

Article

Hypergraph-Driven Heterogeneous Spatial Relationship Learning for Remote Sensing Segmentation

Qihao Zhang, Lankun Peng, Feiyang Hu and Xiaoming Xi *

School of Computer and Artificial Intelligence, Shandong Jianzhu University, Jinan 250101, China; qihaozhang.mail@gmail.com (Q.Z.); penglankun24@gmail.com (L.P.); hufeiyang0616@163.com (F.H.)

* Correspondence: fyzq10@126.com

Abstract

Remote sensing semantic segmentation is critical for extracting fine-grained spatial information in applications such as urban planning and environmental monitoring. However, existing methods face significant challenges in modeling heterogeneous spatial relationships within complex urban scenes, where semantically related regions are spatially dispersed yet functionally interdependent. Conventional convolutional neural networks exhibit limited receptive fields that fail to capture long-range dependencies, while Transformer-based approaches capture global dependencies but do not explicitly model regional heterogeneity, leading to blurred boundaries and category confusion. To address these limitations, this paper proposes a novel multi-relational-aware segmentation framework that leverages hypergraph theory to dynamically model higher-order semantic groupings across non-adjacent regions. The core innovation lies in a hypergraph structure learning unit that employs fuzzy clustering to partition multi-scale features into adaptive hyperedge sets, enabling joint representation of topological associations and functional dependencies among spatially distributed entities. Additionally, a multi-scale co-modeling strategy integrates stochastic feature masking with weighted fusion to bridge semantic abstraction and spatial localization. Experiments demonstrate that the proposed method achieves state-of-the-art mIoU performance on the LoveDA, Vaihingen, and Potsdam datasets, obtaining mIoU scores of 54.70%, 85.01%, and 87.64%, with improvements of 0.30%, 0.91%, and 0.08%, respectively, over the best existing methods.

Keywords: heterogeneous spatial relationships; higher-order relationships; hypergraph learning; remote sensing; semantic segmentation



Academic Editor: Masoud Daneshmand

Received: 22 July 2026

Revised: 22 August 2026

Accepted: 24 August 2026

Published: 27 August 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.

1. Introduction

Remote sensing semantic segmentation [1,2] plays a pivotal role in extracting fine-grained spatial information from high-resolution satellite or aerial imagery, serving as a cornerstone for applications such as urban planning, environmental monitoring, disaster management, and land use analysis. The ability to accurately distinguish and classify various land cover types—such as buildings, roads, vegetation, and water bodies—enables dynamic tracking of the evolution of urban systems and environmental changes [3]. With advancements in remote sensing technology, the resolution of high-resolution imagery has reached sub-meter levels (e.g., 0.3 m for WorldView-3). However, fine-grained categories [4] in complex urban scenes (e.g., distinctions between old and new building types, industrial and storage zones) and spatial heterogeneity (e.g., interwoven distributions of dense urban areas and green spaces) continue to pose significant challenges to existing methods.

Traditional approaches often encounter performance bottlenecks in such scenarios due to ambiguities in local features (e.g., texture similarities between shadows and buildings) or the absence of long-range semantic dependencies (e.g., functional relationships between roads and distant buildings) [5].

Among these challenges, heterogeneous relationships between spatially dispersed but semantically related regions represent a fundamental yet underexplored issue in remote sensing segmentation. In complex urban environments, different spatial regions often exhibit intricate semantic dependencies that extend beyond their immediate proximity. For instance, the functional type of a building (e.g., residential or commercial) can be inferred from its spatial adjacency to roads of specific widths or vegetation patterns, while water bodies and urban structures often form contextual clusters that define land use categories. As illustrated in Figure 1, conventional architectures are inherently constrained in modeling heterogeneous spatial relationships. The localized receptive fields of traditional convolutional neural networks [6] (CNNs), constrained by fixed ranges, struggle to model semantic dependencies between distant regions (e.g., roads and remote buildings). Although Transformers [7] can capture global correlations through self-attention mechanisms, standard self-attention does not explicitly model regional heterogeneity, such as semantic disparities or spatial layout variations among regions. This limitation may lead to suboptimal performance in preserving object boundaries and distinguishing visually similar categories.

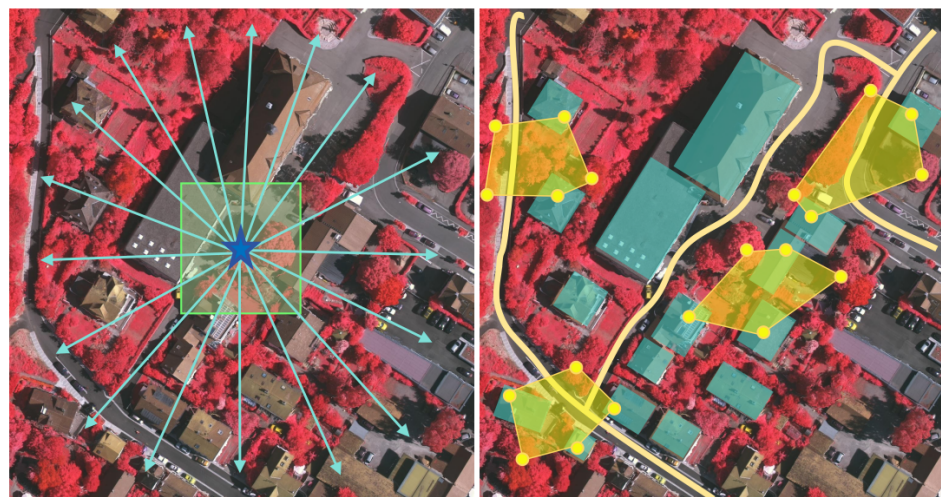


Figure 1. Limitations of conventional architectures in modeling heterogeneous spatial relationships for remote sensing segmentation. **(Left):** CNNs (fixed-range receptive field, green box) and Transformers (global self-attention, cyan arrows) fail to capture semantic dependencies between spatially dispersed regions (e.g., distant buildings and roads). The blue star denotes the reference region used to illustrate its interactions with other spatial regions. **(Right):** Higher-order functional semantic grouping through hyperedges (yellow edges) requires joint modeling of multiple non-adjacent spatial regions (marked by yellow nodes) to achieve accurate segmentation, which cannot be addressed by conventional methods.

Given these limitations, this study aims to address the dual challenges of preserving local details and understanding global context in remote sensing segmentation. While existing hybrid architectures [8] (e.g., UNetFormer) have significantly enhanced the representation capacity of complex scenes by integrating CNN-based local feature extraction with Transformer-based global attention mechanisms, their feature aggregation frameworks still exhibit limitations in explicitly modeling heterogeneous relationships among spatially distributed regions. Although these architectures can adaptively aggregate contextual information, they generally do not construct explicit region-level relational structures for organizing interactions among multiple semantically related regions.

Specifically, hybrid models typically capture local spatial patterns through convolutional operations and global contextual correlations through self-attention. However, such feature-level aggregation does not explicitly organize multiple spatially separated regions into structured relational groups. For example, semantic associations between roads and distant buildings may not be sufficiently represented by local convolutional features, while global self-attention may capture broad correlations without explicitly distinguishing region-specific heterogeneity, such as differences in semantic roles or spatial layouts. Consequently, heterogeneous relationships among spatially distributed regions may remain insufficiently represented, which can contribute to boundary ambiguity and category confusion. Therefore, explicitly constructing adaptive region-level relational structures for modeling higher-order heterogeneous interactions remains an important challenge in remote sensing semantic segmentation.

To overcome the limitations of static modeling frameworks, this study proposes the introduction of hypergraphs [9] as a dynamic modeling tool for heterogeneous spatial relationships. The core motivation lies in the fact that traditional graph structures or attention mechanisms can only model binary node relationships (e.g., interactions between pixel pairs or region pairs), whereas complex semantic associations in remote sensing scenes often involve higher-order relationships among multiple nodes. For instance, the functional category of a building is jointly defined by its roof structure, surrounding road networks, and the distribution patterns of adjacent vegetation—a multi-element collaborative semantic representation that cannot be fully captured through simple binary relationship modeling. Hypergraphs, through the mathematical structure of hyperedges, naturally enable the joint modeling of multiple spatial regions or feature nodes as a unified entity, thereby explicitly encoding synergistic constraints between distant regions (e.g., topological associations between roads and buildings) or semantic dependencies of local structures (e.g., geometric relationships between building contours and shadows). Compared to graph convolutional networks [10] (GCNs) or self-attention mechanisms, hypergraphs excel in modeling higher-order relationships—bypassing the need to approximate complex interactions through combinations of binary edges and instead directly defining joint probability distributions or feature aggregation rules among multiple nodes via hyperedges. This characteristic allows hypergraphs to more accurately characterize spatially dispersed but semantically coupled regional clusters (e.g., the coordinated layout of airport runways and aprons) or distinguish visually similar yet functionally distinct categories (e.g., differentiating between industrial and storage building clusters). Thus, the incorporation of hypergraphs addresses a representational limitation of conventional pairwise relational modeling by providing an adaptive mechanism for higher-order feature aggregation.

To achieve these objectives, this study proposes a multi-relational-aware semantic segmentation framework, the Hypergraph Network (HyperNet) for remote sensing segmentation, designed to address the modeling of heterogeneous spatial relationships in remote sensing scenes. Its core architecture is based on a hybrid Transformer-CNN encoder-decoder structure, which extracts multi-scale features through the synergistic design of global context modeling and local detail capture [11]. To transcend the limitations of static feature aggregation in conventional methods, we introduce a hypergraph modeling module dedicated to capturing heterogeneous spatial relationships, with its core innovation lying in the design of a hypergraph structure learning unit. This module employs a fuzzy clustering algorithm to partition multi-scale feature nodes (e.g., spatial regions or semantic segments) extracted by the decoder into dynamic hyperedge sets, where each hyperedge can jointly model higher-order relationships among multiple nodes (e.g., topological associations between building contours and road networks, semantic dependencies between vegetation distribution and functional zones). Through hypergraph convolution mechanisms, node features undergo

non-local information propagation under the constraints of hyperedges, thereby explicitly encoding multi-faceted relationships among geographical entities (e.g., adjacency, exclusion, or functional synergy). Additionally, we propose a multi-scale co-modeling strategy: dynamic masking of encoder features via a stochastic perturbation mechanism forces the model to learn more robust cross-scale representations, while a weighted feature fusion module adaptively integrates perturbed deep features with shallow decoder details to bridge the representational gap between semantic abstraction and spatial localization. By leveraging the synergy between higher-order relationship modeling via hypergraphs and dynamic multi-scale feature interactions, our model provides novel methodological support for the precise parsing of complex spatial semantics in remote sensing scenes.

