

Fine-Grained Remote Sensing Semantic Segmentation via Optical–DSM Collaborative Enhancement and Frequency-Guided Decoding

Xuebo Jin

jinxuebo@btbu.edu.cn

Beijing Technology and Business University

Zexin Zhai

Beijing Technology and Business University

Jianlei Kong

Beijing Technology and Business University

Tingli Su

Beijing Technology and Business University

Yuting Bai

Beijing Technology and Business University

Research Article

Keywords: Remote sensing imagery, Semantic segmentation, Optical–DSM bimodal fusion, Hierarchical representation, Frequency-guided decoding, Fine-grained segmentation

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11470010/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Fine-Grained Remote Sensing Semantic Segmentation via Optical–DSM Collaborative Enhancement and Frequency-Guided Decoding

Xuebo Jin^{1,2}, Zexin Zhai¹, Jianlei Kong^{1,3}, Tingli Su¹, Yuting Bai¹

1. School of Computer and Artificial Intelligence, Beijing Technology and Business University,
Beijing 100048, China;

2. School of Computer and Artificial Intelligence, Xinjiang Hetian College, Hetian, 848000, China;

3. National Engineering Research Center for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing 100048, China;

Abstract Semantic segmentation of remote sensing imagery remains challenging for fine-grained structures such as vehicles, narrow roads, building contours, and vegetation boundaries, where discriminative evidence is limited and boundary blurring, small-object omission, and confusion between spectrally similar classes are common. Optical imagery provides spectral and textural cues, whereas digital surface models (DSMs) encode height and geometric structure. However, their heterogeneous distributions make simple concatenation or single-stage fusion insufficient, while conventional decoder upsampling may further introduce response diffusion, cross-scale misalignment, and boundary-detail attenuation. We propose an optical–DSM collaborative enhancement and frequency-guided decoding framework for remote sensing semantic segmentation. The network first enhances complementary optical texture and DSM geometry through modality-specific encoding, bidirectional cross-modal channel guidance, multiscale structural perception, and dense feature integration. Hierarchical bimodal interaction is then organized according to feature abstraction: channel recalibration and dual-branch fusion propagate local structure in shallow and middle layers, whereas self-attention and cross-modal attention coordinate long-range intra-modal dependencies and cross-modal class semantics in deep layers. Finally, low-frequency region-consistency reconstruction, similarity-guided offset resampling, and high-frequency detail compensation are integrated into multistage decoding. On Vaihingen and Potsdam, the proposed model achieves mean F1 (mF1) scores of 90.91% and 92.37% and mean intersection over union (mIoU) scores of 83.74% and 86.07%, respectively. Comparative, ablation, and qualitative results demonstrate complementary gains from input-level bimodal enhancement, hierarchical cross-modal interaction, and frequency-guided decoding, particularly for complex structures, class boundaries, and small objects.

Keywords Remote sensing imagery; Semantic segmentation; Optical–DSM bimodal fusion; Hierarchical representation; Frequency-guided decoding; Fine-grained segmentation

1 Introduction

Remote sensing semantic segmentation assigns a land-cover label to each pixel and provides fundamental information for applications such as urban object extraction, land-use surveys, and disaster monitoring [1]. With the widespread adoption of deep learning in remote sensing image understanding, pixel-level scene interpretation has gradually shifted from hand-crafted features to end-to-end representation learning, accompanied by rapid advances in model architectures and application scope [2]. Nevertheless, fine-grained regions in complex urban scenes—such as vehicles, narrow roads, building contours, and vegetation boundaries—often contain only limited discriminative evidence and are vulnerable to scale variation, spectral similarity, shadows, and cluttered backgrounds. These factors can lead to blurred boundaries, missed small objects, fragmented regions, and confusion between adjacent classes. Small objects in remote sensing imagery often occupy only a limited number of pixels, making their weak foreground features vulnerable to background interference, feature attenuation, and repeated downsampling [3, 4]. Accurate recognition of these objects therefore requires not only high-level semantic representations but also effective preservation of local spatial evidence.

For detailed land-cover interpretation, some targets and boundary regions provide only limited spatial evidence. Strengthening the encoder's global modeling capability can improve scene-level semantic understanding, but it cannot fully compensate for insufficient local texture, geometric contours, and small-scale structural cues. This is particularly problematic for spectrally or texturally similar classes, such as buildings versus impervious surfaces and trees versus low vegetation, for

which a single optical modality may not provide stable structural criteria. Therefore, introducing complementary geometric evidence while preserving local structural information remains an important problem in remote sensing semantic segmentation.

A digital surface model (DSM) records height variation, surface relief, and geometric contours, complementing the spectral, color, and textural information provided by optical imagery. In recent years, optical–DSM segmentation has evolved from input-level concatenation toward dual-branch encoding, multiscale fusion, and cross-layer semantic interaction. MMA-Net [5], introduced in 2025, coordinates features at different levels through multimodal fusion and multiscale multi-attention mechanisms; MMFNet-MLFL [3] enhances small-object representations with limited pixel coverage through multilevel feature learning; MSFFNet [6] organizes RGB–DSM fusion jointly in the spatial and frequency domains; and the Mamba-based MMFNet [7] combines convolutional local modeling with state-space global modeling to strengthen bimodal representations. More recently, BATFNet [8] uses DSM structural and boundary priors to guide bidirectional cross-modal attention, while DLGCNet [9] further enhances multimodal interaction through graph-convolutional feature fusion and visual foundation-model adaptation. These studies indicate that multiscale structure preservation, cross-layer information propagation, and boundary awareness have become important directions for optical–DSM remote sensing segmentation.

Although existing methods fuse optical imagery and DSM at multiple encoding levels, the two modalities differ substantially in statistical distribution and physical meaning, and a uniform fusion strategy across all levels may not fully exploit their complementarity. Shallow features focus mainly on edges, textures, and local geometric structures, and premature complete mixing may weaken modality-specific elevation information in DSM. As encoding deepens, both modalities gradually form class semantics, for which simple concatenation or element-wise addition is insufficient to establish long-range intra-modal dependencies and cross-modal class relationships. Recent multimodal approaches have explored multiscale, cross-layer, and boundary-aware fusion, but how to organize input-level enhancement, shallow-to-middle structural propagation, and deep semantic coordination according to representation depth still warrants further investigation.

In addition, the enhanced bimodal encoding features must be progressively reconstructed in the pixel prediction space through multistage upsampling. Bilinear interpolation and transposed convolution mainly change spatial resolution and cannot adapt neighborhood contributions to local content; directly concatenating high-level semantics with shallow details also fails to explicitly address response diffusion within class regions, spatial correspondence errors across scales, and attenuation of high-frequency boundary information. FreqFusion reorganizes cross-scale feature fusion from three perspectives—low-frequency intra-class consistency, similarity-guided resampling, and high-frequency detail enhancement—providing a frequency-domain perspective for improving region continuity and boundary localization during decoding [10]. MSFFNet [6] likewise shows that jointly modeling spatial and frequency information benefits structural detail representation in RGB–DSM scenes. However, existing frequency-aware methods typically focus on a single fusion stage or general spatial–frequency interaction. How to connect frequency-aware processing with hierarchical bimodal representations and organize it as a continuous multistage pixel reconstruction process remains underexplored.

To address these issues, we propose a collaborative optical–DSM enhancement and frequency-guided decoding framework for complex urban remote sensing scenes. The front end combines modality-specific encoding, bidirectional channel guidance, multiscale structural perception, and auxiliary reconstruction supervision to strengthen optical texture and DSM geometry. Hierarchical interaction then propagates local structure in shallow and middle layers and coordinates intra-modal and cross-modal semantics in deep layers. Finally, multistage decoding integrates low-frequency consistency reconstruction, similarity-guided offset resampling, and high-frequency detail compensation to improve region continuity, spatial correspondence, and boundary localization. The entire framework is optimized end-to-end for semantic segmentation. The main contributions of this work are as follows:

1. We propose a hierarchical optical–DSM framework that organizes bimodal interaction by representation depth, spanning input-level enhancement, shallow-to-middle structural propagation, and deep class-semantic coordination.
2. We design complementary interaction mechanisms for these stages: modality-specific encoding and bidirectional channel guidance at the input, controlled DSM propagation and dual-branch fusion in shallow and middle layers, and self-attention with cross-modal attention in deep layers. Together, they strengthen discrimination of small objects, structurally complex targets, and spectrally similar classes.
3. We develop a frequency-guided multistage decoder that reorganizes low-frequency region-consistency reconstruction, similarity-guided spatial correction, and high-frequency detail compensation into a progressive

decoding process for hierarchical optical–DSM representations, thereby improving region continuity, cross-scale correspondence, and boundary-detail preservation.

The remainder of this paper is organized as follows. Section 2 reviews related work on fine-grained remote sensing segmentation, optical–DSM collaborative representation, and frequency-aware feature reconstruction. Section 3 presents the proposed framework, including task-driven bimodal enhancement, hierarchical optical–DSM interaction, frequency-guided multistage decoding, and joint optimization. Section 4 describes the datasets, implementation details, comparative experiments, ablation studies, qualitative analysis, and limitations. Finally, Section 5 concludes the paper and outlines future research directions.

2 Related work

2.1 Fine-grained segmentation and task-driven representation enhancement

Convolutional architectures such as FCN [11], U-Net [12], DeepLabv3+ [13], and UNet++ [14] progressively improved pixel-level segmentation through fully convolutional prediction, skip connections, multiscale context modeling, and nested decoding. Subsequently, ViT [15] introduced Transformer-based visual representation learning, while SETR [16], SegFormer [17], and Swin Transformer [18] further strengthened long-range dependency modeling and hierarchical semantic representation. Despite these advances, fine-grained remote sensing structures remain difficult to segment accurately. Small objects often occupy only a limited number of pixels, while object contours and narrow structures are particularly vulnerable to repeated downsampling, background interference, spectral similarity, and scale variation [3, 4]. Stronger semantic encoders can improve contextual understanding, but local texture, geometric contours, and other spatial evidence relevant to pixel-level discrimination may still be progressively weakened during feature abstraction.

One way to alleviate the loss of local spatial evidence is to introduce reconstruction-oriented constraints that encourage intermediate representations to preserve task-relevant information. Super-resolution research has demonstrated the importance of recovering local texture and structural details in remote sensing imagery [19]. More importantly for semantic segmentation, joint reconstruction and segmentation methods show that reconstruction objectives can be optimized together with pixel-level classification [20], while task-driven restoration further indicates that the recovered information should benefit downstream recognition rather than merely improve visual quality [21]. These findings suggest that reconstruction can serve as an auxiliary representation constraint when it is guided by the downstream semantic objective.

