remote sensing

MDPI

Article

MCPF-Net: Multi-Stage LiDAR-Image Collaborative Perception Fusion Network for Point Cloud Semantic Segmentation of Urban Scenes

Huchen Li [1], Wubiao Huang [1], Xiangda Lei [2], Bin Liu [3], Haibing Liu [1], Shihan Chen [1] and Fei Deng [1,4,*]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China; lihuchen@whu.edu.cn (H.L.); huangwubiao@whu.edu.cn (W.H.); liuhb_whu@whu.edu.cn (H.L.); csh1999@whu.edu.cn (S.C.)
[2] School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China; leixd@hpu.edu.cn
[3] Hinton STAI Institute, East China Normal University, Minhang, Shanghai 200241, China; 52290161014@stu.ecnu.edu.cn
[4] Hubei Luojia Laboratory, Wuhan 430079, China
* Correspondence: fdeng@sgg.whu.edu.cn

Highlights

What are the main findings?

- A multi-stage LiDAR-image collaborative perception fusion network (MCPFNet) was proposed for urban point cloud semantic segmentation.
- The proposed geometric-aware, semantic-aware, and adaptive fusion modules effectively improved multimodal feature complementarity and achieved superior segmentation performance on three benchmark datasets.

What are the implications of the main findings?

- The proposed framework improves cross-modal semantic and geometric representation learning, providing a reliable solution for multimodal urban scene understanding.
- The multi-stage bidirectional fusion strategy demonstrates strong generalization ability across different urban remote sensing datasets and complex scenes.

Abstract

Multimodal fusion unlocks the potential of point cloud semantic segmentation, thereby driving advancements in surface observation and visual perception tasks. Although light detection and ranging (LiDAR) systems capture precise 3D structural geometry and optical images provide rich semantic and textural information, existing fusion methods struggle with limited cross-modal perception and insufficient information complementarity. To address these limitations, we propose a multi-stage LiDAR-image collaborative perception fusion network (MCPFNet) for point cloud semantic segmentation of urban scenes. At the middle fusion stage, the network incorporates an elevation-guided geometric-aware fusion module and a semantic-aware cross-attention fusion module to enable bidirectional feature injection between LiDAR and image modalities. In the late fusion stage, a bidirectional adaptive fusion module further refines semantic representations through gated weighting and bidirectional cross-attention mechanisms. Extensive experiments on three multimodal datasets with different resolutions, i.e., ISPRS Vaihingen, N3C-California, and UAVScenes, demonstrate that MCPFNet outperforms existing fusion methods, achieving mIoUs of 74.51%, 95.15%, and 62.76%, respectively. Hence, our multi-stage fusion and bidirectional interaction strategy is more reliable and accurate than existing methods in performing segmentation across diverse and complex urban scenes.

Check for updates

Academic Editor: Sheng Nie

Received: 18 May 2026
Revised: 24 June 2026
Accepted: 2 July 2026
Published: 6 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: point cloud semantic segmentation; multimodal fusion; multi-stage fusion; geometric-aware fusion; semantic-aware fusion

1. Introduction

Multimodal data fusion is widely employed for semantic segmentation in various applications, including land cover mapping [1], smart agriculture [2], low-altitude intelligent transportation [3], autonomous driving [4,5], and urban planning [6–9]. Light detection and ranging (LiDAR) systems capture precise spatial and geometric information, while optical images provide dense textural patterns and rich semantic content. Although the complementary nature of these modalities facilitates a holistic scene interpretation, substantial differences in sensor properties, data structures, and semantic representations present considerable challenges to effective cross-modal fusion [10,11]. Therefore, the efficient integration of semantic and geometric information is essential for advancing semantic segmentation of complex urban scenes.

Depending on the fusion stage, multimodal fusion methods can be broadly categorized into three paradigms: early-, middle-, and late-stage fusion [12,13]. Early-stage fusion directly combines raw or preprocessed data from multiple sensors, thereby utilizing original information at the input level [14,15]. These fused features are subsequently processed using 3D network backbones such as PointNet [16] and RandLA-Net [17]. However, the inherent heterogeneity between different modalities often introduces interference, making it difficult to fully capture mode-specific characteristics. Recent studies have demonstrated that middle-stage or late-stage fusion strategies are generally more robust and effective in multimodal segmentation tasks [13].