Our contributions are threefold:

- We propose a hypergraph-driven remote sensing segmentation framework, termed HyperNet, which jointly models heterogeneous spatial relationships and multi-scale feature representations for accurate segmentation of complex remote sensing scenes.
- We develop a Hypergraph-Driven Functional Semantics Module (HFSM) and a Multi-Scale Co-Modeling Module (MCM). HFSM dynamically constructs hyperedges through fuzzy clustering to capture higher-order semantic relationships, while MCM combines stochastic feature masking with adaptive weighted fusion to enhance cross-scale feature robustness.
- Extensive mechanism-level and experimental analyses demonstrate the effectiveness of the proposed HyperNet. Specifically, HyperNet achieves state-of-the-art mIoU scores of 54.70%, 85.01%, and 87.64% on LoveDA, Vaihingen, and Potsdam, outperforming existing methods by 0.30%, 0.91%, and 0.08%, respectively.

2. Related Work

Existing remote sensing segmentation methods mainly rely on CNN-based and Transformer-based architectures. CNN-based methods focus on local feature extraction and multi-scale representation, while Transformer-based methods enhance global context modeling. However, explicit modeling of heterogeneous spatial relationships among multiple regions remains limited.

2.1. CNN-Based Methods

The fully convolutional network (FCN) [12] revolutionized semantic segmentation by enabling end-to-end pixel-wise classification, rapidly becoming the cornerstone of remote sensing applications. However, its simplified decoder architecture inherently struggles to preserve spatial details, leading to coarse segmentation results in complex remote scenes. This limitation spurred the development of more sophisticated encoder-decoder frameworks.

The U-Net [13] architecture addressed this challenge by introducing symmetric skip connections to integrate low-level spatial details with high-level semantic features. This design significantly improved segmentation accuracy in medical imaging and was quickly adopted in remote sensing. For example, U-Net achieved over 85% accuracy in building and road extraction on the ISPRS Vaihingen dataset. Recent advancements have further optimized U-Net's skip connections and decoders. RCFSNet [14] addresses these limitations. Another innovation, CM-UNet [15], replaces traditional skip connections with a multi-scale attention aggregation module (MSAA). This enhancement improves feature fusion across resolutions and achieves state-of-the-art performance on high-resolution satellite imagery.

Despite these improvements, traditional CNNs remain limited by their inability to model global dependencies. In urban scenes with dense buildings and roads, CNNs with restricted receptive fields struggle to distinguish complex patterns. To overcome this,

multi-scale feature aggregation and attention mechanisms have been widely explored. By integrating attention gates, Attention U-Net [16] focuses on salient spatial features to enhance segmentation accuracy in difficult areas. DANet [17] utilizes dual attention modules for spatial and channel-wise feature refinement, achieving superior performance on various segmentation benchmarks. Lightweight architectures such as BiSeNetv2 [18] balance accuracy and speed by combining spatial detail with context information through a two-branch structure, suitable for real-time remote sensing applications. MST-DeepLabv3+ [19] introduces the SENet attention mechanism module and transfer learning to better enhance semantic segmentation accuracy. LOGCAN++ [20] introduces GCA attention modules to capture global representations for class-level context modeling, thus reducing background noise interference and improving segmentation mIoU by 1.23% in the Vaihingen dataset. These methods improve local representation through multi-scale aggregation and attention mechanisms. However, CNN-based architectures still rely mainly on local feature extraction and feature-level enhancement. Their ability to model semantic relationships between spatially separated regions remains limited.

2.2. Transformer-Based Methods

Unlike conventional Convolutional Neural Network architectures, Transformer-based methods reformulate two-dimensional image processing tasks as the one-dimensional sequence modeling problems [21]. Leveraging its robust sequence-to-sequence learning capabilities, the Transformer architecture exhibits superior performance in capturing global contextual relationships compared to attention mechanisms operating in isolation [22]. This enhanced modeling capacity has enabled Transformer-based approaches to achieve state-of-the-art performance across core computer vision applications, including but not limited to image classification [23], object detection [24], and semantic segmentation [25].

Transformer-based methods for remote sensing segmentation can be broadly categorized into two groups. The first category consists of pure Transformer architectures, exemplified by ResU-Former [26], which integrates Swin residual Transformer blocks for multi-scale feature fusion. This approach balances visual-semantic spatial information via hierarchical structures and employs dynamic sparse training to reduce complexity. A notable limitation of this approach is its lack of CNN's local feature extraction capabilities, often leading to deteriorated segmentation accuracy for small targets [27]. The second category comprises hybrid CNN-Transformer architectures that address the limitations of pure Transformer designs [28]. For instance, SSNet innovatively integrates Feature Fusion and Feature Injection Modules (FFM & FIM), achieving mIoU scores of 61.93% on the WHDLD and 77.67% on the Potsdam datasets by constructing a dual-branch cooperative mechanism, which effectively enhances the discrimination of similar objects and small targets. UNetFormer [8] employs CNN-based feature encoding and Transformer-based decoding. However, these approaches mainly focus on global context modeling and feature fusion. They do not explicitly model heterogeneous spatial relationships among multiple regions. This limitation reduces their ability to represent complex spatial dependencies in remote sensing scenes.

2.3. Graph-Based and Hypergraph-Based Methods

Graph-based methods have been explored to model structural relationships in remote sensing images. KGGCN [29] constructed graphs based on geographic objects and their topological relationships for semantic segmentation. A hierarchical class graph framework [30] was introduced to capture structured correlations among different ground-object categories. More recently, MGF-GCN [31] employed graph convolution to model spatial correlations between multimodal remote sensing features.

Although graph-based methods improve relational representation, conventional graphs mainly model relationships through pairwise edges between nodes. Hypergraphs extend this representation by associating multiple nodes within a hyperedge. DNHGNN [32] combines multi-scale superpixel segmentation with hypergraph convolution to capture higher-order relationships among remote sensing objects for change detection. HGBT [33] further introduces hypergraph modeling into a Transformer architecture to capture high-order correlations and local topology for remote sensing semantic segmentation.

These methods demonstrate the potential of graph and hypergraph learning for remote sensing interpretation. However, adaptive modeling of heterogeneous spatial relationships among distributed regions remains insufficiently explored. In contrast, HyperNet constructs adaptive hyperedges from feature representations through fuzzy clustering. This enables higher-order relationships among multiple spatial regions to be dynamically modeled.

Overall, existing methods have improved local and global feature representation, while graph and hypergraph learning further enhances relational modeling. However, adaptive modeling of heterogeneous spatial relationships among distributed regions remains insufficiently explored. The robustness of cross-scale feature interactions also requires further investigation. These limitations motivate the design of HyperNet. HFSM models heterogeneous spatial relationships, while MCM enhances robust multi-scale feature interaction.

3. Methodology

3.1. Mechanism Analysis of Hypergraph Modeling

Before introducing the specific methodological design, we analyze the mechanism of hypergraph modeling from the perspective of feature aggregation and relational propagation. Unlike conventional pairwise relations, a hyperedge can associate multiple feature nodes within a unified structure. This property enables information exchange among spatially distributed but semantically related regions. Therefore, hypergraph modeling provides a natural way to represent higher-order relationships in complex remote sensing scenes.

3.1.1. Parameterized Feature Aggregation for Hypergraphs

Let the feature aggregation operation of the hypergraph model be defined as

$$\mathbf{Z}_v = \sigma\left(\mathbf{H} \text{softmax}\left(\mathbf{H}^T \mathbf{X} \mathbf{W}_e\right) \mathbf{W}_v\right). \quad (1)$$

where $\mathbf{H} \in \mathbb{R}^{N \times M}$ is the incidence matrix, $\mathbf{W}_e$ and $\mathbf{W}_v$ are learnable parameters, and $\mathbf{X} \in \mathbb{R}^{N \times D}$ is the node feature matrix. Each column of $\mathbf{H}$ describes the associations between feature nodes and a hyperedge.

The aggregation consists of two stages. First, $\mathbf{H}^T \mathbf{X} \mathbf{W}_e$ aggregates the projected node features into hyperedge representations. The hyperedge features are then propagated back to their associated nodes through $\mathbf{H} \mathbf{Z}_e \mathbf{W}_v$. This node-to-hyperedge and hyperedge-to-node process enables related nodes to exchange information through shared hyperedges.

Unlike conventional pairwise interactions, one hyperedge can connect multiple nodes simultaneously. Thus, the resulting feature representation contains group-level contextual information from several related regions. This property provides the basic mechanism for higher-order relationship modeling in HyperNet.

3.1.2. Intuitive Effect of Shared Hyperedge Aggregation

Nodes belonging to the same hyperedge participate in a shared feature aggregation process. Their representations are jointly integrated into a common hyperedge feature and then redistributed to the associated nodes. This process encourages semantically related nodes to exchange complementary information and learn more consistent representations.

From an intuitive perspective, shared aggregation can also reduce the influence of isolated local feature variations. A node is no longer updated only according to its individual representation. Instead, its updated feature also incorporates information from other related nodes in the same hyperedge. This group-wise interaction may lead to smoother feature updates during training.

However, this effect depends on the learned hypergraph structure, feature distribution, and optimization process. Therefore, we interpret it as a mechanism-level explanation rather than a general guarantee of gradient variance reduction.

3.1.3. Optimization Interpretation

The higher-order connections introduced by hypergraphs provide direct information propagation among multiple related nodes. This property is particularly useful for remote sensing images, where semantically related regions may be spatially separated. Through shared hyperedges, information can be exchanged across these regions without relying solely on isolated local interactions.

Such relational propagation can reduce information fragmentation among related regions. It can also provide more consistent contextual information for feature learning. From an optimization perspective, these properties may facilitate a more stable learning process and improve feature interaction during training.

Nevertheless, the convergence behavior of a deep segmentation network is affected by many factors, including the network architecture, loss function, optimizer, and learned hypergraph structure. Therefore, we do not assume that hypergraph modeling guarantees a faster convergence rate. Its practical optimization behavior is evaluated through the experimental results.

3.1.4. Higher-Order Connectivity Interpretation

Conventional graphs generally describe relationships through pairwise edges between nodes. In contrast, a hyperedge can directly connect multiple nodes. This higher-order connectivity is suitable for representing complex dependencies among geographical entities in remote sensing scenes.

