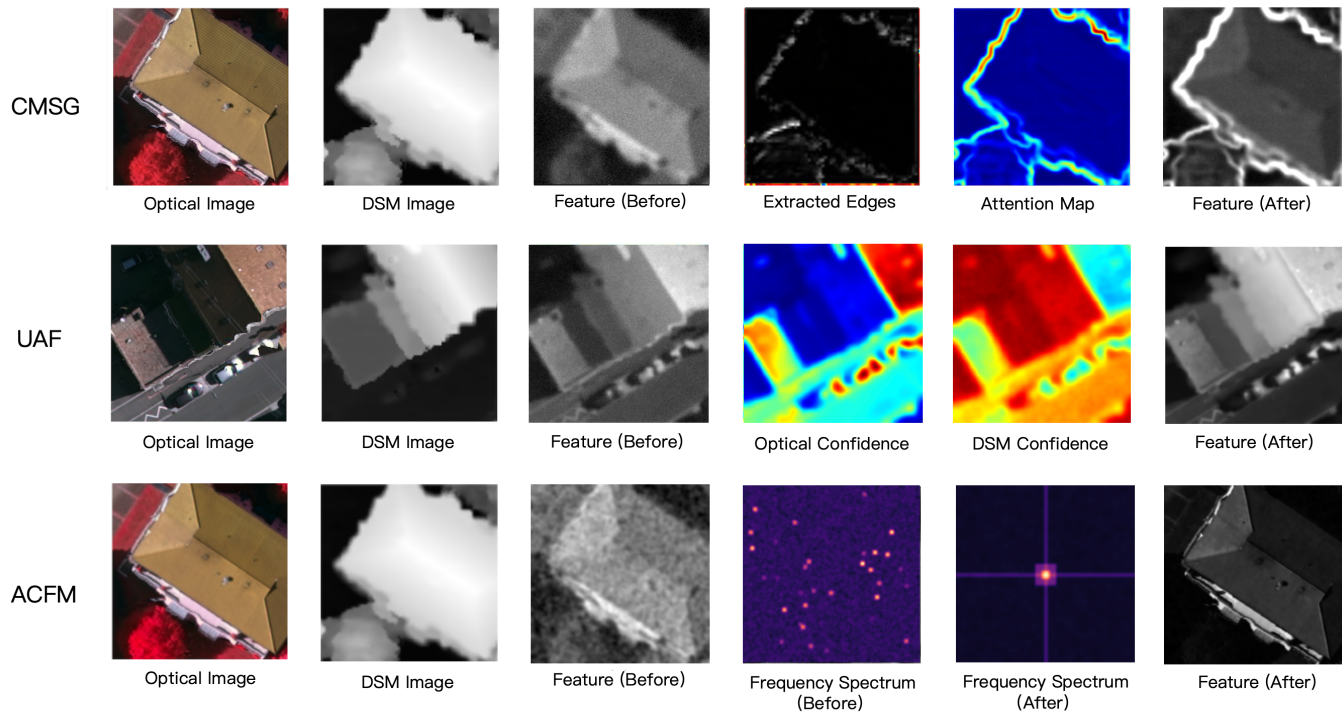


Figure 4. Multi-dimensional visualization of the internal mechanisms of CAF-Net.

5.4. Sensitivity Analysis of Temperature Coefficient

To evaluate the influence of the temperature coefficient T in UAF, we conduct experiments on both the Vaihingen and Potsdam datasets. The temperature parameter controls the sharpness of the Softmax distribution, thereby affecting the strength of modality competition during fusion. As shown in Table 7, the performance remains stable across a wide range of temperature values. Variations in T lead to only minor changes in mIoU and F1-score, indicating that the proposed framework is not sensitive to this hyperparameter. This suggests that the reliability-aware fusion mechanism can adaptively balance modality contributions without requiring precise tuning. Notably, $T = 1.0$ achieves the best Kappa and OA on both datasets, while maintaining competitive performance in mIoU and F1-score. Smaller values of T lead to slightly sharper weighting, whereas larger values produce smoother distributions, but neither significantly degrades performance.