Middle-stage fusion facilitates joint learning between image and LiDAR data through either a 2D-to-3D or 3D-to-2D strategy [18]. The 3D-to-2D approach projects point clouds onto 2D feature maps and fuses them with image features across multiple scales [19]. However, this projection process inevitably distorts the spatial topology and local geometric relationships of the original point cloud. Furthermore, owing to the inherent sparsity of point clouds, interpolation or filling operations often introduce geometric inaccuracies in the resulting dense projection maps. By contrast, the 2D-to-3D strategy transfers image information into the point cloud feature space using techniques such as knowledge distillation and specialized fusion modules. For example, 2DPASS [20] integrated multi-scale fusion within a unified distillation framework, whereas CSFNet [21] introduced a semantic-focused module for adaptive multi-scale fusion. Further, late-stage fusion combines modalities at the decision level after each branch has already been trained independently. Hang et al. [22] employed a weighted-sum strategy based on the classification accuracy of the output of each modality. EGAFNet [23] further advanced this paradigm using an adaptive fusion module that dynamically fuses 2D and 3D features.

Despite multimodal fusion in point cloud semantic segmentation attracting increasing attention, the inherent structural strengths of each modality and the potential for bidirectional feature interaction remain underexplored. Most existing methods [1,20] rely on unidirectional transfer mechanisms, such as projection or distillation, which restrict deeper cross-modal alignment and synergistic representation learning. In addition, although feature characteristics vary considerably across different fusion stages, methods [22,24] largely continue to apply a uniform fusion strategy throughout, failing to account for stage-specific feature heterogeneity and adaptability. A primary challenge lies in effectively balancing geometric and semantic information across stages and dynamically optimizing the fusion strategy.

To fully leverage the complementary semantic and geometric information between LiDAR point clouds and optical images, this study proposes a multi-stage LiDAR-image collaborative perception fusion network (MCPFNet). The model facilitates bidirectional feature interaction across both middle-stage and late-stage fusion processes. At the middle stage, an elevation-guided geometric-aware fusion module (EGFM) and a semantic-aware cross-attention fusion module (SCFM) are designed to explicitly exchange geometric and semantic cues between 3D point clouds and 2D images, thereby enhancing cross-modal consistency and alignment. In the late fusion stage, a bidirectional adaptive fusion module (BAFM) is introduced, which utilizes a gated weighting mechanism to adaptively select high-confidence features from each modality. Additionally, a bidirectional cross-attention mechanism is used for reinforced feature interaction, combined with which the BAFM effectively reduces semantic ambiguity and further improves segmentation accuracy. The main contributions of our study are summarized below.

- A novel model, MCPFNet, is proposed for accurate semantic segmentation of urban scenes, which provides bidirectional and hierarchical cross-modal interaction.
- The EGFM is designed to transfer geometric priors from point clouds to images via perspective projection, markedly enhancing the spatial structural perception of 2D features.
- The SCFM is proposed to propagate rich semantic information from images to point clouds through a multi-scale cross-attention mechanism, achieving effective semantic alignment and complementary enhancement across modalities.
- A BAFM is constructed for the late stage to dynamically balance the contributions of geometric and semantic features, thereby refining cross-modal fusion and improving final segmentation accuracy.

The remainder of this paper is organized as follows. Section 2 reviews point-cloud pseudo-image fusion and point-cloud image fusion methods. Section 3 details the proposed multimodal fusion framework. Section 4 presents experimental results and analysis. Section 5 provides a discussion. Finally, Section 6 concludes the paper.

2. Related Works

2.1. Point-Cloud Pseudo-Image Fusion Methods

To leverage point cloud information from multiple perspectives, various studies have employed point-cloud pseudo-image fusion methods to project 3D point clouds onto 2D planes or spherical surfaces, thereby extracting key semantic features from the resulting 2D views. RPVNet [25] integrates point-, voxel-, and range-view representations and adaptively selects information through a gating mechanism. Building on this, UniSeg [26] introduces a learnable cross-modal correlation module to adaptively fuse voxel-view and range-view features, which are then mapped back to the point space for unified learning. PVI-Net [27] uses a point-voxel cross-attention mechanism combined with a multi-view fusion strategy to enhance cross-modal feature interaction. A2Blend [28] further integrates a cross-modal triplet alignment loss to promote the aggregation of semantically related features. These methods establish view correspondence via indexing. However, LiDAR point cloud sparsity and projection distortion result in fragmented and distorted projected views. Consequently, the blurred boundaries and semantic ambiguity undermine structural fidelity and weaken the stability and continuity of cross-view feature fusion.