For example, the semantic interpretation of a building may depend jointly on its surrounding road network and vegetation distribution. These relationships can be represented by a hyperedge as

$$e_m = \{v_{\text{building}}, v_{\text{road}}, v_{\text{vegetation}}\}. \quad (2)$$

The associated nodes can first contribute information to the shared hyperedge and then receive the aggregated contextual information. In this way, the relationships among multiple regions are modeled within a unified structure rather than through several independent pairwise connections.

The effectiveness of this propagation depends on the quality of the constructed hyperedges and their association strengths. Therefore, we do not assume that hypergraphs universally possess superior spectral properties to conventional graphs. Instead, their advantage in this work lies in the explicit representation and propagation of higher-order information among multiple spatial regions.

In summary, the above analysis provides an intuitive interpretation of hypergraph modeling rather than formal optimization guarantees. Hyperedges organize multiple related spatial regions into shared relational structures and enable group-wise information propagation. These properties facilitate the modeling of higher-order heterogeneous spatial relationships in remote sensing scenes. Based on this motivation, we further design the Hypergraph-Driven Functional Semantics Module in Section 3.5 to dynamically construct and learn hypergraph relationships from remote sensing features.

3.2. Overall Architecture

Our proposed HyperNet architecture is illustrated in Figure 2. The proposed framework HyperNet addresses the limitations of static feature aggregation mechanisms in conventional approaches by integrating hypergraph-driven higher-order relation modeling into a hybrid Transformer-CNN architecture. Its overall structure is formulated in Equation (3) as:

$$\mathcal{F} = \mathcal{D}(\mathcal{E}(\mathbf{X})) \rightarrow \mathcal{H}(\mathcal{F}) \rightarrow \mathcal{L}(\mathbf{Y}_{\text{pred}}), \quad (3)$$

where $\mathcal{E}$ denotes the hybrid encoder, $\mathcal{D}$ represents the multi-scale co-occurrence modeling decoder, $\mathcal{H}$ is the hypergraph-driven functional semantics module, and $\mathcal{L}$ stands for the hybrid loss function. The detailed design of each component is described below.

3.3. Hybrid Transformer-CNN Encoder

The encoder $\mathcal{E}$ employs a hierarchical architecture that balances global context capture with local detail preservation through window-based self-attention and cross-window interaction modules, as expressed in Equation (4):

$$\mathbf{Z}_i = \text{Window-MSA}(\mathbf{Z}_{i-1}) + \text{Cross-Window-Interaction}(\mathbf{Z}_{i-1}), \quad (4)$$

where Window-MSA preserves fine-grained structures via self-attention within localized windows, while cross-window interaction aggregates dependencies across windows through a global-local dual-branch design. The resulting multi-scale feature maps $\{\mathbf{F}_l\}_{l=1}^L$ encode a spatial hierarchy from coarse to fine, with exponentially decreasing spatial resolutions across layers.

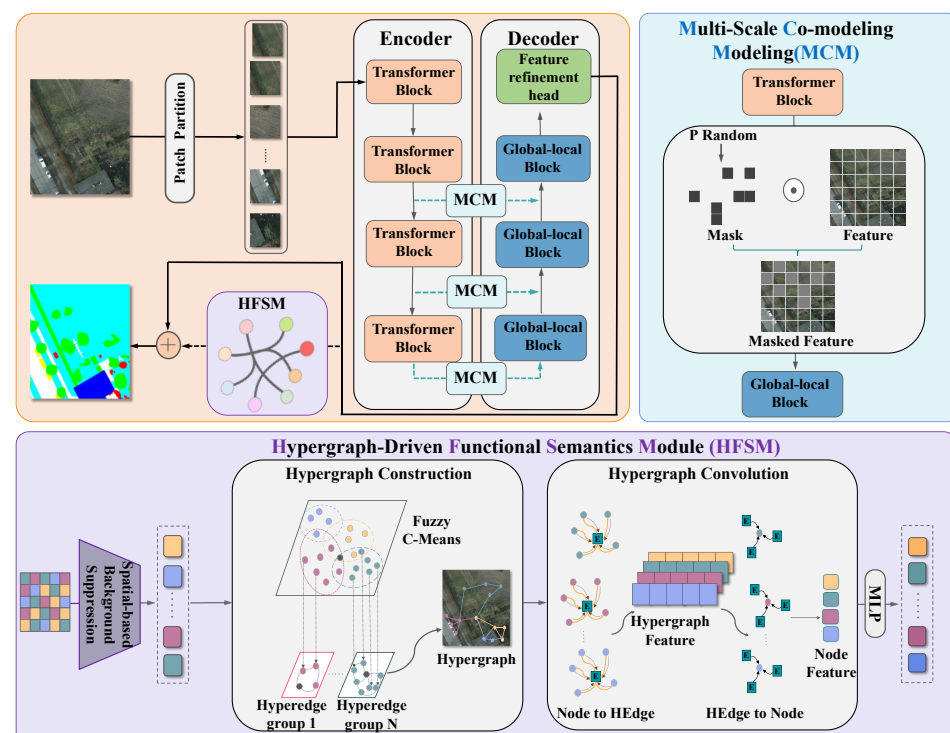


Figure 2. In our proposed HyperNet, the input image is first divided into multiple patches, which are then fed to a Transformer encoder. The encoder consists of four stages, each generating features at a different scale. Stochastic perturbations are then introduced via the MCM module. Subsequently, these features are processed by a decoder and passed to the HFSM module for hypergraph structure learning. Finally, a segmentation head produces the final segmentation map from the learned features.

3.4. Multi-Scale Co-Modeling Module

The decoder $\mathcal{D}$ facilitates dynamic multi-scale feature interaction through perturbation-driven feature refinement and adaptive fusion. For each encoder feature $\mathbf{F}_l$, a stochastic mask $\mathbf{M}_l = \mathcal{P}(\mathbf{F}_l; p)$ is generated with a suppression ratio p :

$$\mathbf{F}_{\text{masked},l} = \mathbf{F}_l \odot \mathbf{M}_l, \quad (5)$$

where p controls the proportion of suppressed feature responses. The stochastic masking introduces controlled feature perturbations and reduces the model's dependence on individual local activations.

The perturbed feature is then adaptively fused with the decoder feature $\mathbf{F}_{\text{dec}}$:

$$\mathbf{F}_{\text{fuse},l} = \alpha \odot \mathbf{F}_{\text{masked},l} + (1 - \alpha) \odot \mathbf{F}_{\text{dec}}. \quad (6)$$

The fusion coefficient α is dynamically generated through a learnable gating mechanism:

$$\alpha = \sigma(\mathbf{W}_\alpha \text{Concat}(\mathbf{F}_{\text{masked},l}, \mathbf{F}_{\text{dec}})), \quad (7)$$

where $\mathbf{W}_\alpha$ denotes a learnable transformation and $\sigma(\cdot)$ is the sigmoid function. The resulting gating map $\alpha \in [0, 1]$ adaptively controls the relative contributions of the perturbed encoder feature and the decoder feature. This design introduces controlled feature perturbations during training while preserving complementary contextual information through adaptive feature fusion.

3.5. Hypergraph-Driven Functional Semantics Module

To model higher-order spatial relationships, the functional semantics module $\mathcal{H}$ constructs a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{H})$. The node set $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ consists of feature tokens extracted from the decoder output $\mathbf{F}_{\text{dec}}$, where each node $v_i \in \mathbb{R}^D$ represents the semantic information of a spatial region.

We adopt fuzzy c-means (FCM) clustering to construct hyperedges. Unlike hard clustering, which assigns each node to a single cluster, FCM represents the association between a node and different clusters using continuous membership degrees. This property is consistent with the heterogeneous relationships in remote sensing scenes, where a spatial region may exhibit semantic associations with multiple surrounding regions rather than belonging to a single relational group.

Given node features $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, FCM minimizes the following objective:

$$J_{\text{FCM}} = \sum_{i=1}^N \sum_{m=1}^M \mu_{im}^q \|x_i - c_m\|_2^2, \quad (8)$$

subject to

$$\sum_{m=1}^M \mu_{im} = 1, \quad 0 \leq \mu_{im} \leq 1, \quad (9)$$

where c_m denotes the center of the m -th cluster, μ_{im} denotes the membership degree of node v_i to cluster m , and $q > 1$ is the fuzziness coefficient.

The continuous membership μ_{im} provides a natural measure of the node–hyperedge association strength. A node can therefore contribute to multiple relational groups with different degrees, which avoids the exclusive assignment imposed by hard clustering. This property is particularly suitable for constructing weighted hyperedges from ambiguous and heterogeneous remote sensing features.

The hyperedge set $\mathcal{E} = \{e_1, e_2, \dots, e_M\}$ is then constructed according to the FCM memberships. Only nodes satisfying $\mu_{im} \geq \tau$ are retained in hyperedge e_m :

$$e_m = \{v_i \mid \mu_{im} \geq \tau\}. \quad (10)$$

In all experiments, the membership threshold τ is fixed at 0.1. This threshold filters weak node–hyperedge associations while allowing a node to participate in multiple hyperedges when meaningful soft memberships are retained. The same value is used across all datasets for consistent hypergraph construction.

The weighted incidence matrix $\mathbf{H} \in \mathbb{R}^{N \times M}$ is defined as

$$H_{im} = \begin{cases} \frac{\mu_{im}}{\sum_{j: \mu_{jm} \geq \tau} \mu_{jm}}, & \mu_{im} \geq \tau, \\ 0, & \mu_{im} < \tau. \end{cases} \quad (11)$$

where the normalized membership degree represents the relative association strength between node v_i and hyperedge e_m .

Hypergraph convolution performs relational feature propagation in two stages of message passing. In the node-to-hyperedge aggregation phase, the feature matrix $\mathbf{X}$ is projected via a learnable weight matrix $\mathbf{W}_e \in \mathbb{R}^{D \times D}$ and aggregated to hyperedges using the transpose of $\mathbf{H}$:

$$\mathbf{Z}_e = \text{softmax}(\mathbf{H}^T \mathbf{X} \mathbf{W}_e), \quad (12)$$

which maps multi-node features into hyperedge-level semantic representations. Subsequently, during the hyperedge-to-node diffusion phase, hyperedge features $\mathbf{Z}_e$ are propagated back to nodes via a learnable weight matrix $\mathbf{W}_v \in \mathbb{R}^{D \times D}$:

$$\mathbf{Z}_v = \text{softmax}(\mathbf{H} \mathbf{Z}_e \mathbf{W}_v). \quad (13)$$

where $\mathbf{W}_v$ enforces consistency among hyperedge features. The final relational embedding is obtained through a residual connection:

$$\mathbf{Z}' = \mathbf{Z}_v + \mathbf{X}, \quad (14)$$

which preserves the original node features while incorporating higher-order relational information. The resulting representation is then used for downstream segmentation.