Overall, these results demonstrate that UAF is robust to the choice of temperature coefficient, and the proposed fusion strategy maintains stable performance across different settings. Therefore, $T = 1.0$ is adopted as the default configuration in all experiments.

Table 7. Effect of temperature coefficient T in UAF on Vaihingen and Potsdam datasets. The best results are highlighted in bold.

T	Vaihingen				Potsdam			
	Kappa	OA	F1	mIoU	Kappa	OA	F1	mIoU
0.5	89.45	92.15	91.32	84.41	88.81	91.51	92.85	86.92
1.0	89.93	92.51	91.28	84.35	89.29	91.87	92.79	86.86
1.5	89.85	92.46	91.29	84.37	89.21	91.82	92.81	86.88
2.0	89.52	92.22	91.21	84.25	88.88	91.58	92.69	86.76

6. Conclusions

In this paper, we present CAF-Net, a unified framework for optical–DSM semantic segmentation that explicitly addresses spatial–frequency representation conflicts in multimodal learning. The proposed architecture replaces standard feature aggregation with a coordinated conflict resolution mechanism. Specifically, the CMSG component extracts elevation priors to correct spatial boundary ambiguities. The ACFM aligns high-frequency structural details with low-frequency global semantics. These modules are complemented by the UAF, which computes pixel-wise reliability maps to prevent the propagation of degraded data during the fusion process.

Extensive experiments on the ISPRS Vaihingen and Potsdam datasets demonstrate that CAF-Net achieves superior performance, reaching 84.35% and 86.86% mIoU, respectively. Beyond standard accuracy improvements, computational complexity analysis suggests that the framework maintains a highly favorable accuracy–efficiency trade-off, requiring fewer parameters and floating-point operations than recent multimodal baselines. Quantitative evaluations based on the proposed conflict measurement metric show a substantial reduction in cross-domain inconsistencies, particularly along sensitive object boundaries. Controlled robustness tests further show improved performance retention under optical noise, DSM elevation noise, local and complete DSM missingness, and optical–DSM misalignment.

Overall, this work demonstrates that resolving structural and semantic conflicts, rather than solely increasing feature extraction capacity, is an important factor for effective multimodal integration. This insight provides a useful perspective for designing multimodal segmentation models and highlights the potential benefits of jointly modeling spatial and frequency representations. Future work will evaluate CAF-Net under naturally occurring registration errors, sensor artifacts, seasonal changes, and cross-domain scenarios and extend the framework to more heterogeneous multimodal settings such as SAR and LiDAR data.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/rs18142284/s1>, The source code and implementation details of the proposed CAF-Net.

Author Contributions: Conceptualization, M.M. and S.W.; methodology, S.W.; software, S.W.; validation, X.L.; formal analysis, S.W. and X.L.; investigation, S.W.; data curation, X.L.; writing—original draft preparation, S.W.; writing—review and editing, M.M. and H.W.; supervision, M.M. and H.W.; project administration, H.W. All authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Inner Mongolia, China (Grant No. 2026MS0228), and the National Natural Science Foundation of China (Grant No. 62462049).

Data Availability Statement: The ISPRS Potsdam and Vaihingen datasets used in this study are openly available from the International Society for Photogrammetry and Remote Sensing (ISPRS). The Potsdam dataset is available at <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>, and the Vaihingen dataset is available at <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx> (both accessed on 3 May 2026).