Beyond these multi-view fusion strategies, other methods construct pseudo aerial images to leverage large-scale 2D spatial structures. Zou and Li [29] and Wei et al. [30] first generate pseudo aerial images via projection and align multimodal features using a geometry-aware fusion module. These approaches preserve the maximum value at each pixel and produce dense feature maps through max pooling. DAPSS [31] introduces

a digital orthophoto map rendering strategy based on RANSAC plane fitting, which fills occluded regions by searching for the nearest points in spherical space. PCE [32] incorporates intensity and multiple elevation cues into projection maps to enhance context perception, whereas EGAFNet [23] accounts for terrain variation and dynamically fuses multimodal features via an elevation guidance adaptive fusion module.

Although these methods improve segmentation performance by incorporating supplementary 2D representations, they face inherent limitations. The sparsity of point clouds leads to empty grid cells in the projection map, and interpolation-based filling often introduces geometric distortions. Furthermore, point-cloud pseudo-image fusion lacks generalizability for cross-modal segmentation tasks involving real optical images [12].

2.2. Point-Cloud Image Fusion Methods

The integration of multimodal sensors has promoted the development of point-cloud image fusion methods, which leverage complementary characteristics across modalities to address the limitations of single-modality perception. PMF [33,34] incorporates image features into point clouds through perspective projection but compromises 3D structural representation. To mitigate this, 2DPASS [20] reverses the fusion direction by injecting point cloud features into the image branch, thereby preserving 3D spatial awareness. Building on this concept, DFAMNet [35] enhances inter-modal interaction via a dual-fusion attention mechanism, whereas TPV-IGKD [36] projects point clouds onto a tri-plane view to integrate structural information from multiple perspectives. Adopting a different approach, VPANet [37] introduces a 3D spatial feature enhancement module to alleviate boundary blurring caused by spatial information loss in PMF. Despite these advances, these methods predominantly rely on unidirectional information transfer and overlook the potential of bidirectional modal interaction.

To enable a more balanced fusion, CMDFusion [18] introduces a bidirectional fusion mechanism primarily conducted in 3D space, with the image branch serving as an auxiliary for feature distillation. U2MKD [38] integrates a unified bidirectional single-to-multimodal knowledge distillation framework, and MSeg3D [39] incorporates both semantic and geometric feature fusion to mitigate heterogeneity between point clouds and images.

In the context of airborne laser scanning, IKDNet [1] employs global and category knowledge-guided gated models to refine feature maps from weaker modalities. PIFNet [12] incorporates a hierarchical fusion module with an efficient interaction mechanism for adaptive integration between images and point clouds. MVGNet [40] computes normal vectors to derive multi-view weighting scores for integrating multi-perspective images. IPCE-Net [2] utilizes a heterogeneous conversion module to facilitate mutual access and feature enhancement between 2D images and 3D point clouds. In contrast to the above methods, our method performs bidirectional feature interaction at the middle stage and further integrates semantic and geometric modalities at the late stage, thereby achieving more comprehensive cross-modal complementarity.

3. Methodology

3.1. Network Overview

The overall architecture of the proposed MCPFNet is shown in Figure 1. The network adopts a dual-stream structure to extract semantic and geometric features from images and point clouds and performs hierarchical fusion across multiple stages. In the middle fusion stage, the EGFM first projects the 3D geometric features F_P onto a 2D image plane to obtain a geometric projection feature F_{PI}, which is then deeply fused with image features F_I to enhance geometric perception. Subsequently, the SCFM integrates semantic and geometric information across multiple scales, improving cross-modal alignment and feature com-

plementarity. In the late fusion stage, the BAFM further refines the fused representations from both modalities, effectively enhancing the performance of semantic segmentation. Sections 3.2–3.5 describe MCPFNet in detail.