3.6. Hybrid Loss Function

The loss function $\mathcal{L}$ combines Softmax cross-entropy ($\mathcal{L}_{ce}$) and Dice loss ($\mathcal{L}_{dice}$). Cross-entropy loss provides pixel-wise classification supervision, whereas Dice loss encourages region-level overlap. The two losses are defined as

$$\begin{aligned} \mathcal{L}_{ce} &= - \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log \hat{y}_{ik}, \\ \mathcal{L}_{dice} &= 1 - \frac{1}{K} \sum_{k=1}^K \frac{2 \sum_{i=1}^N \hat{y}_{ik} y_{ik}}{\sum_{i=1}^N (\hat{y}_{ik}^2 + y_{ik}^2)}, \\ \mathcal{L} &= \mathcal{L}_{ce} + \lambda \mathcal{L}_{dice}, \end{aligned} \quad (15)$$

where N and K denote the number of pixels and semantic classes, respectively. The weighting coefficient λ is set to 1.5 to balance the contributions of the two loss terms.

4. Datasets

We evaluate HyperNet on three widely used remote sensing semantic segmentation benchmarks, including the ISPRS Vaihingen dataset, the ISPRS Potsdam dataset, and LoveDA. These datasets cover different geographical regions, spatial resolutions, and scene characteristics. Their main imaging properties are summarized in Table 1.

Table 1. Geographical and imaging characteristics of the datasets used in this study. The reported coordinates represent approximate central locations of the corresponding study areas rather than the coordinates of individual image tiles.

Dataset	Study Area	Latitude	Longitude	GSD	Image Source/Product	Bands Used	Atmospheric Correction
Vaihingen	Vaihingen, Germany	48.936° N	8.960° E	0.09 m	Airborne TOP	NIR, R, G	Not specified
Potsdam	Potsdam, Germany	52.399° N	13.066° E	0.05 m	Airborne TOP	R, G, B	Not specified
LoveDA	Nanjing, China	32.062° N	118.778° E	0.30 m	Google Earth	R, G, B	Not specified
LoveDA	Changzhou, China	31.774° N	119.954° E	0.30 m	Google Earth	R, G, B	Not specified
LoveDA	Wuhan, China	30.583° N	114.267° E	0.30 m	Google Earth	R, G, B	Not specified

4.1. Vaihingen Dataset

The ISPRS Vaihingen dataset consists of 33 very-high-resolution true orthophoto (TOP) tiles acquired over Vaihingen an der Enz, Germany. The tiles have varying spatial sizes, with an average size of approximately 2494×2064 pixels, and a ground sampling distance (GSD) of 0.09 m. The TOP images are provided as three-band images corresponding to near-infrared (NIR), red, and green. Although the benchmark also provides digital surface model (DSM) data, only the three optical TOP bands are used in our experiments.

The semantic annotations contain six categories: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background.

4.2. Potsdam Dataset

The ISPRS Potsdam dataset contains 38 very-high-resolution TOP tiles acquired over Potsdam, Germany. Each tile has a spatial size of 6000×6000 pixels with a GSD of 0.05 m. The benchmark provides the TOP imagery in RGB, IRRG, and RGBIR channel compositions, together with DSM data. In this study, only the RGB bands are used, without incorporating DSM or normalized DSM information. The semantic categories are identical to those of the Vaihingen dataset.

4.3. LoveDA Dataset

LoveDA is a high-spatial-resolution land-cover dataset containing 5987 remote sensing images collected from urban and rural areas of Nanjing, Changzhou, and Wuhan, China. The images were obtained from the Google Earth platform and have a GSD of 0.3 m and a spatial size of 1024×1024 pixels. The released images are three-channel RGB images. The dataset contains seven semantic categories: background, building, road, water, barren land, forest, and agriculture.

We follow the official training-validation-test split, consisting of 2522 training images, 1669 validation images, and 1796 test images. The inclusion of both urban and rural scenes introduces substantial variations in object scale, background appearance, and class distribution.

The exact wavelength ranges or spectral response functions of the released benchmark images are not specified in the official dataset documentation. Therefore, Table 1 reports the spectral bands actually used in our experiments rather than assigning approximate wavelength intervals that are not provided by the dataset sources.

The Vaihingen and Potsdam images are distributed as true orthophoto products, indicating that geometric and orthorectification processing has been performed during dataset

generation. However, the official benchmark documentation does not explicitly specify whether atmospheric correction was applied to the released image products. Atmospheric correction is likewise not explicitly documented for the released LoveDA images. We therefore used the officially released benchmark images directly and did not perform any additional atmospheric correction.

5. Experiments

5.1. Evaluation Metrics

We adopted the overall accuracy (OA), mean F1 Score (F1), and mean intersection over union (mIoU) as the evaluation metrics. For LoveDA, only mIoU is reported following its standard evaluation protocol. The model is trained using only these datasets, without any additional dataset for supervised training.

5.2. Implementation Details

To ensure a fair comparison with other methods, we followed the settings used in previous methods. For the Vaihingen dataset, we exclusively utilized the optical spectral bands of TOP imagery, excluding DSM/NDSM data. The dataset was divided into 17 test tiles (IDs: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38) and 16 training tiles. For the Potsdam dataset, we selected 14 test tiles (IDs: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13) and 24 training tiles. The listed IDs are the official identifiers of spatially distinct orthophoto tiles in the ISPRS benchmarks and are used only to specify the training/test split. They do not represent individual surface types; each tile contains multiple land-cover classes, including impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. LoveDA follows the official training-validation-test split. All original tiles were cropped into 1024×1024 patches. During training, the input images were randomly cropped into 512×512 patches. Data augmentation included random scaling with factors of $\{0.5, 0.75, 1.0, 1.25, 1.5\}$, random vertical flipping, random horizontal flipping, and random rotation.

All models were trained for 80 epochs with a batch size of 2 using the AdamW optimizer. The initial learning rate was set to 6×10^{-4} for Vaihingen and 5×10^{-4} for LoveDA and Potsdam. The weight decay was set to 2.5×10^{-4} for Vaihingen and Potsdam and 0.01 for LoveDA. The backbone weight decay was set to the same value as the corresponding weight decay. No learning-rate warmup strategy was used. The learning rate was adjusted using the CosineAnnealingWarmRestarts scheduler with $T_0 = 15$ and $T_{\text{mult}} = 2$. The number of data-loading workers was set to 4. The main comparison experiments and key component ablation experiments were independently repeated three times using random seeds 42, 43, and 44. The corresponding results are reported as mean $\pm$ standard deviation.

During inference, each full-resolution test tile is first padded along the bottom and right boundaries so that its spatial dimensions are divisible by the patch size. The padded tile is then partitioned into non-overlapping 512×512 patches with a stride of 512 pixels. Each patch is processed independently, and the predictions are stitched back to their original spatial locations and cropped to the original tile size. Therefore, inference is performed using fixed-size, non-overlapping grid patches rather than feeding the entire tile into the network in a single forward pass. The same test-time flip augmentation settings were applied to HyperNet, the internally evaluated FT-UNetFormer baseline, and all ablation configurations. The results of the other state-of-the-art methods in the comparison tables were taken from their original publications; therefore, their test-time augmentation settings follow the corresponding original implementations and were not controlled by us.

In the HFSM module, the ratio of selected patch tokens a in spatial-based background suppression and the number of node clusters b for hypergraph construction were set to

0.3 and 5, respectively. The membership threshold τ for hyperedge construction was fixed at 0.1. In the MCM, the perturbation ratio p for each stage was set to 0.1. All experiments were conducted using PyTorch 1.8.1 on NVIDIA GeForce RTX 2080 Ti GPUs (NVIDIA Corporation, Santa Clara, CA, USA). To facilitate reproducibility, the source code, training and evaluation configurations, and trained model checkpoints will be publicly released upon publication.

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

We compare the proposed HyperNet with existing state-of-the-art CNN-based and Transformer-based methods on the LoveDA, Vaihingen, and Potsdam datasets. The results of HyperNet are reported as mean $\pm$ standard deviation over three independent runs with random seeds 42, 43, and 44, while the results of the compared methods are taken from their original publications.

LoveDA: Table 2 compares HyperNet with existing CNN-based and Transformer-based methods. HyperNet achieves an mIoU of $54.70 \pm 0.12\%$, obtaining state-of-the-art performance on the LoveDA dataset. Compared with UNetFormer and the best existing result of ViT-G12X4, HyperNet achieves numerical mIoU improvements of 2.30% and 0.30%, respectively. These results demonstrate the effectiveness of modeling heterogeneous spatial relationships in complex remote sensing scenes.

Vaihingen: As shown in Table 3, HyperNet achieves an F1 score of $91.70 \pm 0.18\%$, an mIoU of $85.01 \pm 0.11\%$, and an OA of $93.50 \pm 0.06\%$. In particular, HyperNet improves the mIoU from 84.10% for FT-UNetFormer to 85.01%, corresponding to an improvement of 0.91%. It also achieves the highest mIoU among the compared methods. The improvement indicates that incorporating higher-order heterogeneous spatial relationship modeling into the baseline provides additional benefits for remote sensing segmentation.

Potsdam: Table 4 compares HyperNet with existing CNN-based and Transformer-based methods on the Potsdam dataset. HyperNet achieves an F1 score of $93.29 \pm 0.05\%$, an mIoU of $87.64 \pm 0.08\%$, and an OA of $92.13 \pm 0.04\%$.

Table 2. Comparison experiments with other state-of-the-art methods on LoveDA dataset.