However, reconstruction based only on optical imagery cannot provide structural information that is intrinsically absent or ambiguous in spectral observations. In urban remote sensing scenes, buildings and impervious surfaces, or trees and low vegetation, may share similar spectral and textural characteristics, and shadows and local contrast variations further increase uncertainty. DSM provides complementary height, relief, and geometric information that can assist local structural discrimination. Existing remote sensing segmentation studies have mainly focused on encoder design, multiscale context modeling, attention mechanisms, and global semantic representation [22], but they have less explored how auxiliary reconstruction and complementary geometric information can jointly enhance representations throughout the segmentation network. In this work, reconstruction is therefore used as a task-oriented auxiliary constraint at the original spatial resolution rather than as an image super-resolution objective, while DSM is introduced as complementary structural evidence for subsequent hierarchical bimodal representation learning.

2.2 Optical–DSM collaborative representation

Optical imagery mainly describes spectral, color, and textural properties, whereas DSM provides height variation, surface relief, and geometric contours. Previous studies have demonstrated the complementarity of optical and elevation information for urban land-cover segmentation [23]. The two modalities offer different advantages in separating buildings from roads, distinguishing vegetation levels, interpreting shadowed regions, and preserving small objects, but they also differ substantially in data distribution, noise characteristics, and physical meaning. Multimodal modeling should therefore not merely increase the number of input channels; it should choose appropriate interaction locations, information directions, and fusion strengths according to representation level.

Multimodal remote sensing segmentation has evolved from input-level or decision-level combination toward dual-branch encoding and stage-wise cross-modal interaction. Existing methods can be broadly organized into three technical

directions. First, multiscale and cross-level fusion methods progressively exchange complementary information across different representation levels. CMFNet [24] models cross-modal dependencies over multiscale optical and DSM features through cross-attention; FTransUNet [25] combines local structural fusion with Transformer-based semantic interaction across multiple levels; ASMFNet [26] strengthens multimodal context through information transfer between adjacent scales; and TMFNet [27] further supplements spatial details through cross-layer boundary interaction. Second, adaptive modality fusion focuses on dynamically regulating the contributions of different modalities. LMF-Net [28] introduces learnable fusion to adjust modality contributions rather than relying on fixed feature combination. More recently, hierarchical encoder–decoder collaboration extends multimodal interaction across multiple network stages. HDC-Net [29] combines shallow spatial alignment, deep cross-modal modeling, and decoder bridging; MFMINet [30] coordinates encoder–decoder representations through multimodal fusion and cross-layer interaction; and WAMNet [31] organizes cross-modal features using wavelet enhancement and asymmetric sequence modeling. These developments show that optical–DSM fusion has progressed from fixed-position feature combination toward adaptive and hierarchical interaction across multiple representation levels and network stages.

Despite this progress, different representation stages require different types of bimodal information. Shallow features mainly emphasize edges, textures, and local geometry, whereas deep features increasingly encode class semantics and long-range dependencies. Although recent hierarchical methods extend multimodal interaction across multiple network stages, they mainly focus on expanding the interaction range across scales or encoder–decoder levels. A remaining issue is how to assign different interaction roles according to representation depth while preserving modality-specific structural information. Accordingly, our framework organizes bimodal collaboration into input-level texture–geometry enhancement, controlled shallow-to-middle DSM structural propagation, and deep class-semantic coordination. This hierarchical representation is further connected to the frequency-guided decoding process described in the following sections.

2.3 Decoder upsampling and frequency-aware feature reconstruction

Deep semantic features produced by the encoder must be progressively restored toward pixel-level feature representations during decoding. Bilinear interpolation and transposed convolution can enlarge feature maps, but the former cannot adapt reconstruction weights to local content and the latter may produce uneven responses. Direct concatenation of high-level semantics and shallow details also cannot explicitly correct positional deviations between features at different scales. Content-aware upsampling such as CARAFE [32] improves local aggregation using position-dependent reconstruction weights, while feature-alignment methods learn spatial correspondence through offset-based resampling [33]. These studies show that decoding is more than a change in spatial resolution; it also involves reconstructing regional structure, correcting cross-level correspondence, and reconstructing detailed boundaries.

Frequency-aware methods further organize cross-scale fusion according to the distinction between regional semantics and structural details. FreqFusion coordinates intra-class consistency and boundary localization through adaptive low-pass processing, similarity-guided resampling, and high-frequency enhancement [10], whereas FeatUp focuses on restoring spatial information missing from deep features [34]. In remote sensing, FILNet improves pixel discrimination near boundaries from the perspectives of frequency and feature integrity [35], while F2Net constructs high- and low-frequency branches to balance global context, local detail, and computational cost [36]. These studies demonstrate the value of frequency modeling for dense prediction, particularly for alleviating inconsistent intra-region responses, attenuated boundary details, and fragmented small-object structures. However, they do not directly address continuous optical–DSM collaboration across input enhancement, hierarchical encoding, and multistage decoding.

In summary, task-driven representation enhancement, multimodal fusion, and frequency-aware decoding have each made substantial progress, but these technical directions have largely developed under different problem settings. For optical–DSM segmentation of detailed land-cover structures, two connected problems remain. First, DSM should contribute to hierarchical representation from the input-level enhancement stage so that local geometry is preserved while deep class semantics are strengthened. Second, the enhanced bimodal representation must be accurately decoded back to pixel space to improve region continuity, cross-scale correspondence, and boundary localization. These considerations motivate the optical–DSM collaborative enhancement and frequency-guided decoding framework developed in this work.

3 Method

3.1 Problem formulation and overall framework

This work addresses optical–DSM semantic segmentation for fine-grained urban land-cover interpretation. Given spatially registered optical imagery and DSM, the model predicts a pixel-wise class map, with emphasis on vehicles, narrow roads, building contours, and complex vegetation boundaries where local evidence is limited. The objective is to leverage DSM height and geometry to complement optical texture and accurately decode the resulting bimodal representation for pixel-level prediction.

The proposed method performs bimodal feature enhancement jointly with encoding and decoding. Segmentation supervision is back-propagated through the entire network, while auxiliary reconstruction constraints encourage the front end to retain textures, edges, and geometric cues useful for class discrimination. DSM structure is introduced from the input stage and propagated through the encoder to reduce ambiguity caused by spectral similarity, shadows, and weak local texture.

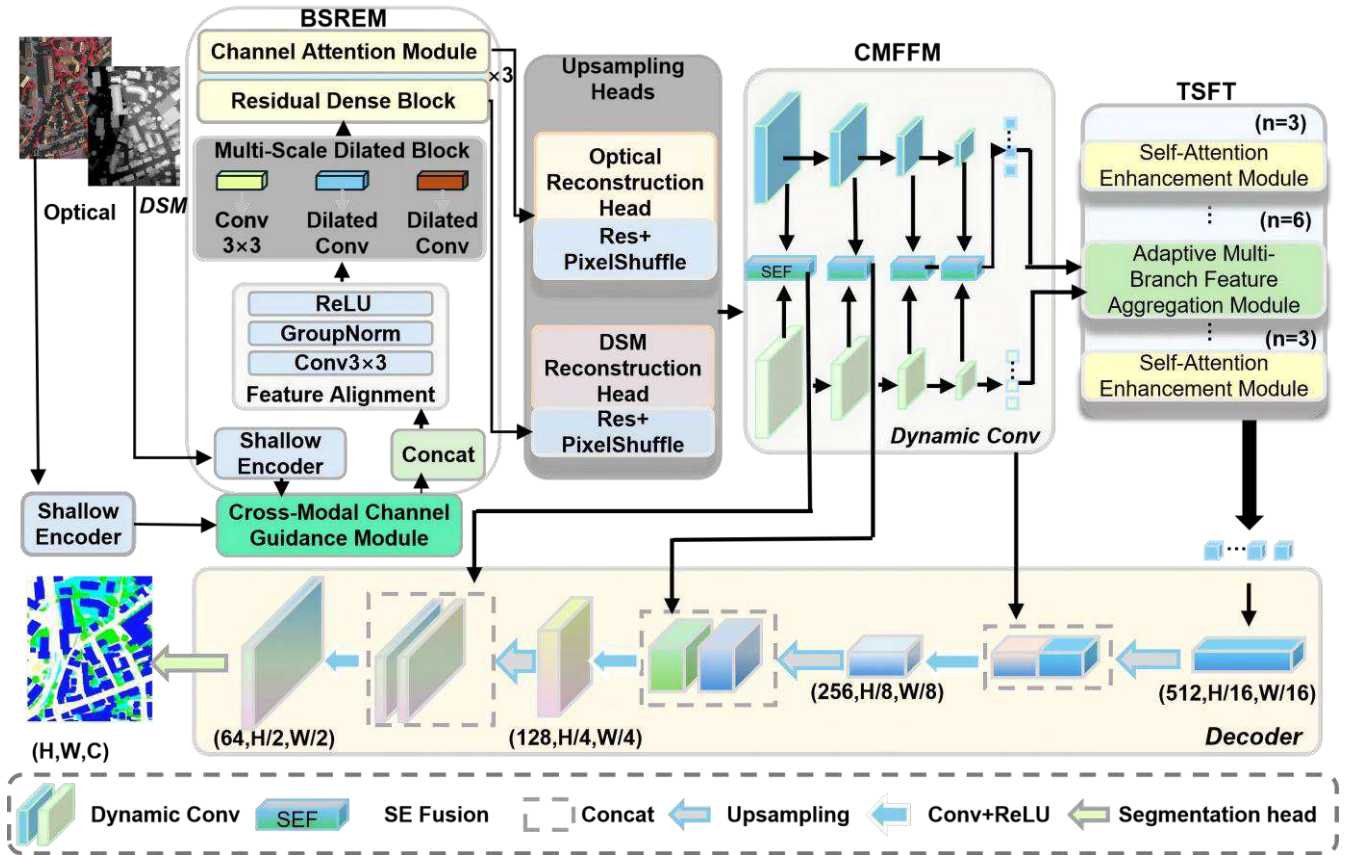


Fig. 1 Overall framework of the proposed optical–DSM collaborative enhancement and frequency-guided decoding network

The overall framework is shown in Fig. 1. Optical imagery and DSM first undergo task-driven collaborative enhancement, followed by hierarchical interaction from shallow-to-middle structural propagation to deep semantic coordination. The decoder then combines deep semantic features with multiscale skip features and performs frequency-guided multistage decoding to improve region continuity, spatial correspondence, and boundary detail. These components form an end-to-end path from bimodal enhancement to pixel-level prediction.

3.2 Task-driven optical–DSM collaborative enhancement and hierarchical representation

Bimodal collaboration is organized by representation depth: input-level enhancement strengthens complementary texture and geometry, shallow and middle layers preserve local structure, and deep layers coordinate intra-modal dependencies and cross-modal semantics.