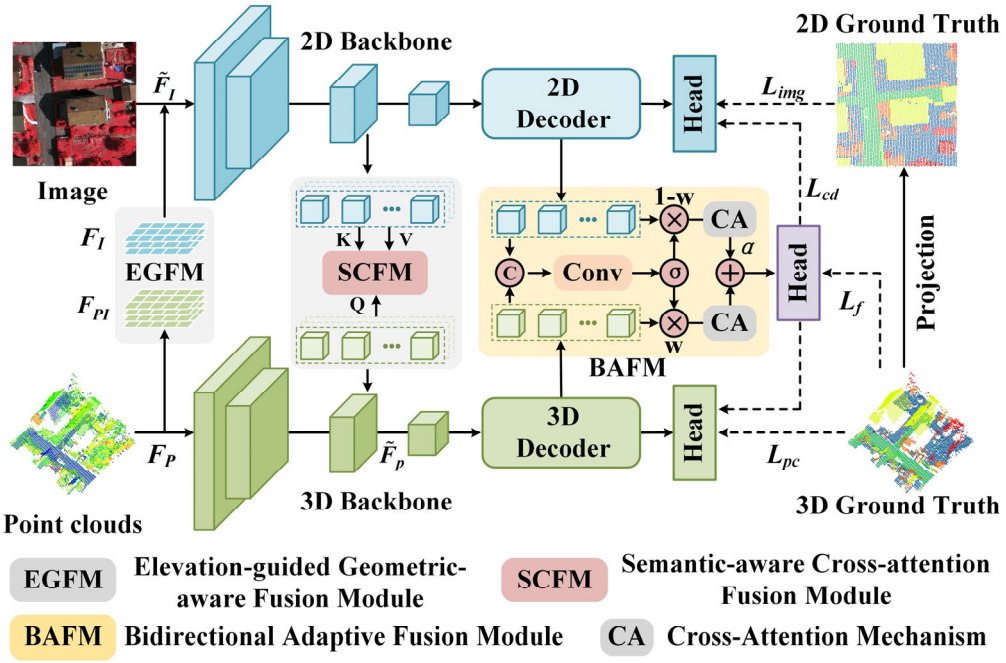


Figure 1. Framework overview of MCPFNet.

3.2. EGFM

To inject accurate 3D geometric information into the image branch, the EGFM projects LiDAR points onto the 2D image plane and fuses them with image features, as shown in Figure 2. Specifically, consider a LiDAR point cloud $P = \{P_i\}_{i=1}^{N} \in \mathbf{R}^{N\times 4}$, where each point $P_i = (x_i, y_i, z_i, r_i)$ consists of Cartesian coordinates and reflection intensity, and N is the total number of points. The corresponding image is denoted as $I \in \mathbf{R}^{H\times W\times 3}$, with H and W representing its height and width, respectively. Due to variations in sensor calibration and acquisition setups, the geometric relationship between LiDAR point clouds and images differs across datasets. When LiDAR and images are captured on the same platform, the standard perspective projection of each P_i onto the image plane is computed as follows:

$$[u_i, v_i, 1]^{\mathrm{T}} = \frac{1}{z_i} \times \mathbf{K} \times \mathbf{T} \times [x_i, y_i, z_i, 1]^{\mathrm{T}}, \tag{1}$$

where $\mathbf{K} \in \mathbf{R}^{3\times 4}$ and $\mathbf{T} \in \mathbf{R}^{4\times 4}$ denote the intrinsic and extrinsic matrices of the camera, respectively. For data collected from different platforms, such as UAV LiDAR and satellite or aerial orthorectified imagery, the projection is computed using the standard affine geospatial mapping as follows:

$$\begin{aligned} u_i &= ax_i + by_i + c \\ v_i &= dx_i + ey_i + f \end{aligned} \tag{2}$$

where (a, b, c, d, e, f) are the geotransform coefficients provided by the orthophoto metadata. The resulting 2D coordinates are given by $P_{i,j}^{2\mathrm{D}} = (\lfloor u_i \rfloor, \lfloor v_j \rfloor) \in \mathbf{R}^2$, with $0 \le \lfloor u_i \rfloor \le H$ and $0 \le \lfloor v_j \rfloor \le W$.