Method	Model	Year	mIoU
HRNet32 [34]	CNN	2021	49.79
LSKNet-S [35]	CNN	2023	54.00
LSKNet-T [35]	CNN	2023	53.20
Hi-ResNet [36]	CNN	2024	52.60
LWGANet L2 [37]	CNN	2025	53.60
CM-UNet [15]	CNN	2024	52.17
LOGCAN++ [20]	CNN	2025	53.35
UNetFormer [8]	Transformer	2021	52.40
ViT-B+RVSA-UperNet [38]	Transformer	2022	51.95
VITAE-BRVSA-UperNet [38]	Transformer	2022	52.44
ViT-G12X4 [39]	Transformer	2023	54.40
AerialFormer [40]	Transformer	2023	54.10
SelectiveMAE+ViT-L [41]	Transformer	2024	54.31
MAE+MTP(ViT-L+RVSA) [42]	Transformer	2024	54.17
IMP+MTP(internimage-XL) [42]	Transformer	2024	54.17
DecoupleNet D2 [43]	Transformer	2024	53.10
MAE+MTP(ViT-B+RVSA) [42]	Transformer	2024	52.39
HyperNet (Ours)	Transformer	2025	54.70 $\pm$ 0.12

Note: Bold indicates the best result.

Compared with LSKNet-S, which achieves the highest mIoU among the CNN-based methods in Table 4, HyperNet improves the mIoU from 87.20% to 87.64%, corresponding to a numerical improvement of 0.44%. Compared with FT-UNetFormer, HyperNet obtains numerical improvements of 0.14% in mIoU and 0.13% in OA, while its F1 score is 0.01% lower. Compared with DC-Swin, HyperNet improves F1, mIoU, and OA by 0.04%, 0.08%, and 0.13%, respectively. In addition, HyperNet improves the mIoU over PDSSNet by 0.09%.

Table 3. Comparison experiments with other state-of-the-art methods on the Vaihingen dataset.

Method	Model	Year	F1	mIoU	OA
DANet [17]	CNN	2019	79.60	69.40	88.20
ABCNet [44]	CNN	2021	-	-	90.70
MANet [45]	CNN	2021	-	-	90.96
CMTFNet [46]	CNN	2023	87.42	77.95	88.71
UPerNet(SAP) [47]	CNN	2024	-	73.27	90.1
ESDINet [48]	CNN	2024	89.99	82.03	90.88
DPFE [49]	CNN	2024	-	-	91.30
RS3Mamba [50]	CNN	2024	90.70	83.30	93.20
BANet [51]	Transformer	2021	-	-	90.5
DC-Swin [52]	Transformer	2021	90.7	-	91.6
Segmenter [25]	Transformer	2021	84.10	73.60	88.10
UNetFormer [8]	Transformer	2021	90.4	82.7	91.0
FT-UNetFormer [8]	Transformer	2021	91.32	84.10	91.62
SFA-Net [53]	Transformer	2024	91.20	-	-
HyperNet (Ours)	Transformer	2025	91.70 ± 0.18	85.01 ± 0.11	93.50 ± 0.06

Note: The best results are highlighted in bold.

Table 4. Comparison experiments with other state-of-the-art methods on the Potsdam dataset.

Method	Model	Year	F1	mIoU	OA
MAResU-Net [54]	CNN	2022	91.89	85.22	90.28
MANet [45]	CNN	2021	-	-	91.32
ABCNet [44]	CNN	2021	-	-	91.30
PSPNet(SAP) [47]	CNN	2024	-	74.30	88.56
LSKNet-S [35]	CNN	2024	93.10	87.20	92.00
DPFE [49]	CNN	2024	-	-	89.70
BANet [51]	Transformer	2021	-	-	91.06
UNetFormer [8]	Transformer	2021	92.80	86.80	91.30
FT-UNetFormer [8]	Transformer	2021	93.30	87.50	92.00
DC-Swin [52]	Transformer	2021	93.25	87.56	92.00
Mask2Former [55]	Transformer	2022	92.37	86.03	90.83
MPCNet [56]	Transformer	2023	92.29	85.91	90.56
RS-Mamba [57]	Transformer	2024	90.50	82.87	90.98
PyMamba [58]	Transformer	2024	90.30	82.52	90.91
Samba [59]	Transformer	2024	90.42	82.75	90.99
SFA-Net [53]	Transformer	2024	93.50	-	-
PDSSNet [60]	Transformer	2025	93.25	87.55	91.84
HyperNet (Ours)	Transformer	2025	93.29 ± 0.05	87.64 ± 0.08	92.13 ± 0.04

Note: The best results are highlighted in bold.

We also note that SFA-Net achieves a higher F1 score of 93.50%, exceeding HyperNet by 0.21%. Therefore, HyperNet does not achieve the best result on every evaluation metric. Nevertheless, it achieves the highest reported mIoU and OA among the compared methods while maintaining a competitive F1 score.

Class-wise error analysis. To further investigate the remaining segmentation errors, we compare the per-class performance of HyperNet with the FT-UNetFormer baseline on the Vaihingen dataset, as shown in Table 5. Among all categories, Low Vegetation exhibits the lowest performance, with an IoU of 73.19% and an F1 score of 84.52%, followed by Tree with an IoU of 81.94% and an F1 score of 90.07%. Compared with the baseline, the IoU of Low Vegetation and Tree decreases by 1.81% and 1.21%, respectively, indicating that vegetation-related categories remain the main source of segmentation errors.

Table 5. Per-class IoU and F1 score comparison on the Vaihingen dataset. (Imp. Surf = Impervious Surfaces, Low Veg = Low Vegetation).

Method	Imp. Surf		Building		Low Veg		Tree		Car	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
FT-UNetFormer	87.79	93.50	92.31	96.00	75.00	85.60	83.15	90.80	82.77	90.40
HyperNet (Ours)	94.23	97.03	92.48	96.09	73.19	84.52	81.94	90.07	83.18	90.82

Note: The best results are highlighted in bold.

In contrast, HyperNet improves the IoU of Impervious Surfaces, Buildings, and Cars by 6.44%, 0.17%, and 0.41%, respectively. In particular, the substantial improvement on Impervious Surfaces indicates that the proposed heterogeneous spatial relationship modeling is effective for large-scale structured regions. The relatively lower performance on Low Vegetation and Tree indicates that distinguishing different vegetation categories remains a limitation of the current model.

5.4. Ablation Experiments

We perform ablation experiments to evaluate the contributions of different components in HyperNet and test the effects of various hyperparameter settings on the final segmentation performance.

Effectiveness of key components. We conduct ablation studies to evaluate the individual contributions of HFSM and MCM, as shown in Table 6. Compared with the baseline, introducing HFSM improves the mIoU by 0.71%, while MCM yields an improvement of 0.52%. Although the individual gains are moderate, both modules consistently improve the segmentation performance over the strong baseline. When HFSM and MCM are jointly employed, the mIoU improvement reaches 0.91%. More specifically, adding MCM on top of HFSM further improves the F1, mIoU, and OA by 0.07%, 0.20%, and 0.03%, respectively. Given the relatively small magnitude of these marginal differences, we interpret them as numerical improvements rather than evidence of statistically significant gains. HFSM provides the larger individual gain by introducing higher-order heterogeneous spatial relationship modeling, whereas MCM provides a complementary benefit through robust multi-scale feature interaction.

Table 6. Effectiveness of different modules in HyperNet on the Vaihingen dataset. Results are reported as mean $\pm$ standard deviation over three independent runs.

Baseline	HFSM	MCM	F1 (%)	mIoU (%)	Δ mIoU (%)	OA (%)
✓			91.32 $\pm$ 0.18	84.10 $\pm$ 0.22	–	91.62 $\pm$ 0.15
✓	✓		91.63 $\pm$ 0.12	84.81 $\pm$ 0.24	+0.71	93.47 $\pm$ 0.10
✓		✓	91.52 $\pm$ 0.20	84.62 $\pm$ 0.16	+0.52	92.31 $\pm$ 0.17
✓	✓	✓	91.70 $\pm$ 0.18	85.01 $\pm$ 0.11	+0.91	93.50 $\pm$ 0.06

Note: ✓ indicates that the corresponding component is included. The best results are highlighted in bold.

To further evaluate the generalizability of the two modules, we repeat the component ablation study on the LoveDA dataset. As shown in Table 7, HFSM and MCM individually improve the mIoU by 0.48% and 0.34%, respectively, while their combination yields an overall improvement of 0.86%. The consistent improvement trends on both Vaihingen and LoveDA indicate that HFSM and MCM contribute from different perspectives and that their effectiveness is not limited to a single dataset.

Table 7. Ablation results of HFSM and MCM on the LoveDA dataset. Results are reported as mean $\pm$ standard deviation over three independent runs.

Baseline	HFSM	MCM	mIoU (%)	Δ mIoU (%)
✓			53.84 ± 0.29	–
✓	✓		54.32 ± 0.34	+0.48
✓		✓	54.18 ± 0.16	+0.34
✓	✓	✓	54.70 ± 0.12	+0.86

Note: ✓ indicates that the corresponding component is included. The best results are highlighted in bold.

Impact of the ratio of perturbation strategy. The experimental results in Figure 3 indicate that a perturbation ratio of 0.1 yields the optimal segmentation performance on the Vaihingen dataset. This phenomenon can be explained by the balance between feature diversity augmentation and semantic consistency preservation in remote sensing (RS) image segmentation. RS images are characterized by complex spatial layouts and fine-grained semantic details (e.g., small-scale objects like vehicles adjacent to large buildings). A perturbation ratio that is too low fails to introduce sufficient diversity in scale variations. In such cases, the model is exposed to limited perturbations during training, leading to overfitting to the original scale distribution of the dataset. Consequently, it struggles to generalize to unseen scale variations in test data (e.g., objects captured at slightly different resolutions), resulting in degraded robustness. Conversely, a perturbation ratio that is too high disrupts the inherent semantic consistency of RS features. For instance, excessive perturbations may distort the positional relationships between correlated objects (e.g., roads connecting buildings) or blur critical boundary information (e.g., the edge between vegetation and bare soil). This destruction of semantic integrity misleads the model during training, as it learns noisy or irrelevant scale-dependent patterns, ultimately reducing segmentation accuracy. Thus, a ratio of 0.1 strikes a critical balance. It introduces moderate scale perturbations to simulate real-world variations while preserving the core semantic structure of RS scenes. This enables the model to learn both robust scale-invariant features and accurate spatial relationships.

Impact of the number of node clusters. Figure 4 evaluates the effect of the number of node clusters b on the Vaihingen dataset. Each configuration was independently repeated three times, and the results are reported as mean $\pm$ standard deviation. Among the evaluated settings, $b = 5$ achieves the highest mean performance and is therefore adopted in the remaining experiments.