3.2.1 Input-level bimodal collaborative enhancement

When task-driven enhancement relies only on optical imagery, spectral similarity, shadows, and reduced local contrast can lead to ambiguity in structural representation. We first encode optical imagery and DSM using separate modality-specific shallow branches to preserve their respective feature distributions. The global channel response of one modality then modulates the other, allowing elevation contours to strengthen optical structural representation while optical texture provides reciprocal information to DSM features. The enhanced bimodal features are subsequently concatenated and processed by multiscale structural perception and dense feature integration to form a shared representation. This process is implemented by the Bimodal Spatial Representation Enhancement Module (BSREM), whose main components are shown in Fig. 2. Specifically, BSREM consists of three sequential operations: modality-specific shallow encoding and dynamic convolution, bidirectional cross-modal channel guidance, and multiscale structural perception with dense feature integration.

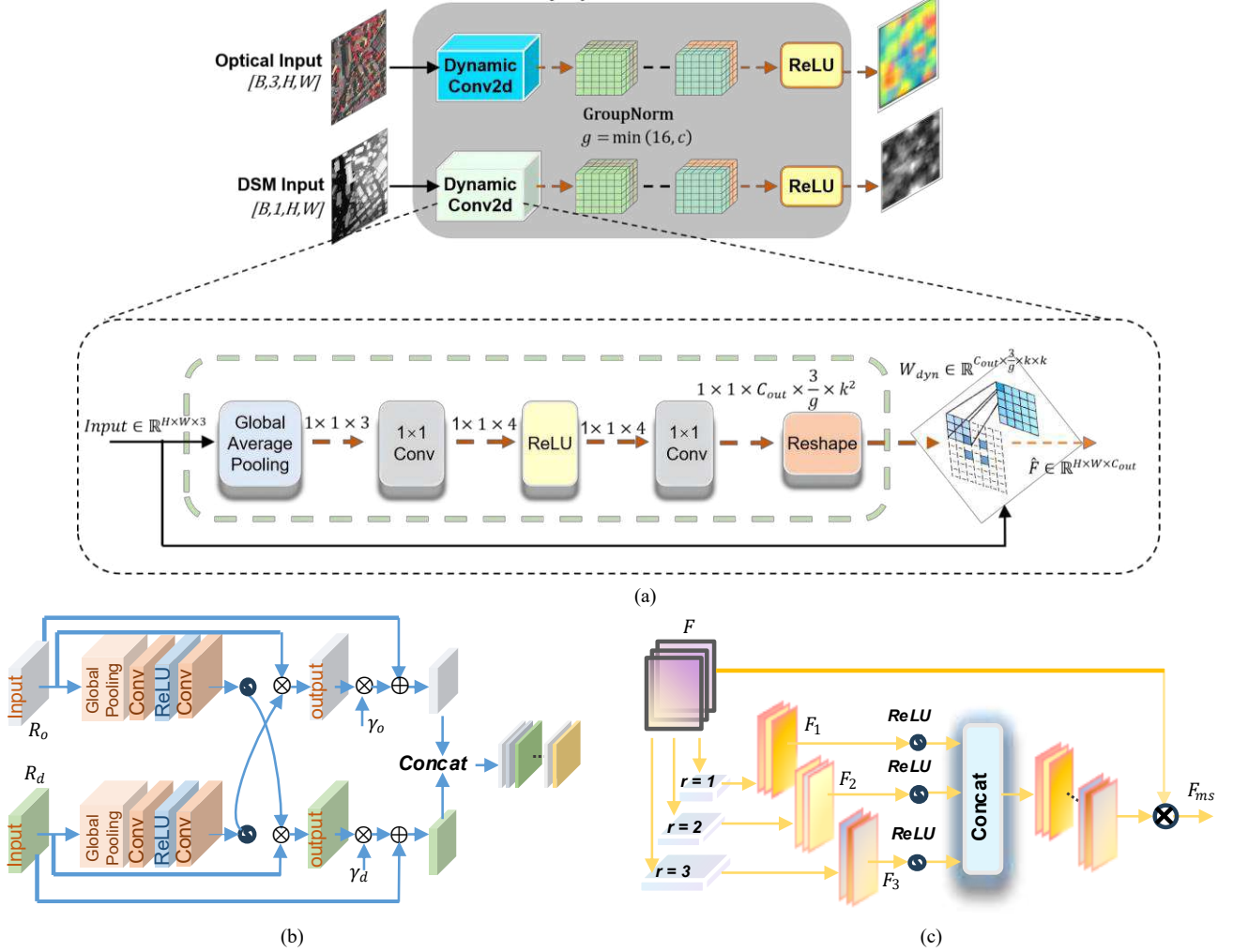


Fig. 2 Bimodal collaborative enhancement process for input-level discriminative representation: (a) modality-specific shallow encoding and dynamic convolution; (b) bidirectional cross-modal channel guidance; (c) multiscale structural perception and dense feature integration

After shallow encoding, the DSM-guided modulation of optical features and the reciprocal optical-guided modulation of DSM features are defined as follows:

$$\tilde{R}_o = R_o + \gamma_o(R_o \odot \mathbf{w}_{d \rightarrow o}) \quad (1)$$

$$\tilde{R}_d = R_d + \gamma_d(R_d \odot \mathbf{w}_{o \rightarrow d}) \quad (2)$$

where R_o and R_d denote the shallow optical and DSM features, respectively, and $\tilde{R}_o$ and $\tilde{R}_d$ are the corresponding enhanced features; $\mathbf{w}_{d \rightarrow o}$ is the channel-guidance weight generated from DSM and applied to the optical feature, whereas $\mathbf{w}_{o \rightarrow d}$ is generated from the optical feature and applied to the DSM feature; $\odot$ denotes channel-wise broadcast multiplication, meaning that the channel weights are shared across spatial locations; and γ_o and γ_d are learnable residual scaling factors initialized to 0 so that the module remains close to an identity mapping at the beginning of training. Equations (1) and (2)

jointly define the bidirectional cross-modal channel guidance in BSREM. As shown in Fig. 2(c), the enhanced bimodal features are concatenated and processed by multiscale structural perception and dense feature integration to form a shared representation, which is then fed into modality-specific optical and DSM reconstruction heads. The reconstructed modality features serve as inputs to the subsequent segmentation encoder. During training, additional reconstruction and edge supervision is imposed on these outputs; during inference, only the auxiliary loss terms are removed, while the reconstruction heads remain in the forward path. The heads operate at the original spatial resolution and serve task-driven representation enhancement rather than image super-resolution.

3.2.2 Shallow-to-middle local structural propagation

Input-level enhancement supplements discriminative evidence, but completely mixing the two modalities in early encoding may weaken modality-specific elevation structure in DSM (shown in Fig. 3). Therefore, shallow and middle encoding stages first select informative channel responses from optical and DSM features and then form a fused main branch at the same spatial scale. The fused result updates the optical main branch and is passed to the decoder as a skip feature, while the DSM feature continues along an independent branch. This design supplements local structure without prematurely discarding elevation information. The process is implemented by the Cross-Modal Feature Fusion Module (CMFFM).

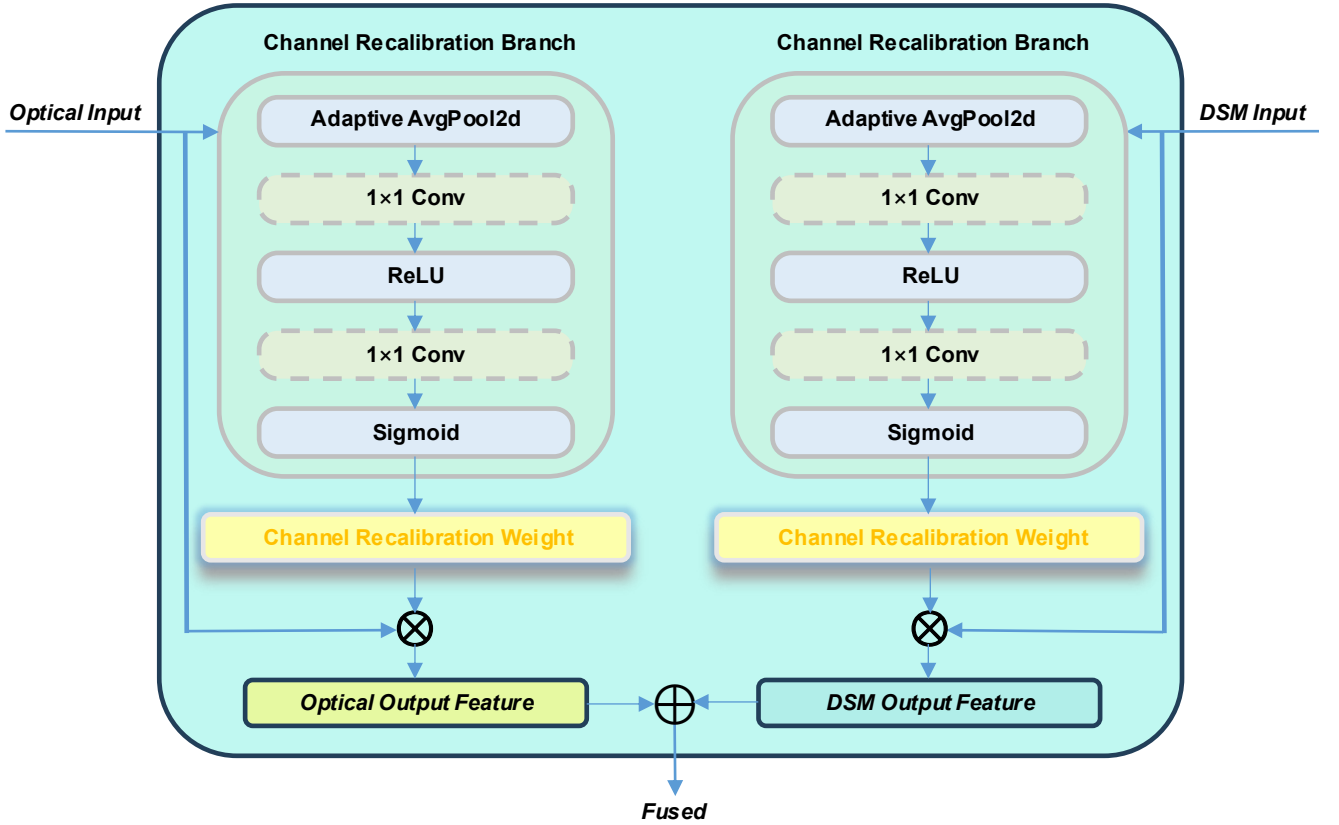


Fig. 3 Local structural propagation and dual-branch interaction in shallow and middle layers

At the r -th encoding stage, the two modalities are first recalibrated along the channel dimension and then combined into a fused main branch at the same scale:

$$\tilde{F}_{r,o} = F_{r,o} \odot \mathbf{a}_{r,o}, \quad \tilde{F}_{r,d} = F_{r,d} \odot \mathbf{a}_{r,d} \quad (3)$$

$$E_r = \tilde{F}_{r,o} + \tilde{F}_{r,d} \quad (4)$$

where $F_{r,o}$ and $F_{r,d}$ denote the optical and DSM features at the r -th encoding stage where CMFFM is deployed; $\mathbf{a}_{r,o}$ and $\mathbf{a}_{r,d}$ are the channel-recalibration weights of the two modalities; $\tilde{F}_{r,o}$ and $\tilde{F}_{r,d}$ denote the recalibrated bimodal features; and E_r is the fused main-branch feature obtained by adding the two recalibrated results. This feature is used both to update the subsequent optical main branch and as a decoder skip feature at the corresponding scale.