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

References

1. Wang, M.; Su, L.; Yan, C.; Xu, S.; Yuan, P.; Jiang, X.; Zhang, B. Rsbuilding: Toward general remote sensing image building extraction and change detection with foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17.
2. Bountos, N.I.; Ouaknine, A.; Papoutsis, I.; Rolnick, D. Fomo: Multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring. *arXiv* **2025**, arXiv:2312.10114.
3. Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; Chanussot, J. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102926.
4. Zhou, G.; Lihuang, Q.; Gamba, P. Advances on multimodal remote sensing foundation models for Earth observation downstream tasks: A survey. *Remote Sens.* **2025**, *17*, 3532.
5. Yang, Y.; Yuan, G.; Li, J. SFFNet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17.
6. Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; Stiefelhausen, R. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14679–14694.
7. Zhang, P.; Peng, B.; Lu, C.; Huang, Q.; Liu, D. ASANet: Asymmetric semantic aligning network for RGB and SAR image land cover classification. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 574–587.
8. Ma, X.; Zhang, X.; Pun, M.O. RS3Mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6011405.
9. Liu, X.; Wang, T.; Jin, F.; Rui, J.; Wang, S.; Huang, Z.; Zou, Y.; Yu, X. Multimodal cross fusion Mamba network for remote sensing image semantic segmentation with complementary masked self-supervision. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *145*, 104960.
10. Li, P.; Chen, Z.; Long, C.; Zhang, H.; Lv, Y.; Huang, R.; Dong, Z.; Yang, B. A novel projection map driven multimodal fusion framework for ALS point cloud semantic segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104907.
11. Hosseinpour, H.; Samadzadegan, F.; Javan, F.D. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 96–115.
12. Huang, X.; Li, H.; Cao, M.; Chen, L.; You, C.; An, D. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Trans. Med. Imaging* **2024**, *44*, 1821–1835.
13. Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; Zhou, J. Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 980–993.
14. Li, X.; Xu, F.; Yu, A.; Lyu, X.; Gao, H.; Zhou, J. A frequency decoupling network for semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–21.
15. Yao, T.; Pan, Y.; Li, Y.; Ngo, C.W.; Mei, T. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 328–345.
16. Dai, T.; Wang, J.; Guo, H.; Li, J.; Wang, J.; Zhu, Z. FreqFormer: Frequency-aware Transformer for Lightweight Image Super-resolution. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Main Track*, Jeju, Republic of Korea, 3–9 August 2024; pp. 731–739.
17. Wang, Q.; Li, Z.; Zhang, S.; Chi, N.; Dai, Q. WaveFusion: A novel wavelet vision transformer with saliency-guided enhancement for multimodal image fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 7526–7542.
18. Zhang, Y.; Ma, Y.; Zhao, F.; Lai, R.; Yan, T.; Liu, T. FAENet: Autoencoding-driven Feature Adaptation for Multi-class Brain Tissue Segmentation. In *Proceedings of the 2025 19th International Conference on Complex Medical Engineering (CME)*; IEEE: New York, NY, USA, 2025; pp. 343–347.
19. Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; Catanzaro, B. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv* **2021**, arXiv:2111.13587.
20. Zhou, Y.; Liu, G.; Yang, Y.; Ru, L.; Liu, D.; Li, X. UFCNet: Unsupervised network based on Fourier transform and convolutional attention for oracle character recognition. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, Bangkok, Thailand, 15 August 2024; pp. 98–106.
21. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.

22. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D. *ISPRS Semantic Labeling Contest*; ISPRS: Leopoldshöhe, Germany, 2014; Volume 1, p. 4.
23. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
24. 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.
25. Wang, L.; Li, D.; Dong, S.; Meng, X.; Zhang, X.; Hong, D. PyramidMamba: Rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104884.
26. Gao, F.; Fu, M.; Cao, J.; Dong, J.; Du, Q. Adaptive frequency enhancement network for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5619415.
27. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638.
28. Ma, X.; Zhang, X.; Pun, M.O. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3463–3474.
29. Wang, Q.; Chen, W.; Huang, Z.; Tang, H.; Yang, L. MultiSenseSeg: A cost-effective unified multimodal semantic segmentation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–24.
30. Ma, X.; Zhang, X.; Pun, M.O.; Liu, M. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15.
31. Ma, X.; Zhang, X.; Pun, M.O.; Huang, B. A unified framework with multimodal fine-tuning for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5405015.
32. Zhong, H.; Tang, F.; Chen, Z.; Chang, H.J.; Gao, Y. AMDANet: Attention-Driven Multi-Perspective Discrepancy Alignment for RGB-Infrared Image Fusion and Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 10645–10655.
33. Zhang, H.; Zuo, X.; Jiang, J.; Guo, C.; Ma, J. Mrfs: Mutually reinforcing image fusion and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle WA, USA, 17–21 June 2024; pp. 26974–26983.
34. Chen, G.; Sun, F.; Jing, W.; Zou, W.; Di, D.; Song, Y.; Fan, L. Learning Frequency-Domain Fusion for Multimodal Remote Sensing Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4420216.

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.