The results suggest that the number of clusters affects the granularity of hypergraph construction. When the number of clusters is relatively small, semantically different feature nodes may be grouped into coarse hyperedges, which can weaken the discrimination of heterogeneous spatial relationships. In contrast, increasing the number of clusters produces finer partitions and may fragment semantically related nodes into different hyperedges. Therefore, an intermediate number of clusters can provide a better balance between coarse relational grouping and excessive hyperedge fragmentation.

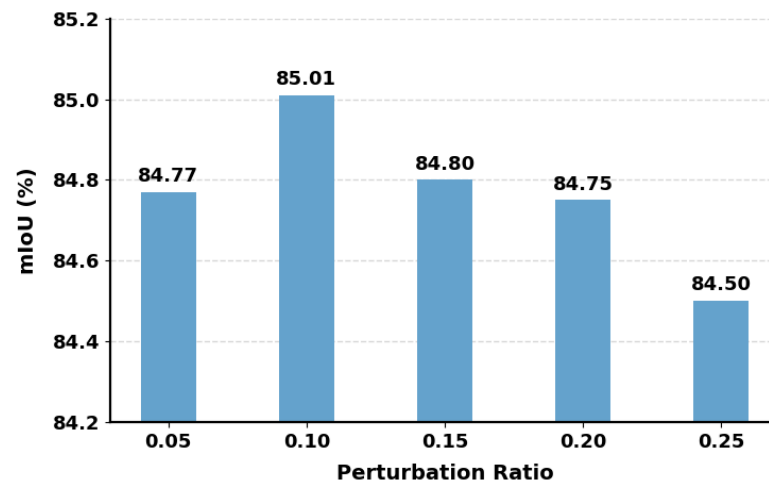


Figure 3. Evaluation of different ratios of the perturbation strategy on the Vaihingen dataset.

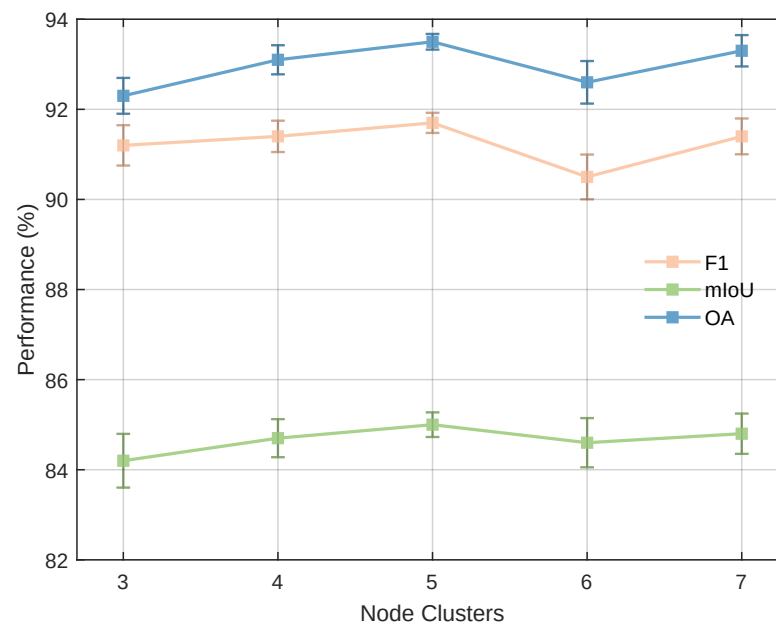


Figure 4. Evaluation of different clusters on the Vaihingen dataset.

It is also worth noting that the performance does not vary monotonically with b . In particular, a noticeable decrease is observed at $b = 6$, followed by a partial recovery at $b = 7$. Changing b causes a discrete reorganization of fuzzy cluster memberships rather than a continuous refinement of the previous partition. Consequently, different values of b can produce different hyperedge compositions and higher-order information propagation patterns. This non-monotonic behavior indicates that the model exhibits some sensitivity to the cluster configuration. Nevertheless, among the evaluated discrete settings, $b = 5$ consistently provides the highest mean segmentation performance and is selected as the default configuration.

Impact of different clustering algorithms. Table 8 compares different strategies for constructing the hypergraph on the Vaihingen dataset. Among them, FCM achieves the best overall performance, with F1, mIoU, and OA scores of 91.70%, 85.01%, and 93.50%, respectively.

Table 8. Comparison of different clustering algorithms on the Vaihingen dataset.

Clustering Algorithm	F1 (%)	mIoU (%)	OA (%)
K-Nearest Neighbors	91.32	84.52	93.18
Spectral Clustering	91.38	84.61	93.32
K-Means Clustering	91.53	84.63	92.62
Fuzzy C-Means	91.70	85.01	93.50

Note: The best results are highlighted in bold.

Different grouping strategies characterize feature relationships in different ways. K-Nearest Neighbors constructs local associations according to feature similarity, which is effective for capturing neighborhood relationships but may provide limited flexibility for modeling nodes with ambiguous associations. Spectral Clustering exploits the similarity structure of the feature space and can capture non-linear grouping patterns, but ultimately produces discrete cluster assignments. K-Means partitions feature nodes according to their distances to cluster centers, providing a simple and efficient grouping strategy, but each node is assigned to only one cluster.

In contrast, FCM assigns each feature node a continuous membership degree to multiple clusters. Such soft assignments naturally correspond to the association strengths between nodes and hyperedges, allowing one node to participate in multiple relational groups with different weights. This property is particularly suitable for heterogeneous remote sensing feature representations, where the associations among spatial regions may not always be strictly exclusive. Therefore, FCM provides more flexible node–hyperedge associations and achieves the best overall performance among the evaluated strategies. It is also worth noting that OA does not always follow the same trend as mean F1 and mIoU. For example, K-Means achieves the second-best F1 and mIoU scores of 91.53% and 84.63%, respectively, but obtains a lower OA of 92.62%. This difference results from the distinct weighting characteristics of these metrics. OA is calculated over all pixels and is therefore more sensitive to performance changes in classes occupying larger pixel proportions, whereas mean F1 and mIoU average class-wise performance and assign equal importance to individual semantic classes. Consequently, different distributions of class-wise errors can lead to different rankings between OA and the class-averaged metrics.

Impact of the loss weighting coefficient. We further evaluate the influence of the weighting coefficient λ in the hybrid loss function. As shown in Table 9, the model achieves the best performance when $\lambda = 1.5$, with an mIoU of 85.01%. Increasing λ from 0.5 to 1.5 gradually improves the segmentation performance, indicating that an appropriate contribution of Dice loss benefits region-level prediction. However, further increasing λ leads to a slight performance decrease. This suggests that an excessively large Dice weight may weaken the balance between pixel-wise classification and region-level supervision. Therefore, $\lambda = 1.5$ is adopted in all experiments.

Computational complexity analysis. To quantify the computational overhead introduced by the proposed modules, we compare the baseline, HFSM-only, MCM-only, and full HyperNet in terms of parameter count, FLOPs, inference latency, peak inference memory, and mIoU, as shown in Table 10. FLOPs, inference latency, and peak inference memory are measured using 512×512 input images under the same inference setting.

Compared with the baseline, the full HyperNet increases the parameter count from 95.98 M to 107.47 M and the FLOPs from 507.19 G to 559.66 G, corresponding to increases of approximately 12.0% and 10.3%, respectively. The inference latency increases from 137.06 ms to 181.20 ms per image, while the peak inference memory increases from 0.88 GB to 1.29 GB. In return, the mIoU improves from 84.10% to 85.01%.

Table 9. Sensitivity analysis of the loss weighting coefficient λ on the Vaihingen dataset.

λ	F1 (%)	mIoU (%)	OA (%)
0.5	91.48	84.67	93.29
1.0	91.60	84.86	93.41
1.5	91.70	85.01	93.50
2.0	91.66	84.93	93.46
2.5	91.53	84.74	93.31

Note: The best results are highlighted in bold.

Table 10. Computational complexity and segmentation performance of different configurations on the Vaihingen dataset.

Method	Params (M)	FLOPs (G)	Inference Latency (ms)	Peak Inference Memory (GB)	mIoU (%)
Baseline	95.98	507.19	137.06	0.88	84.10
Baseline+HFSM	107.42	558.73	177.84	1.24	84.81
Baseline+MCM	96.03	508.12	140.12	0.92	84.62
HyperNet	107.47	559.66	181.20	1.29	85.01

The module-wise comparison further shows that HFSM accounts for most of the additional computational cost due to fuzzy clustering, hypergraph construction, and node–hyperedge message propagation. In contrast, MCM introduces relatively limited overhead, increasing the parameter count from 95.98 M to 96.03 M and the FLOPs from 507.19 G to 508.12 G, while improving the mIoU from 84.10% to 84.62%. These results provide a quantitative characterization of the accuracy–efficiency trade-off of HyperNet.

Visualization. We further present qualitative comparisons of our method across the three employed datasets in Figures 5–7.

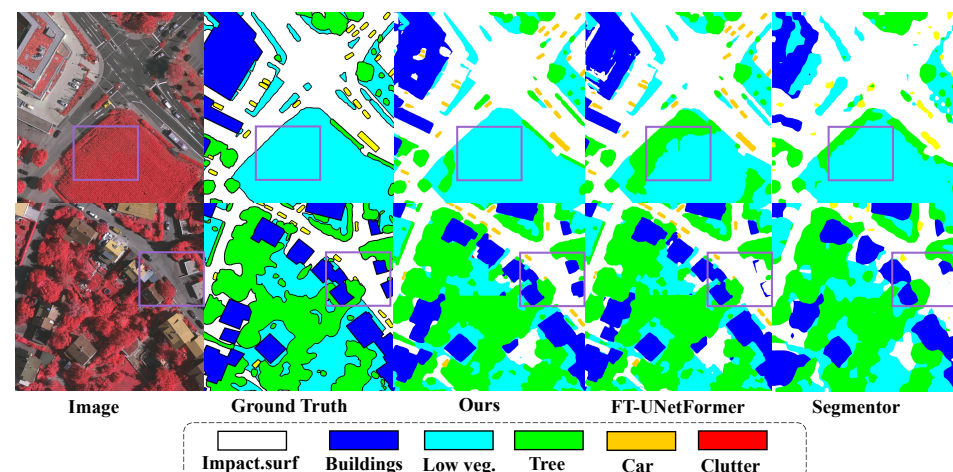


Figure 5. Visualization results of IDs 10 and 14 from the Vaihingen test set. The first column denotes the input NIR–R–G TOP images, the second column shows the ground truth, and the remaining columns show the segmentation results of different methods. The purple boxes highlight representative regions for qualitative comparison.