Unlike the bidirectional cross-modal guidance in BSREM, CMFFM does not generate cross-modal channel weights. Instead, each modality is independently recalibrated before their responses are fused into the main branch, while the DSM branch is retained for subsequent structural propagation.

3.2.3 Deep class-semantic coordination

As encoding deepens, local texture and geometry are increasingly abstracted into class semantics, making same-scale additive fusion insufficient for long-range intra-modal and cross-modal modeling. The deep stage therefore applies self-attention within each modality and cross-modal attention between optical and DSM representations, followed by reintegration of the enhanced features. This process is implemented by the Two-Stage Fusion Transformer (TSFT) in Fig. 4. Its ACF units retain self-attention and cross-modal branches with learnable scaling factors, while residual connections, normalization, and feed-forward networks complete the feature updates.

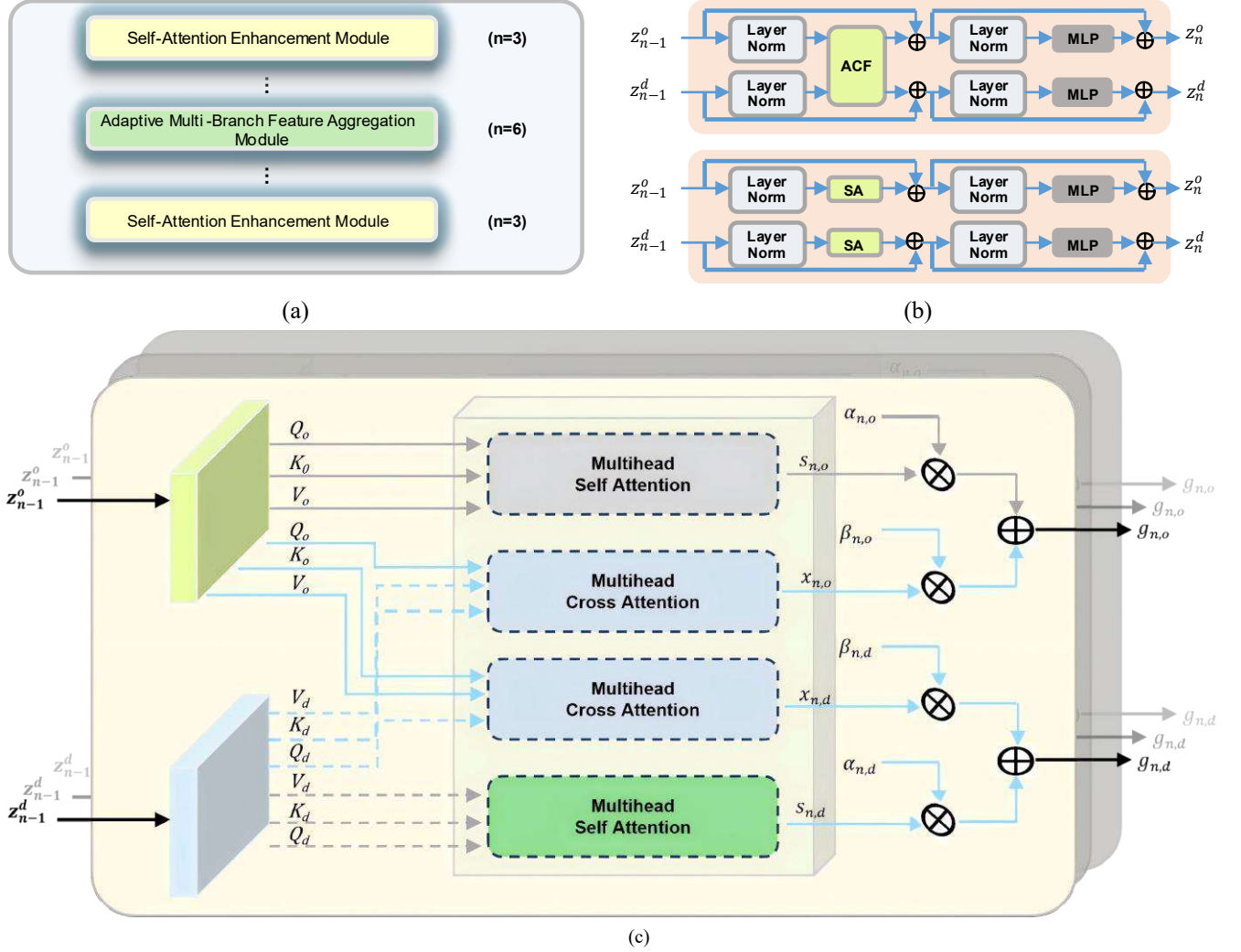


Fig. 4 TSFT architecture for deep class-semantic coordination: (a) overall TSFT architecture; (b) update structures of the self-attention enhancement and ACF modules; (c) multi-branch attention fusion in ACF

The ACF in Fig. 4 retains both self-attention and cross-modal attention branches and uses independent learnable scaling factors to regulate the two types of responses. The core fusion relationships are given in Eqs. (5) and (6).

$$g_{n,o} = \alpha_{n,o}s_{n,o} + \beta_{n,o}x_{n,o} \quad (5)$$

$$g_{n,d} = \alpha_{n,d}s_{n,d} + \beta_{n,d}x_{n,d} \quad (6)$$

where n denotes the ACF module index; $s_{n,o}$ and $s_{n,d}$ are the self-attention responses of the optical and DSM modalities, respectively; $x_{n,o}$ and $x_{n,d}$ are the cross-modal attention responses obtained using the key-value information from the other modality; $\alpha_{n,o}$, $\beta_{n,o}$, $\alpha_{n,d}$, and $\beta_{n,d}$ are independent learnable scaling factors for the two attention branches; and $g_{n,o}$ and $g_{n,d}$ are the bimodal attention-fusion responses output by ACF. Equations (5) and (6) describe only the key fusion relationship

in the ACF shown in Fig. 4; residual connections, normalization, and feed-forward networks perform the subsequent updates according to the TSFT architecture. After multiple interaction layers, the two deep representations are merged at the output and used as the initial semantic feature of the decoder.

3.3 Frequency-guided multistage decoding

Hierarchical encoding produces deep class semantics and multiscale structural features, but these representations must still be progressively decoded toward pixel-level prediction space. Fixed interpolation cannot adapt neighborhood contributions to local content, while direct cross-scale concatenation does not explicitly correct spatial correspondence or compensate for high-frequency detail attenuation. Our decoder is inspired by the three complementary operations introduced in FreqFusion [10], namely low-frequency adaptive filtering, similarity-guided spatial resampling, and high-frequency enhancement. We reorganize these operations into a continuous multistage decoding process and couple them with the hierarchical bimodal skip representations produced by the proposed optical-DSM encoder. This integrates frequency-aware reconstruction stage by stage with hierarchical bimodal representations to progressively improve region consistency, cross-scale spatial correspondence, and boundary-detail preservation. At each decoding stage, the current high-level semantic feature is reconstructed together with the corresponding skip feature, and the resulting representation is passed to the next stage, as shown in Fig. 5.

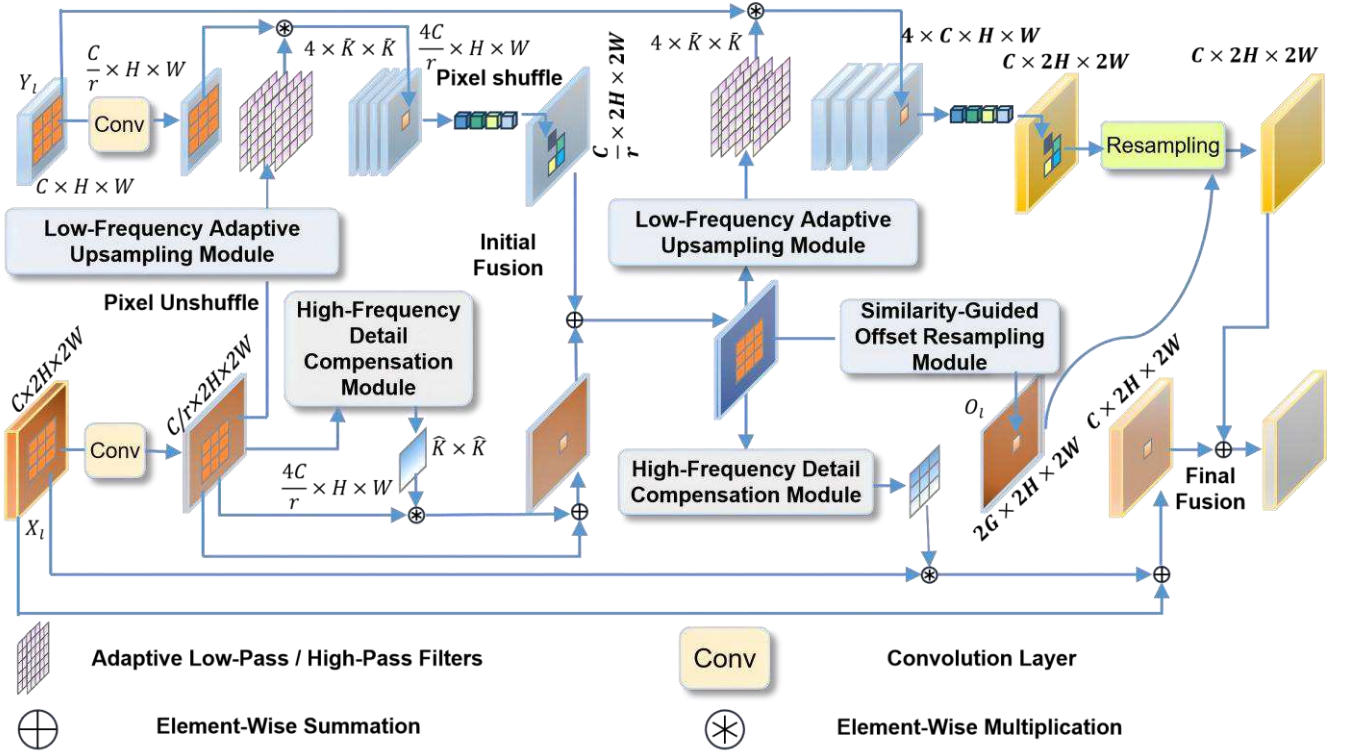


Fig. 5 Frequency-guided multistage decoding process

First, the decoder predicts position-dependent dynamic low-pass kernels from the current local content, allowing each output location to use adaptive normalized neighborhood weights for region-consistent reconstruction of high-level semantics. The dynamic low-pass kernel is normalized as follows:

$$W_{\ell,t}(i,j;p,q) = \frac{\exp(V_{\ell,t}(i,j;p,q))}{\sum_{(u,v) \in \Omega} \exp(V_{\ell,t}(i,j;u,v))} \quad (7)$$