By integrating the hypergraph-perception module we developed, the model effectively augments the feature embedding with rich, multi-faceted spatial relationships among entities, thereby strengthening inter-entity coherence. Compared to the baseline FT-UNetFormer [8], our newly designed multi-scale modeling component perturbs features at each scale in a targeted manner. This encourages the network to distinguish between disparate resolutions and produce more precise boundaries—even when the scene contains objects of varying sizes (such as buildings and vehicles). The qualitative results show that

HyperNet produces more coherent predictions in scenes containing objects with different scales and spatial layouts. These observations are consistent with the intended roles of MCM in improving cross-scale feature interaction and HFSM in enhancing higher-order contextual information propagation.

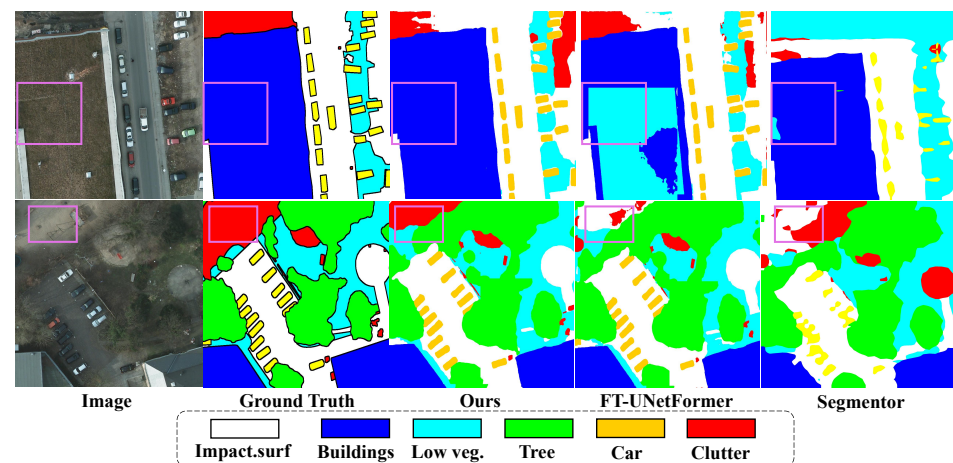


Figure 6. Visualization results of ID 4_13 and 6_15 from the Potsdam test set. The first column denotes the input RGB images. The third column shows the segmentation results of the proposed HyperNet. The purple boxes highlight representative regions for qualitative comparison.

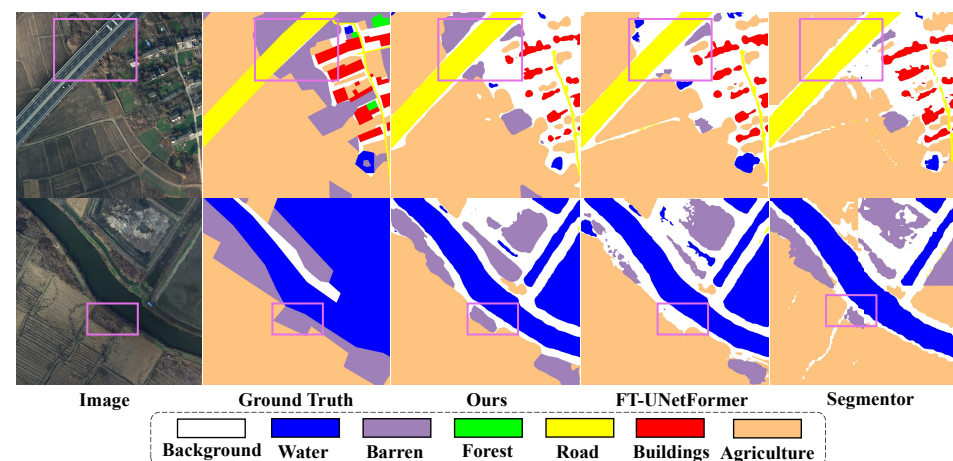


Figure 7. Visualization results of ID 2558 and 2585 from the LoveDA test set. The first column denotes the input RGB images. The third column shows the segmentation results of the proposed HyperNet. The purple boxes highlight representative regions for qualitative comparison.

6. Discussion

In addition to the quantitative comparisons, the class-wise and failure-case analyses provide further insights into the behavior of HyperNet. As shown in Figure 8, segmentation errors mainly occur in low-quality regions with blurred boundaries and in scenes containing small or densely distributed objects. These observations are consistent with the class-wise results, where vegetation-related categories remain relatively challenging. Together with the quantitative experiments, these results reveal both the strengths and remaining limitations of the proposed higher-order relational modeling strategy.

The experimental results across LoveDA, Vaihingen, and Potsdam demonstrate that HyperNet achieves competitive segmentation performance across different remote sensing scenes. The ablation experiments further show that both HFSM and MCM contribute to the performance improvement over the baseline. HFSM provides the larger individual improvement, while MCM offers a complementary contribution when combined with HFSM.

Nevertheless, the magnitude of the improvement varies across datasets and evaluation metrics. In particular, on the Potsdam dataset, SFA-Net achieves a higher F1 score of 93.50% than the $93.29 \pm 0.05\%$ obtained by HyperNet, whereas HyperNet achieves the highest reported mIoU and OA among the compared methods. Therefore, the experimental results support the effectiveness of the proposed modules, but do not indicate uniform superiority over existing methods across all evaluation criteria.

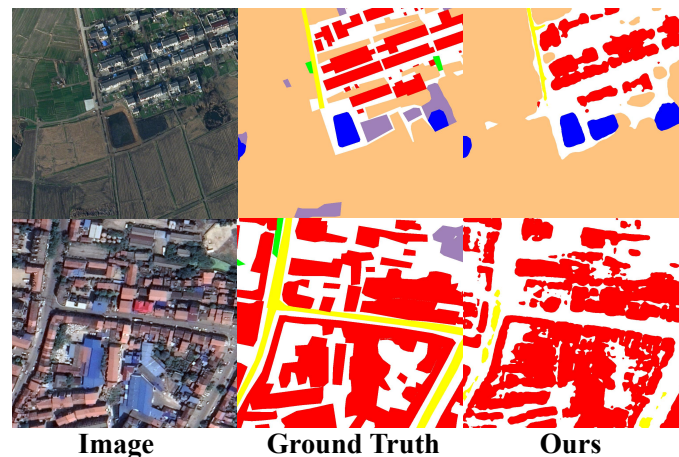


Figure 8. Representative failure cases on the LoveDA dataset. The colors correspond to the semantic classes and follow the same color scheme as in Figure 7.

The experiments also reveal that different evaluation metrics may respond differently to the same model configuration. In the clustering comparison, K-Means achieves competitive mean F1 and mIoU but a relatively lower OA. A similar behavior is observed for the MCM-only configuration, which obtains competitive mean F1 and mIoU while showing a lower OA than the HFSM-only configuration. This discrepancy is consistent with the different weighting mechanisms of these metrics. OA is calculated over all pixels and is therefore more strongly influenced by classes occupying larger pixel proportions, whereas mean F1 and mIoU average class-wise performance and assign equal importance to individual semantic classes. The class-wise analysis on the Vaihingen dataset further shows that the performance changes are not uniform across categories. HyperNet improves the segmentation of Impervious Surfaces, Buildings, and Cars, while lower performance is observed for Low Vegetation and Trees relative to the baseline. These results indicate that the performance of a segmentation model should be assessed jointly using multiple metrics and class-wise behavior rather than relying on a single aggregate measure.

The node-cluster sensitivity analysis provides another important observation. The segmentation performance does not vary monotonically with the number of clusters. Although $b = 5$ provides the best overall mean performance among the evaluated settings, a noticeable decrease is observed at $b = 6$, followed by a partial recovery at $b = 7$. Changing b does not continuously refine a fixed partition; instead, it causes a discrete reorganization of fuzzy memberships and consequently changes the composition of the constructed hyperedges. Different cluster configurations can therefore lead to different higher-order information propagation patterns. This behavior indicates that the current hypergraph construction retains some sensitivity to the cluster configuration and that the selected value of $b = 5$ should be regarded as the best-performing configuration among the evaluated settings rather than a universally optimal value.

The computational analysis further clarifies the accuracy–efficiency trade-off introduced by the proposed modules. Compared with the baseline, the full HyperNet increases the parameter count from 95.98 M to 107.47 M and the FLOPs from 507.19 G to 559.66 G, corresponding to increases of approximately 12.0% and 10.3%, respectively. The inference

latency increases from 137.06 ms to 181.20 ms per image, while the peak inference memory increases from 0.88 GB to 1.29 GB. Meanwhile, the mIoU improves from 84.10% to 85.01%. The module-wise comparison shows that HFSM accounts for most of the additional computational cost because it involves fuzzy clustering, hypergraph construction, and node–hyperedge message propagation, whereas MCM introduces relatively limited additional overhead. These results indicate that the improved segmentation performance is accompanied by additional computational and memory costs, which should be considered in practical deployment.

Limitations. Several limitations remain in the current framework. First, although HyperNet consistently improves the baseline in the conducted ablation experiments, the numerical margins over some existing methods are relatively small, and the proposed method does not achieve the best result on every metric, as illustrated by the higher F1 score of SFA-Net on Potsdam. Second, the non-monotonic cluster-number analysis indicates that hypergraph construction is sensitive to the cluster configuration, suggesting that a more adaptive strategy for determining the hypergraph structure would be desirable. Third, HFSM introduces additional computational and memory overhead, which may limit its applicability in resource-constrained scenarios. Finally, the class-wise and failure-case analyses show that vegetation-related categories and small or fragmented objects remain challenging for the current model. Future work will therefore focus on lightweight and adaptive hypergraph construction, reducing sensitivity to clustering hyperparameters, and improving the representation of fine-grained and ambiguous semantic regions.

7. Conclusions

This paper proposes HyperNet, a hypergraph-driven framework for remote sensing semantic segmentation that models higher-order heterogeneous spatial relationships among geographically distributed regions. HFSM constructs adaptive hyperedges through fuzzy clustering to learn higher-order relational information, while MCM enhances robust multi-scale feature interaction. Experiments on the LoveDA, Vaihingen, and Potsdam datasets demonstrate state-of-the-art mIoU performance, with mIoU scores of 54.70%, 85.01%, and 87.64%, respectively. Ablation studies further confirm the individual and complementary contributions of HFSM and MCM.

Several limitations remain. The current experiments do not separately quantify specific topological or functional relationships. In addition, HFSM introduces additional computational overhead, and the constructed hypergraph depends on the clustering strategy and associated hyperparameters. The performance on very small and fragmented objects also requires further investigation. Future work will focus on lightweight and robust hypergraph modeling, object-scale-aware representation learning, and more direct relation-level analysis.

Author Contributions: Conceptualization, F.H. and X.X.; methodology, Q.Z.; software, Q.Z.; validation, Q.Z.; formal analysis, F.H.; investigation, Q.Z.; resources, X.X.; writing—original draft preparation, Q.Z. and L.P.; writing—review & editing, Q.Z. and X.X.; visualization, L.P.; supervision, L.P. and X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key R&D Program of Shandong Province, China (No. 2025CXGC020307) and the National Natural Science Foundation of China (No. 62506213).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: In the process of manuscript preparation, the authors utilized Qwen 3.7 for assistance in linguistic refinement, grammatical correction, and enhancing overall clarity and fluency. The authors have thoroughly reviewed and revised the AI-generated content and assume complete responsibility for the publication's content.

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

References

1. Liu, P. A multimodal fusion framework for semantic segmentation of remote sensing based on multilevel feature fusion learning. *Neurocomputing* **2025**, *653*, 131233. [\[CrossRef\]](#)
2. Zeng, W.; Cheng, M.; Yuan, Z.; Dai, W.; Wu, Y.; Liu, W.; Wang, C. Domain adaptive remote sensing image semantic segmentation with prototype guidance. *Neurocomputing* **2024**, *580*, 127484. [\[CrossRef\]](#)
3. Qin, Z.; Yu, D.; Luo, C.; Chen, Z. Sliced Wasserstein Bridge for Open-Vocabulary Video Instance Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 12470–12478.
4. Chu, Y.; Ye, M.; Qian, Y. Fine-Grained Image Recognition Methods and Their Applications in Remote Sensing Images: A Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 19640–19667. [\[CrossRef\]](#)
5. Li, J.; Cai, Y.; Li, Q.; Kou, M.; Zhang, T. A review of remote sensing image segmentation by deep learning methods. *Int. J. Digit. Earth* **2024**, *17*, 2328827. [\[CrossRef\]](#)
6. Alam, M.; Wang, J.F.; Guangpei, C.; Yunrong, L.; Chen, Y. Convolutional Neural Network for the Semantic Segmentation of Remote Sensing Images. *Mob. Netw. Appl.* **2021**, *26*, 200–215. [\[CrossRef\]](#)
7. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [\[CrossRef\]](#)
8. 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. [\[CrossRef\]](#)
9. Cai, D.; Song, M.; Sun, C.; Zhang, B.; Hong, S.; Li, H. Hypergraph Structure Learning for Hypergraph Neural Networks. In Proceedings of the IJCAI, Vienna, Austria, 23–29 July 2022; pp. 1923–1929.
10. Ouyang, S.; Li, Y. Combining Deep Semantic Segmentation Network and Graph Convolutional Neural Network for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.* **2020**, *13*, 119. [\[CrossRef\]](#)
11. Qin, Z.; Yu, D.; Shi, Y.; Wang, Q.; Chen, Z. Video Instance Segmentation by Weighted Structure Inference. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025; pp. 7616–7624.
12. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
14. Yang, Z.; Zhou, D.; Yang, Y.; Zhang, J.; Chen, Z. Road extraction from satellite imagery by road context and full-stage feature. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 8000405. [\[CrossRef\]](#)
15. Liu, M.; Dan, J.; Lu, Z.; Yu, Y.; Li, Y.; Li, X. CM-UNet: Hybrid CNN-Mamba UNet for Remote Sensing Image Semantic Segmentation. *arXiv* **2024**, arXiv:2405.10530.
16. John, D.; Zhang, C. An attention-based U-Net for detecting deforestation within satellite sensor imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102685. [\[CrossRef\]](#)
17. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
18. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [\[CrossRef\]](#)
19. Wang, Y.; Yang, L.; Liu, X.; Yan, P. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. *Sci. Rep.* **2024**, *14*, 9716. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Ma, X.; Lian, R.; Wu, Z.; Guo, H.; Yang, F.; Ma, M.; Wu, S.; Du, Z.; Zhang, W.; Song, S. LOGCAN++: Adaptive Local-Global Class-Aware Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4404216. [\[CrossRef\]](#)
21. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.

23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
24. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany; pp. 213–229.
25. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 7262–7272.
26. Li, H.; Li, L.; Zhao, L.; Liu, F. ResU-Former: Advancing Remote Sensing Image Segmentation with Swin Residual Transformer for Precise Global–Local Feature Recognition and Visual–Semantic Space Learning. *Electronics* **2024**, *13*, 436. [\[CrossRef\]](#)
27. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. HRFormer: High-Resolution Transformer for Dense Prediction. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021.
28. Fan, L.; Zhou, Y.; Liu, H.; Li, Y.; Cao, D. Combining Swin Transformer with UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5530111. [\[CrossRef\]](#)
29. Cui, W.; Yao, M.; Hao, Y.; Wang, Z.; He, X.; Wu, W.; Li, J.; Zhao, H.; Xia, C.; Wang, J. Knowledge and geo-object based graph convolutional network for remote sensing semantic segmentation. *Sensors* **2021**, *21*, 3848. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Kang, X.; Hong, Y.; Duan, P.; Li, S. Fusion of hierarchical class graphs for remote sensing semantic segmentation. *Inf. Fusion* **2024**, *109*, 102409. [\[CrossRef\]](#)
31. Zhao, Y.; Qiu, L.; Yang, Z.; Chen, Y.; Zhang, Y. MGF-GCN: Multimodal interaction Mamba-aided graph convolutional fusion network for semantic segmentation of remote sensing images. *Inf. Fusion* **2025**, *122*, 103150. [\[CrossRef\]](#)
32. Wu, J.; Fu, R.; Liu, Q.; Ni, W.; Cheng, K.; Li, B.; Sun, Y. A dual neighborhood hypergraph neural network for change detection in VHR remote sensing images. *Remote Sens.* **2023**, *15*, 694. [\[CrossRef\]](#)
33. Jing, W.; Zhang, W.; Di, D.; Li, C.; Emam, M.; Mian, A. Hypergraph biformer for semantic segmentation of high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [\[CrossRef\]](#)
34. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In *Neural Information Processing Systems Track on Datasets and Benchmarks*; Curran Associates, Inc.: Red Hook, NY, USA, 2021.
35. Li, Y.; Li, X.; Dai, Y.; Hou, Q.; Liu, L.; Liu, Y.; Cheng, M.M.; Yang, J. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *Int. J. Comput. Vis.* **2024**, *133*, 1410–1431. [\[CrossRef\]](#)
36. Chen, Y.; Fang, P.; Zhong, X.; Yu, J.; Zhang, X.; Li, T. Hi-ResNet: Edge Detail Enhancement for High-Resolution Remote Sensing Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 15024–15040. [\[CrossRef\]](#)
37. Lu, W.; Chen, S.B.; Ding, C.H.Q.; Tang, J.; Luo, B. LWGANet: A Lightweight Group Attention Backbone for Remote Sensing Visual Tasks. *arXiv* **2025**, arXiv:2501.10040.
38. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5607315. [\[CrossRef\]](#)
39. Cha, K.; Seo, J.; Lee, T. A Billion-scale Foundation Model for Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *19*, 16363–16378. [\[CrossRef\]](#)
40. Hanyu, T.; Yamazaki, K.; Tran, M.; McCann, R.A.; Liao, H.; Rainwater, C.; Adkins, M.; Cothren, J.; Le, N. AerialFormer: Multi-Resolution Transformer for Aerial Image Segmentation. *Remote Sens.* **2024**, *16*, 2930. [\[CrossRef\]](#)
41. Wang, F.; Wang, H.; Wang, D.; Guo, Z.; Zhong, Z.; Lan, L.; Yang, W.; Zhang, J. Harnessing Massive Satellite Imagery with Efficient Masked Image Modeling. *arXiv* **2024**, arXiv:2406.11933.
42. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing remote sensing foundation model via multitask pretraining. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11632–11654. [\[CrossRef\]](#)
43. Lu, W.; Chen, S.B.; Shu, Q.L.; Tang, J.; Luo, B. DecoupleNet: A Lightweight Backbone Network with Efficient Feature Decoupling for Remote Sensing Visual Tasks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4414613. [\[CrossRef\]](#)
44. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [\[CrossRef\]](#)
45. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [\[CrossRef\]](#)
46. Wu, H.; Huang, P.; Zhang, M.; Tang, W.; Yu, X. CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 2004612. [\[CrossRef\]](#)
47. Kim, B.J.; Kim, S.W. Stochastic Subsampling with Average Pooling. *arXiv* **2024**, arXiv:2409.16630.
48. Zhang, X.; Weng, Z.; Zhu, P.; Han, X.; Zhu, J.; Jiao, L. ESDINet: Efficient Shallow-Deep Interaction Network for Semantic Segmentation of High-Resolution Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5607615. [\[CrossRef\]](#)

49. Li, B.; Zhang, Y.; Zhang, Y.; Li, B.; Li, Z. Dual-Path Feature Fusion Network for Semantic Segmentation of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2503205. [[CrossRef](#)]
50. 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. [[CrossRef](#)]
51. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [[CrossRef](#)]
52. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [[CrossRef](#)]
53. Hwang, G.; Jeong, J.; Lee, S.J. SFA-Net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sens.* **2024**, *16*, 3278. [[CrossRef](#)]
54. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [[CrossRef](#)]
55. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 1290–1299.
56. Wang, Q.; Luo, X.; Feng, J.; Zhang, G.; Jia, X.; Yin, J. Multiscale Prototype Contrast Network for High-Resolution Aerial Imagery Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615114. [[CrossRef](#)]
57. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. RS-Mamba for Large Remote Sensing Image Dense Prediction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5633314. [[CrossRef](#)]
58. 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. *arXiv* **2024**, arXiv:2406.10828.
59. Zhu, Q.; Cai, Y.; Fang, Y.; Yang, Y.; Chen, C.; Fan, L.; Nguyen, A. Samba: Semantic segmentation of remotely sensed images with state space model. *Heliyon* **2024**, *10*, e38495. [[CrossRef](#)] [[PubMed](#)]
60. Wang, J.; Li, J.; Fan, G.; Ju, Y.; Fang, X.; Kot, A.C. Prototype-Driven Structure Synergy Network for Remote Sensing Images Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4708216. [[CrossRef](#)]

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.