After normalization, the local reconstruction result for the t -th subpixel group is given by:

$$\hat{Y}_{\ell,t}(i,j) = \sum_{(p,q) \in \Omega} W_{\ell,t}(i,j;p,q) Y_{\ell}(i+p,j+q) \quad (8)$$

where ℓ denotes the decoding stage, (i,j) is the current spatial location, and (p,q) is the local neighborhood offset relative to the current position; $t \in \{1,2,3,4\}$ denotes the subpixel-group index for $2\times$ upsampling; $V_{\ell,t}(i,j;p,q)$ represents the

unnormalized kernel parameter, Ω is the $K \times K$ local-neighborhood index set, and $W_{\ell,t}(i, j; p, q)$ is the normalized dynamic low-pass weight; Y_ℓ denotes the high-level feature to be upsampled, and $\hat{Y}_{\ell,t}(i, j)$ is the local reconstruction result of the t -th subpixel group. Equation (7) ensures that kernel weights within the same location and subpixel group are nonnegative and sum to 1. Equation (8) uses scalar weights for local weighted reconstruction of neighboring features, followed by subpixel rearrangement to restore the spatial resolution. This operation suppresses abnormal high-frequency fluctuations within class regions while maintaining semantic continuity. The low-pass kernel size is set to 5.

Region reconstruction improves intra-class continuity, but spatial misalignment may remain between high-level semantics and shallow structure. The decoder therefore computes cosine similarity between each center location and its local neighborhood, concatenates the similarity descriptor with the current feature along the channel dimension, and uses the result to predict candidate offset components and their amplitude modulation coefficients. The local similarity and offset-generation relationships are defined in Eqs. (9)–(12).

$$\rho_{\ell}(i, j; p, q) = \frac{\langle Z_{\ell}(i, j), Z_{\ell}(i+p, j+q) \rangle}{\|Z_{\ell}(i, j)\|_2 \|Z_{\ell}(i+p, j+q)\|_2} \quad (9)$$

$$\mathbf{P}_\ell = \text{Conv}_p(\text{Concat}[Z_\ell, S_\ell]) \quad (10)$$

$$\mathbf{A}_\ell = \sigma(\text{Conv}_a(\text{Concat}[Z_\ell, S_\ell])) \quad (11)$$

$$\mathbf{O}_\ell(i, j) = \mathbf{P}_\ell(i, j) \odot \mathbf{A}_\ell(i, j) \quad (12)$$

where $Z_\ell(i, j) \in \mathbb{R}^C$ denotes the feature vector at location (i, j) in the ℓ -th decoding stage; $\rho_\ell(i, j; p, q)$ is the cosine similarity between this position and neighboring location $(i + p, j + q)$; C is the number of feature channels; S_ℓ denotes the descriptor formed by local similarities; $\mathbf{P}_\ell$ is the candidate offset-component field; $\mathbf{A}_\ell$ is the component-wise amplitude-modulation field; σ denotes the sigmoid function; Conv_p and Conv_a denote the convolutional predictors for the offset components and amplitude modulation, respectively, and Concat denotes channel-wise concatenation; $\mathbf{O}_\ell(i, j)$ is the final offset field; $\odot$ denotes element-wise multiplication between tensors of the same shape. The final offset field contains the horizontal and vertical displacement components at each spatial location: $\mathbf{O}_\ell(i, j) = (\Delta i_\ell(i, j), \Delta j_\ell(i, j))$, where $\Delta i_\ell(i, j)$ and $\Delta j_\ell(i, j)$ are the horizontal and vertical offset components for each group, respectively. The current-scale feature is differentially resampled according to the horizontal and vertical offsets of each group. Bilinear interpolation is used for non-integer coordinates, as expressed in Eq. (13).

$$\tilde{Z}_\ell(i, j) = Z_\ell(i + \Delta i_\ell(i, j), j + \Delta j_\ell(i, j)) \quad (13)$$

where $\tilde{Z}_\ell$ is the corrected feature after resampling. Equation (9) establishes the local semantic-consistency descriptor, Eqs. (10) and (11) predict the candidate offset components and amplitude-modulation coefficients, Eq. (12) produces the final offset field, and Eq. (13) performs bilinear resampling according to the offset coordinates. This process adjusts cross-scale spatial correspondence based on local semantic consistency without changing class semantics.

After spatial correction, low-pass reconstruction may still attenuate boundary responses. To recover local structural details, the High-Frequency Detail Compensation module (HFDC) independently predicts a normalized 3×3 smoothing kernel at each spatial location. Following the adaptive high-pass formulation, the corresponding dynamic high-pass kernel is constructed by subtracting the normalized smoothing kernel from a 3×3 discrete unit impulse kernel:

$$H_\ell(i, j; p, q) = \delta(p, q) - W_\ell^S(i, j; p, q), (p, q) \in \Omega_{(3 \times 3)} \quad (14)$$

where $W_\ell^S(i, j; p, q)$ denotes the normalized 3×3 smoothing-kernel weight independently predicted by HFDC at spatial location (i, j) of the ℓ -th decoding stage for neighborhood offset (p, q) ; $\delta(p, q)$ denotes the 3×3 discrete unit impulse kernel, whose center is 1 and all other positions are 0; and $H_\ell(i, j; p, q)$ denotes the resulting dynamic high-pass kernel.

The high-frequency response $F_{\ell, \text{high}}$ is obtained by applying the dynamic high-pass kernel H_ℓ to the current-scale feature F_ℓ . It is then added back to the original feature through residual compensation:

$$\tilde{F}_\ell = F_\ell + \eta F_{\ell, \text{high}} \quad (15)$$

where $F_{\ell, \text{high}}$ denotes the extracted high-frequency detail response, η is a learnable high-frequency compensation coefficient, and F_ℓ and $\tilde{F}_\ell$ denote the current-scale features before and after high-frequency compensation, respectively. Because the weights of the normalized 3×3 smoothing kernel W_ℓ^S sum to 1, the resulting high-pass kernel H_ℓ has zero-

sum weights and therefore emphasizes local structural differences such as boundaries and fine details. The high-pass kernel size is set to 3.

3.4 Joint auxiliary reconstruction and segmentation optimization

To jointly optimize front-end representation enhancement and pixel prediction under a shared task objective, the model jointly uses a segmentation loss, bimodal reconstruction loss, and edge loss. The segmentation loss combines class-weighted cross-entropy and Dice losses; the reconstruction loss constrains the auxiliary reconstruction outputs of both optical imagery and DSM; and the edge loss compares the Sobel-gradient differences between the auxiliary reconstruction outputs and their corresponding targets. The total loss is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} \quad (16)$$

where $\mathcal{L}_{\text{seg}}$ denotes the semantic segmentation loss, $\mathcal{L}_{\text{rec}}$ is the reconstruction loss for the optical and DSM auxiliary reconstruction outputs, and $\mathcal{L}_{\text{edge}}$ is the edge loss based on Sobel-gradient differences; λ_{seg} , λ_{rec} , and λ_{edge} are the weights of the three losses and are set to 1.0, 0.1, and 0.05, respectively. Through end-to-end back-propagation, segmentation supervision updates the front-end enhancement, encoding, and decoding processes so that the bimodal representation retains information useful for pixel classification while satisfying the auxiliary reconstruction constraints.

4 Experiments and results

4.1 Datasets and input processing

Experiments are conducted on the Vaihingen and Potsdam 2D Semantic Labeling datasets of the International Society for Photogrammetry and Remote Sensing (ISPRS). Both datasets provide true orthophotos (TOP), DSMs, and pixel-level land-cover annotations, with classes including impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. The Vaihingen data used in the ISPRS benchmark originate from the DGPF digital airborne camera evaluation project [37]. For Vaihingen, we use 16 publicly annotated tiles; the optical imagery consists of near-infrared, red, and green bands with a ground sampling distance of 9 cm. The Potsdam dataset contains 38 image tiles of 6000×6000 pixels with a ground sampling distance of 5 cm. Following the experimental protocol adopted by FTransUNet [25], we use the same 24-tile subset for our experiments. RGB imagery and the corresponding DSM are used as input modalities.

For Vaihingen, tiles 1, 3, 7, 11, 13, 17, 23, 26, 28, 32, 34, and 37 are used for training, while tiles 5, 15, 21, and 30 are used for testing. The 24 selected Potsdam tiles are divided into 18 training tiles and six testing tiles. Specifically, tiles 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, and 4_11 are used for training, while tiles 4_10, 5_11, 2_11, 3_10, 6_11, and 7_12 are used for testing.

During training, 256×256 local patches are randomly cropped from the original tiles, and horizontal and vertical flipping are each applied with a probability of 0.5. Optical pixel intensities are normalized to $[0, 1]$. For each tile, the DSM is independently min-max normalized to $[0, 1]$. This tile-wise normalization maps the elevation range of each DSM to a common numerical scale while preserving the relative height ordering and local relief patterns within the tile. As a result, the DSM branch mainly exploits relative geometric structure and local elevation contrast, whereas absolute elevation magnitudes are not directly comparable across different tiles after normalization. During testing, full-tile prediction is performed using a 256×256 sliding window with strides of 32 and 128 pixels for Vaihingen and Potsdam, respectively. They were empirically selected to balance overlap-based boundary smoothing and inference efficiency. Predictions from overlapping windows are accumulated pixel-wise, and the final class prediction is obtained by argmax.

4.2 Implementation details and evaluation metrics

The model is trained on Ubuntu 22.04 using a single NVIDIA RTX 4090 GPU with 24 GB memory, Python 3.10, PyTorch 2.1.0, and CUDA 12.1. The encoder adopts a hybrid ResNet50 and ViT-B/16 architecture initialized with the R50+ViT-B_16.npz pretrained weights. Training uses stochastic gradient descent (SGD) with a momentum of 0.9, a weight decay of 0.0005, and an initial learning rate of 0.01. A MultiStep learning-rate schedule is adopted, multiplying the current learning rate by 0.1 at epochs 25, 35, and 45. All models are trained for 50 epochs, and the checkpoint from the final epoch is used for testing. The main implementation settings are summarized in Table 1.

Table 1 Main implementation settings

Item	Setting
Batch size	8
Training epochs	50
Initial learning rate	0.01
Encoder configuration	ResNet50 + ViT-B/16; input size 256×256 ; three skip connections
Frequency-guided decoding	LFAU kernel: 5×5 ; HFDC kernel: 3×3 ; compressed channels: 64; both low- and high-frequency branches enabled with residual connections

The evaluation metrics include overall accuracy (OA), F1-score, intersection over union (IoU), mean F1-score (mF1), and mean intersection over union (mIoU). OA is the proportion of correctly predicted pixels among all valid pixels. Per-class F1 and IoU are calculated using Eq. (17), and mF1 and mIoU are obtained by averaging over the five foreground classes according to Eq. (18).

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \quad IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (17)$$

$$mF1 = \frac{1}{N_f} \sum_{c \in \mathcal{C}_f} F1_c, \quad mIoU = \frac{1}{N_f} \sum_{c \in \mathcal{C}_f} IoU_c \quad (18)$$

In Eqs. (17) and (18), TP_c , FP_c , and FN_c denote the true positives, false positives, and false negatives of class c , respectively, and c is the class index. $\mathcal{C}_f$ is the foreground-class set used for averaging, including Building, Tree, Low Vegetation, Car, and Impervious Surface. $N_f = |\mathcal{C}_f| = 5$ denotes the number of foreground classes included in the average. Clutter/background is excluded from the calculation of mF1 and mIoU.

4.3 Comparison with representative methods

To evaluate the overall competitiveness of the proposed framework for urban land-cover segmentation with complex structures and small objects, we compare it with representative methods from different semantic segmentation paradigms. PIDNet [38] represents real-time convolutional segmentation, while GCNet [39] focuses on global context modeling. Swin Transformer [18], SegFormer [17], and SETR [16] represent different Transformer-based segmentation strategies, and MaskFormer [40] is included to compare the mask-classification paradigm. We further include recent optical–DSM bimodal methods, MSFFNet [6] and BATFNet [8], which represent spatial–frequency fusion and boundary-aware cross-modal modeling, respectively. For each competing method, we follow the corresponding official implementation and the model-specific settings recommended in the original publication, including pretrained initialization and architecture-specific hyperparameters. Meanwhile, all methods are retrained using the same training/test split, input size, data augmentation, training duration, and evaluation protocol. Generic single-modality methods use optical imagery as input, whereas MSFFNet [6], BATFNet [8], and our method use spatially registered optical imagery and DSM. Therefore, MSFFNet and BATFNet provide the most direct modality-matched comparisons, while the single-modality methods are included to assess overall competitiveness across representative segmentation paradigms.

The staged and component-level analyses presented below further examine the contributions of hierarchical bimodal collaboration and frequency-guided decoding.

4.3.1 Results on Vaihingen

The full model achieves an OA of 92.26%, an mF1 of 90.91%, and an mIoU of 83.74% on Vaihingen. Among generic segmentation methods, Swin obtains the highest mF1 (86.43%) and mIoU (76.59%), while among recent bimodal methods BATFNet reaches an mF1 of 86.27% and an mIoU of 76.43%, showing overall performance comparable to Swin. Relative to the best result among all external methods, our model improves mF1 and mIoU by 4.48 and 7.15 percentage points, respectively. The class-wise results further show that our model achieves the highest F1 score for all five foreground classes—Building, Tree, Low Vegetation, Car, and Impervious Surface. Compared with the best external results, F1 for Car and Low Vegetation increases by 5.79 and 3.65 percentage points, respectively. These results indicate that the gains are reflected not

only in overall region-overlap quality but also in pixel discrimination for small vehicles, vegetation classes, and man-made surface categories.

Table 2 Segmentation results on Vaihingen (%)

Model category	Model	Bui.	Tre.	Low.	Car	Imp.	OA	mF1	mIoU
Single-modality methods	PIDNet	90.13	83.16	65.12	53.51	84.37	82.27	75.26	62.20
	GCNet	92.54	85.16	73.59	79.87	87.27	85.70	83.68	72.48
	Swin	94.74	88.27	77.98	81.59	89.59	88.68	86.43	76.59
	MaskFormer	93.28	87.20	74.31	78.68	87.32	86.66	84.16	73.24
	SegFormer	94.85	87.99	77.44	78.12	89.27	88.28	85.53	75.33
	SETR	91.04	84.86	72.65	54.84	84.88	84.06	77.65	65.16
Optical–DSM bimodal methods	MSFFNet	93.88	86.83	74.51	82.90	88.70	87.28	85.36	75.01
	BATFNet	95.00	87.22	75.05	84.63	89.44	88.01	86.27	76.43
	Ours	97.52	91.54	81.63	90.42	93.42	92.26	90.91	83.74

Table 3 Segmentation results on Potsdam (%)

Model category	Model	Bui.	Tre.	Low.	Car	Imp.	OA	mF1	mIoU
Single-modality methods	PIDNet	94.55	81.07	81.80	88.88	89.39	86.81	87.14	77.57
	GCNet	87.50	78.88	68.61	88.62	84.08	80.27	81.54	69.44
	Swin	96.09	84.47	84.27	90.86	90.91	88.66	89.32	81.00
	MaskFormer	95.48	84.48	83.64	90.96	90.45	88.40	88.99	80.47
	SegFormer	95.40	83.63	83.21	89.03	89.94	88.00	88.24	79.25
	SETR	88.72	62.39	73.89	80.18	83.86	78.52	77.81	64.56
Optical–DSM bimodal methods	MSFFNet	94.91	82.22	82.27	90.39	89.29	87.12	87.81	78.62
	BATFNet	96.44	84.19	83.81	91.55	91.09	88.92	89.41	81.20
	Ours	97.42	87.91	87.96	95.84	92.73	90.98	92.37	86.07

4.3.2 Results on Potsdam

On Potsdam, the full model achieves an OA of 90.98%, an mF1 of 92.37%, and an mIoU of 86.07%. Among the recent bimodal methods, BATFNet provides the strongest external overall result, with an OA of 88.92%, an mF1 of 89.41%, and an mIoU of 81.20%. Our model improves these values by 2.06, 2.96, and 4.87 percentage points, respectively. At the class level, our model also maintains the highest F1 score for all five foreground classes. Compared with the best external results, F1 improves by 4.29 percentage points for Car, 3.69 points for Low Vegetation, and 3.43 points for Tree. Together with the Vaihingen results, these findings show that the proposed method consistently improves both region-level consistency and detailed class discrimination across the two urban remote sensing datasets, particularly for vehicles, vegetation classes, and complex impervious-surface regions.

The following stage-wise and component-level analyses isolate the contributions of bimodal collaboration and frequency-guided decoding.

4.4 Stage-wise analysis of the proposed framework

We define four progressive configurations to evaluate the proposed framework (Table 4): Stage A is the optical segmentation baseline; Stage B introduces the front-end task-driven representation enhancement pathway without the complete hierarchical bimodal interaction; Stage C extends Stage B with the full optical–DSM hierarchical representation pathway, including BSREM, CMFFM, and TSFT; and Stage D adds frequency-guided multistage decoding to form the complete model. Table 5 further decomposes the internal components of this hierarchical representation stage under a dedicated fixed ablation configuration.

We can find that from Stage A to Stage B, mF1/mIoU increase by 3.64/7.48 points on Vaihingen and 3.99/7.53 points on Potsdam. These gains support the effectiveness of task-driven representation enhancement before hierarchical bimodal interaction. Stage C further improves mF1/mIoU by 1.77/2.82 points on Vaihingen and 0.74/1.92 points on Potsdam. These

gains indicate that DSM geometry provides complementary structural cues when it participates throughout input enhancement, shallow-to-middle propagation, and deep semantic coordination instead of being fused only at a single location.

Adding frequency-guided multistage decoding in Stage D yields further gains of 2.29/1.58 points on Vaihingen and 2.71/4.17 points on Potsdam in mF1/mIoU. The result indicates that region-consistency reconstruction, spatial correction, and high-frequency compensation provide complementary benefits after hierarchical bimodal encoding.

Performance increases monotonically from Stage A to Stage D on both datasets, supporting the complementary roles of task-driven enhancement, hierarchical bimodal representation, and frequency-guided decoding. Because Stage C contains several interaction mechanisms, the next section isolates the contributions of BSREM, CMFFM, and TSFT.

Table 4 Stage-wise analysis of the proposed framework (%)

Stage	Configuration	Vaihingen		Potsdam	
		mF1	mIoU	mF1	mIoU
A	Optical-only baseline	83.21	71.86	84.93	72.45
B	A + task-driven representation enhancement	86.85	79.34	88.92	79.98
C	B + optical–DSM hierarchical collaboration	88.62	82.16	89.66	81.90
D	C + frequency-guided multistage decoding	90.91	83.74	92.37	86.07

4.5 Component analysis of hierarchical bimodal representation

Because Table 4 shows consistent monotonic stage-wise trends on both datasets, the subsequent finer-grained component analyses are reported on Vaihingen as a representative benchmark. Stage C contains BSREM, CMFFM, and TSFT. To examine the incremental contribution of each component, we conduct a progressive component ablation (Table 5), sequentially introducing input-level enhancement, shallow-to-middle interaction, and deep semantic coordination while keeping the remaining settings unchanged.

Table 5 Component ablation of hierarchical bimodal representation on Vaihingen (%)

BSREM	CMFFM	TSFT	mF1	mIoU
×	×	×	86.91	79.58
✓	×	×	87.45	80.41
✓	✓	×	87.63	80.76
✓	✓	✓	88.62	82.16

The component-wise results reveal a clear progression from input-level enhancement to structural propagation and finally to deep semantic coordination. Introducing BSREM first improves mF1 from 86.91% to 87.45% and mIoU from 79.58% to 80.41%. This indicates that strengthening modality-specific optical texture and DSM geometry before semantic encoding provides a more discriminative starting representation.

The subsequent addition of CMFFM brings smaller but consistent gains, reaching 87.63% mF1 and 80.76% mIoU. Rather than producing a large standalone improvement, CMFFM primarily preserves DSM structural information during shallow-to-middle encoding and supplies recalibrated multimodal features to later stages, preventing feature abstraction from weakening the early gains from BSREM.

When TSFT is further introduced, mF1 and mIoU rise to 88.62% and 82.16%, corresponding to the largest incremental gains of 0.99 and 1.40 percentage points. This stronger improvement suggests that, once local texture and geometric structure have been sufficiently enhanced and propagated, the remaining performance bottleneck lies mainly in high-level class confusion. Deep self-attention and cross-modal interaction therefore provide the greatest additional benefit by coordinating long-range intra-modal dependencies and cross-modal class semantics. Overall, the progressive gains support the intended hierarchical design: BSREM establishes stronger bimodal evidence, CMFFM preserves and transfers that evidence through intermediate representations, and TSFT converts it into more discriminative high-level semantics.

CMFFM provides a smaller incremental gain than those observed after introducing BSREM and TSFT, suggesting that its primary role is to bridge input enhancement and deep interaction by retaining DSM structure and providing recalibrated skip features. The larger incremental gain associated with TSFT further highlights the importance of deep cross-modal semantic coordination for reducing class confusion.

This component analysis uses a dedicated fixed baseline configuration to isolate the incremental effects of BSREM, CMFFM, and TSFT. Therefore, the baseline scores in Table 5 are not directly compared with Stage B in Table 4, whereas the complete configuration is consistent with Stage C.

4.6 Progressive component analysis of the frequency-guided decoder

We further analyze the progressive effects of the Low-Frequency Adaptive Upsampling module (LFAU), Similarity-Guided Offset Resampling module (SGOR), and High-Frequency Detail Compensation module (HFDC), which target region consistency, cross-scale spatial correction, and boundary/detail recovery, respectively (Table 6). The three modules are introduced sequentially while the remaining settings are kept unchanged to evaluate their incremental contributions to the frequency-guided decoder.

Table 6 Progressive component analysis of the frequency-guided multistage decoder on Vaihingen (%)

LFAU	SGOR	HFDC	mF1 (%)	mIoU (%)
×	×	×	88.62	82.16
✓	×	×	89.58	82.23
✓	✓	×	90.28	82.88
✓	✓	✓	90.91	83.74

Introducing LFAU increases mF1 from 88.62% to 89.58% and mIoU from 82.16% to 82.23%, corresponding to gains of 0.96 and 0.07 percentage points, respectively. The result indicates that adaptive low-frequency reconstruction provides a modest benefit at this stage, with a clearer improvement in mF1 than in mIoU.

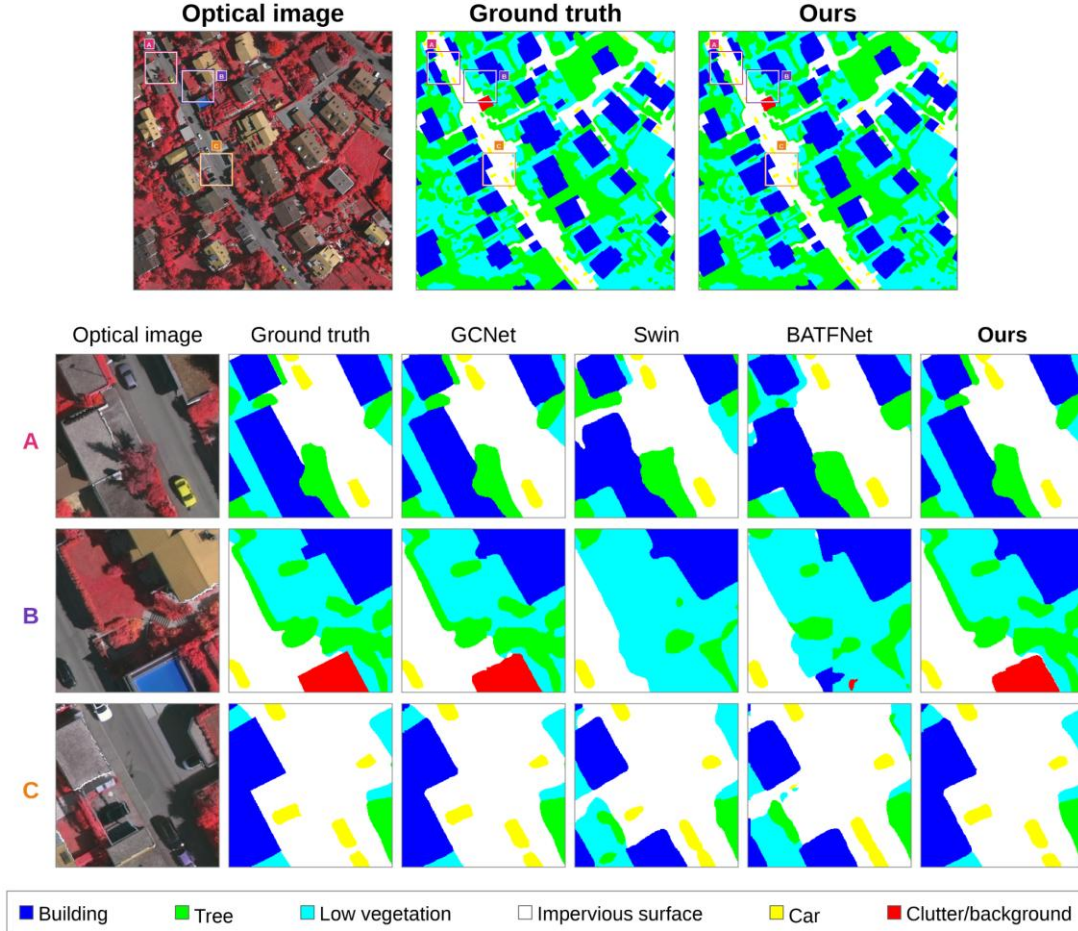


Fig. 6 Qualitative segmentation comparisons on the Vaihingen test set

Adding SGOR on top of LFAU further increases mF1 from 89.58% to 90.28% and mIoU from 82.23% to 82.88%, yielding additional gains of 0.70 and 0.65 percentage points, respectively. These improvements indicate that similarity-guided offset resampling improves cross-scale spatial correspondence and region-overlap quality after low-frequency reconstruction. Finally, incorporating HFDC increases mF1 from 90.28% to 90.91% and mIoU from 82.88% to 83.74%, adding 0.63 and 0.86 percentage points, respectively. The relatively larger gain in mIoU is consistent with the role of HFDC in compensating high-frequency structural responses and strengthening boundary and fine-detail representation.

Overall, progressively introducing LFAU, SGOR, and HFDC improves mF1 from 88.62% to 90.91% and mIoU from 82.16% to 83.74%, corresponding to overall gains of 2.29 and 1.58 percentage points, respectively. These results indicate that region-consistency reconstruction, spatial correction, and high-frequency detail compensation provide complementary benefits when integrated sequentially into the decoder.

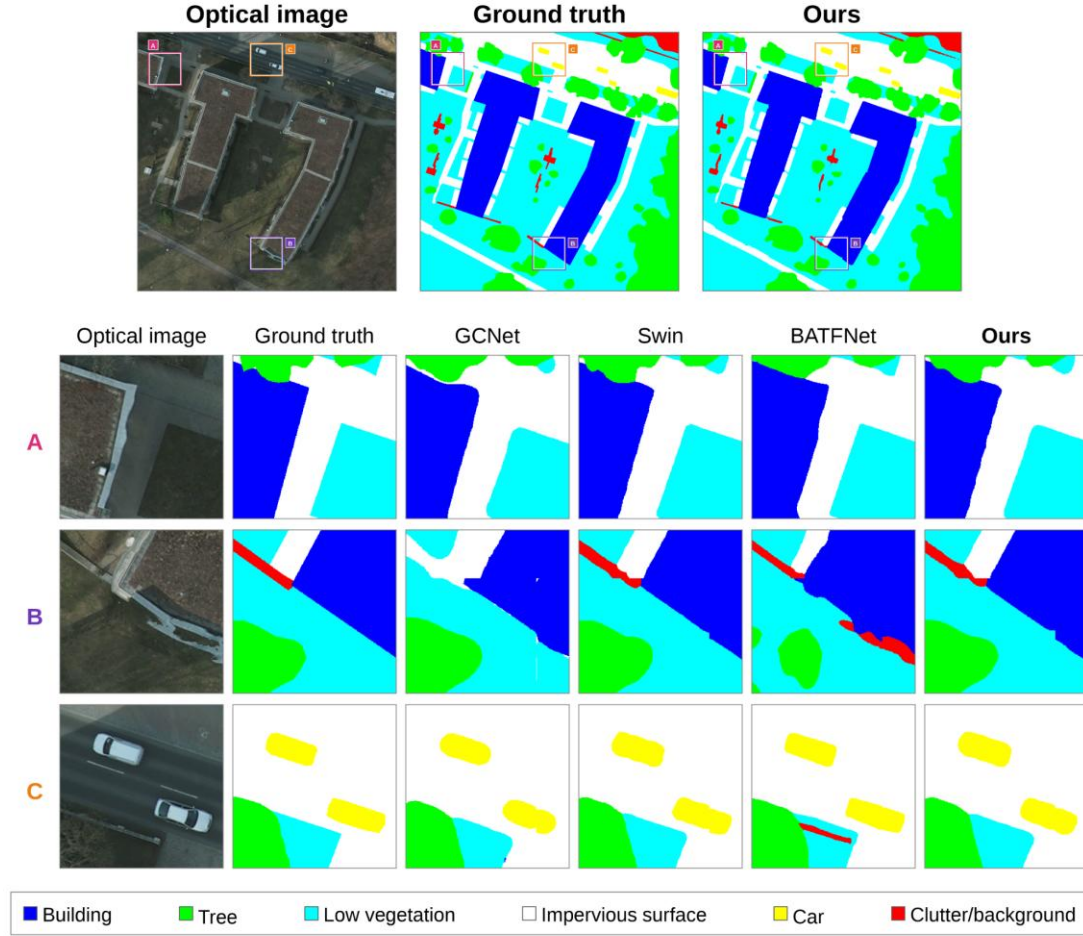


Fig. 7 Qualitative segmentation comparisons on the Potsdam test set

4.7 Qualitative comparison

After the quantitative comparison, representative regions from the ISPRS Vaihingen and ISPRS Potsdam datasets are selected for qualitative comparisons. Direct visual inspection of predictions from different models highlights differences in complex land-cover structures, small-object recognition, and class-boundary preservation.

Fig. 6 shows qualitative results on Vaihingen. All methods can identify the major land-cover categories in regions where buildings, vegetation, and impervious surfaces are adjacent, but different degrees of class leakage and contour displacement remain near local boundaries. Locations with complex vegetation cover are particularly prone to confusion among trees, low vegetation, and surrounding classes. In regions with multiple adjacent classes, some predictions show scattered responses or locally discontinuous structures, whereas our method produces more coherent main regions and clearer boundaries between buildings and surrounding land-cover categories. The vehicle regions further reveal differences in recognizing small objects against cluttered backgrounds. Our prediction preserves vehicle shapes more completely and reduces erroneous merging

between vehicles and surrounding impervious surfaces. Overall, the proposed method performs stably in preserving complex class interfaces, land-cover contours, and small-scale structures.

The main differences in the Potsdam scenes occur at regular man-made boundaries, vegetation–artificial-surface interfaces, and densely distributed small objects, as shown in Fig. 7. Buildings are generally identified well by all methods, but local contours still exhibit varying degrees of outward expansion, gaps, and boundary displacement. At locations where vegetation intersects impervious surfaces, some methods produce fragmented class responses and less continuous region boundaries. Parking areas contain multiple small and closely spaced vehicles, placing greater demands on shape preservation and spatial separation between adjacent vehicles. The methods show visible differences in vehicle-contour completeness and merging between adjacent vehicle regions. Overall, our predictions preserve more regular building boundaries and more continuous class regions while maintaining the structures of dense vehicles. Qualitative results on both datasets indicate that combining optical texture and DSM geometry improves spatial discrimination in complex urban scenes and better preserves local structural details at the pixel level.

5 Conclusion

This work presents a collaborative optical–DSM enhancement and frequency-guided decoding framework for fine-grained remote sensing semantic segmentation. The framework organizes multimodal interaction according to representation depth, ranging from input-level texture–geometry enhancement and shallow-to-middle structural propagation to deep cross-modal semantic coordination. A frequency-guided multistage decoder further reconstructs hierarchical representations toward pixel-level prediction by combining region-consistency processing, similarity-guided spatial correction, and high-frequency detail compensation.

On Vaihingen and Potsdam, the model achieves mF1 scores of 90.91% and 92.37% and mIoU scores of 83.74% and 86.07%, respectively, outperforming the selected representative methods. Stage-wise experiments show consistent gains as task-driven enhancement, hierarchical optical–DSM representation, and frequency-guided reconstruction are introduced. Remaining errors near tree shadows and low-vegetation/impervious-surface interfaces indicate that class confusion at difficult boundaries is not fully resolved.

Future work will investigate cross-city and cross-sensor generalization and systematically evaluate robustness to DSM misregistration, missing data, and quality variation. We will also explore lightweight bimodal encoding and more efficient frequency-guided decoding to reduce computational overhead for practical deployment.

Statements and Declarations

Funding

This work was supported by the National Natural Science Foundation of China (No. 62473008), Beijing Nova Program (No. 20240484710), Xinjiang Talent Development Fund (No. XJRC-2025-JY-YJ-GX-RX-001), Project of Beijing Municipal University Teacher Team Construction Support Plan (No. BPHR20220104), Beijing Natural Science Foundation Joint Fund Key Project (No. L25034), and Beijing Scholars Program (No. 099).

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Data Availability

The Vaihingen and Potsdam datasets used in this study are available from the ISPRS 2D Semantic Labeling Benchmark at <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/>. Access and reuse are subject to the applicable ISPRS/DGPF terms of use. No new datasets were generated in this study.

References

References

- [1] Li, J., Cai, Y., Li, Q., et al.: A review of remote sensing image segmentation by deep learning methods. *Int. J. Digit. Earth* 17(1), 2328827 (2024). <https://doi.org/10.1080/17538947.2024.2328827>

- [2] Zhu, X.X., Tuia, D., Mou, L., et al.: Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* 5(4), 8–36 (2017). <https://doi.org/10.1109/MGRS.2017.2762307>
- [3] Liu, P.: A multimodal fusion framework for semantic segmentation of remote sensing based on multilevel feature fusion learning. *Neurocomputing* 653, 131233 (2025). <https://doi.org/10.1016/j.neucom.2025.131233>
- [4] Ma, A., Wang, J., Zhong, Y., Zheng, Z.: FactSeg: foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* 60, 1–16 (2022). <https://doi.org/10.1109/TGRS.2021.3097148>
- [5] Huang, X., Zhang, X., Wang, L., et al.: MMA-Net: a semantic segmentation network for high-resolution remote sensing images based on multimodal fusion and multi-scale multi-attention mechanisms. *Remote Sens.* 17(21), 3572 (2025). <https://doi.org/10.3390/rs17213572>
- [6] Zhi, Y., Wang, Y., Zhang, F., et al.: MSFFNet: multimodal spatial–frequency fusion network for RGB-DSM remote sensing image segmentation. *Remote Sens.* 17(22), 3745 (2025). <https://doi.org/10.3390/rs17223745>
- [7] Qiu, J., Chang, W., Ren, W., et al.: MMFNet: a Mamba-based multimodal fusion network for remote sensing image semantic segmentation. *Sensors* 25(19), 6225 (2025). <https://doi.org/10.3390/s25196225>
- [8] Tong, Y., Tang, M., Zhang, Y., et al.: BATFNet: boundary-aware transformer fusion network for RGB-DSM semantic segmentation of remote sensing images. *Sensors* 26(10), 3205 (2026). <https://doi.org/10.3390/s26103205>
- [9] Zeng, J., Xu, J., Jia, X., et al.: DLGCNet: multimodal remote sensing semantic segmentation via dual diagonal low-rank adaptation and graph convolutional feature fusion. *Knowl.-Based Syst.* 347, 116299 (2026). <https://doi.org/10.1016/j.knosys.2026.116299>
- [10] Chen, L., Fu, Y., Gu, L., et al.: Frequency-aware feature fusion for dense image prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* 46(12), 10763–10780 (2024). <https://doi.org/10.1109/TPAMI.2024.3449959>
- [11] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440 (2015). <https://doi.org/10.1109/CVPR.2015.7298965>
- [12] Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
- [13] Chen, L.C., Zhu, Y., Papandreou, G., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Computer Vision – ECCV 2018*, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
- [14] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., et al.: UNet++: a nested U-Net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
- [15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)* (2021)
- [16] Zheng, S., Lu, J., Zhao, H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6877–6886 (2021). <https://doi.org/10.1109/CVPR46437.2021.00681>
- [17] Xie, E., Wang, W., Yu, Z., et al.: SegFormer: simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* 34, 12077–12090 (2021)
- [18] Liu, Z., Lin, Y., Cao, Y., et al.: Swin Transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002 (2021). <https://doi.org/10.1109/ICCV48922.2021.00986>
- [19] Liu, H., Qian, Y., Zhong, X., et al.: Research on super-resolution reconstruction of remote sensing images: a comprehensive review. *Opt. Eng.* 60(10), 100901 (2021). <https://doi.org/10.1117/1.OE.60.10.100901>
- [20] Wang, L., Wang, X., Liu, W., et al.: A unified architecture for super-resolution and segmentation of remote sensing images based on similarity feature fusion. *Displays* 84, 102800 (2024). <https://doi.org/10.1016/j.displa.2024.102800>
- [21] Kim, J., Oh, J., Lee, K.M.: Beyond image super-resolution for image recognition with task-driven perceptual loss. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2651–2661 (2024). <https://doi.org/10.1109/CVPR52733.2024.00256>
- [22] Yuan, X., Shi, J., Gu, L.: A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* 169, 114417 (2021). <https://doi.org/10.1016/j.eswa.2020.114417>
- [23] Audebert, N., Le Saux, B., Lefèvre, S.: Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* 140, 20–32 (2018). <https://doi.org/10.1016/j.isprsjprs.2017.11.011>
- [24] 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.* 15, 3463–3474 (2022). <https://doi.org/10.1109/JSTARS.2022.3165005>
- [25] Ma, X., Zhang, X., Pun, M.O., et al.: A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* 62, 5403215 (2024). <https://doi.org/10.1109/TGRS.2024.3373033>

- [26] Ma, X., Xu, X., Zhang, X., et al.: Adjacent-scale multimodal fusion networks for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 17, 20116–20128 (2024). <https://doi.org/10.1109/JSTARS.2024.3486906>
- [27] Liu, Y., Gao, K., Wang, H., et al.: A Transformer-based multi-modal fusion network for semantic segmentation of high-resolution remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* 133, 104083 (2024). <https://doi.org/10.1016/j.jag.2024.104083>
- [28] Li, J., Zhang, W., Zhang, W., et al.: LMF-Net: a learnable multimodal fusion network for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 18, 3905–3920 (2025). <https://doi.org/10.1109/JSTARS.2025.3527213>
- [29] Li, Z., Zhang, Y., Hu, J., et al.: HDC-Net: a multimodal remote sensing semantic segmentation network with hierarchical dual-stream fusion and cross-token interaction. *Knowl.-Based Syst.* 337, 115416 (2026). <https://doi.org/10.1016/j.knosys.2026.115416>
- [30] Chen, J., Wang, F., Fang, J., et al.: MFMINet: multimodal fusion and cross-layer interaction network for semantic segmentation of high-resolution remote sensing images. *iScience* 29(2), 114705 (2026). <https://doi.org/10.1016/j.isci.2026.114705>
- [31] Wang, F., Yang, Y., Zhang, H., et al.: WAMNet: wavelet-enhanced asymmetric Mamba network for semantic segmentation of multimodal remote sensing images. *Neurocomputing* 685, 133665 (2026). <https://doi.org/10.1016/j.neucom.2026.133665>
- [32] Wang, J., Chen, K., Xu, R., et al.: CARAFE: content-aware reassembly of features. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3007–3016 (2019). <https://doi.org/10.1109/ICCV.2019.00310>
- [33] Liu, W., Lu, H., Fu, H., et al.: Learning to upsample by learning to sample. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6004–6014 (2023). <https://doi.org/10.1109/ICCV51070.2023.00554>
- [34] Fu, S., Hamilton, M., Brandt, L.E., et al.: FeatUp: a model-agnostic framework for features at any resolution. In: *International Conference on Learning Representations (ICLR)* (2024)
- [35] Yang, P., Zhou, W., Liu, Y.: Frequency-aware integrity learning network for semantic segmentation of remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 18, 3398–3409 (2025). <https://doi.org/10.1109/JSTARS.2024.3524753>
- [36] Chen, H., Feng, L., Wu, W., et al.: F2Net: a frequency-fused network for ultra-high resolution remote sensing segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13275–13284 (2026)
- [37] Cramer, M.: The DGPF-test on digital airborne camera evaluation—overview and test design. *Photogramm. Fernerkund. Geoinf.* 2010(2), 73–82 (2010). <https://doi.org/10.1127/1432-8364/2010/0041>
- [38] Xu, J., Xiong, Z., Bhattacharyya, S.P.: PIDNet: a real-time semantic segmentation network inspired by PID controllers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19529–19539 (2023). <https://doi.org/10.1109/CVPR52729.2023.01871>
- [39] Cao, Y., Xu, J., Lin, S., et al.: GCNet: non-local networks meet squeeze-excitation networks and beyond. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1971–1980 (2019). <https://doi.org/10.1109/ICCVW.2019.00246>
- [40] Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* 34, 17864–17875 (2021)