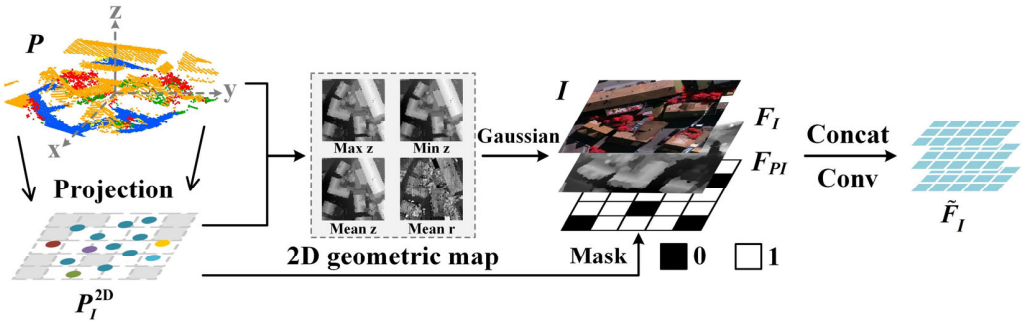


Figure 2. Elevation-guided geometric-aware fusion module.

Subsequently, a 2D grid is constructed at the image resolution, and elevation and intensity information of all projected points within each cell are aggregated. To capture the local geometric structure, the maximum, minimum, and mean elevations $(z_{max}, z_{min}, z_{mean})$ are computed, together with the mean intensity, r_{mean}, of points in each cell to form four geometric channels, as shown below:

$$\begin{cases} z_{max} = \max(z_{p_k}) \\ z_{min} = \min(z_{p_k}) \\ z_{mean} = \frac{1}{m}\sum_{k=1}^{m} z_{p_k} \\ r_{mean} = \frac{1}{m}\sum_{k=1}^{m} r_{p_k} \end{cases}, \tag{3}$$

where z_{p_k} and r_{p_k} denote the elevation and intensity of the k-th point P_k in the cell $P_{i,j}^{2D}$, and m represents the number of points in that cell. The resulting 2D geometric feature map $F_{PI} = [z_{max}, z_{min}, z_{mean}, r_{mean}] \in \mathbf{R}^{4\times H\times W}$ provides diverse geometric cues aligned with the image grid.

However, owing to point cloud sparsity, some grid cells are empty. Inspired by APNet [30] and EGAFNet [23], which fill empty cells using max-pooling or minimum-value strategies, we introduce a mask-guided Gaussian interpolation to preserve geometric accuracy while maintaining boundary sharpness.

$$F_{PI}(u,v) = \begin{cases} F_{PI}(u,v), & \tau = 1 \\ \frac{\sum_{(i,j)\in N(u,v)} G_\sigma(i,j)\left(\theta_{i,j} F_{PI}(u+i,v+j)\right)}{\sum_{(i,j)\in N(u,v)} G_\sigma(i,j)\theta_{i,j}+\epsilon}, & \tau = 0 \end{cases} \tag{4}$$

where $N(u,v)$ denotes the set of neighboring pixels, G_σ is a Gaussian kernel weight with $\sigma = 1.5$, $\theta_{i,j} = mask(u+i, v+j)$, $\tau = mask(u,v)$, and $mask \in \{0,1\}$ represents the occupancy of projected 3D points in each grid cell. Pixels that remain empty after the interpolation are filled with zeroes [2].

Finally, the refined geometric map F_{PI} is concatenated with the image feature $F_I \in \mathbf{R}^{3\times H\times W}$ to form the geometric-aware image features $\tilde{F}_I \in \mathbf{R}^{C_e\times H\times W}$, where C_e denotes the initial embedding channel that is embedded through a convolutional layer, which enriches the image branch with geometric priors from the point cloud. The process is formulated as follows:

$$\tilde{F}_I = \varphi(concat[F_I, F_{PI}]), \tag{5}$$

where φ denotes a convolutional embedding layer. This fusion enhances spatial correspondence and strengthens the geometric consistency between the two modalities.

3.3. SCFM

Although the EGFM injects geometric priors into the image branch to enhance geometric perception, the point cloud branch lacks guidance from the rich semantic information

in images. To effectively utilize semantic cues from images and integrate them with point cloud features, an SCFM was designed, as shown in Figure 3. Specifically, the multi-scale image features are first upsampled to their original resolution to align spatially with the point cloud domain. A convolutional block attention mechanism (CBAM) [41] is then applied to enhance both the channel-wise and spatial representations of these features, resulting in semantically augmented feature maps $\widehat{F}_I^l \in \mathbf{R}^{C_l \times H \times W}$.

$$\widehat{F}_I^l = \mathrm{CBAM}\Big(\mathrm{Unsample}(\tilde{F}_I^l)\Big), \tag{6}$$

where $l \in \{1, 2, \ldots, L\}$ denotes the layer index and $L = 4$ is the total number of network layers.

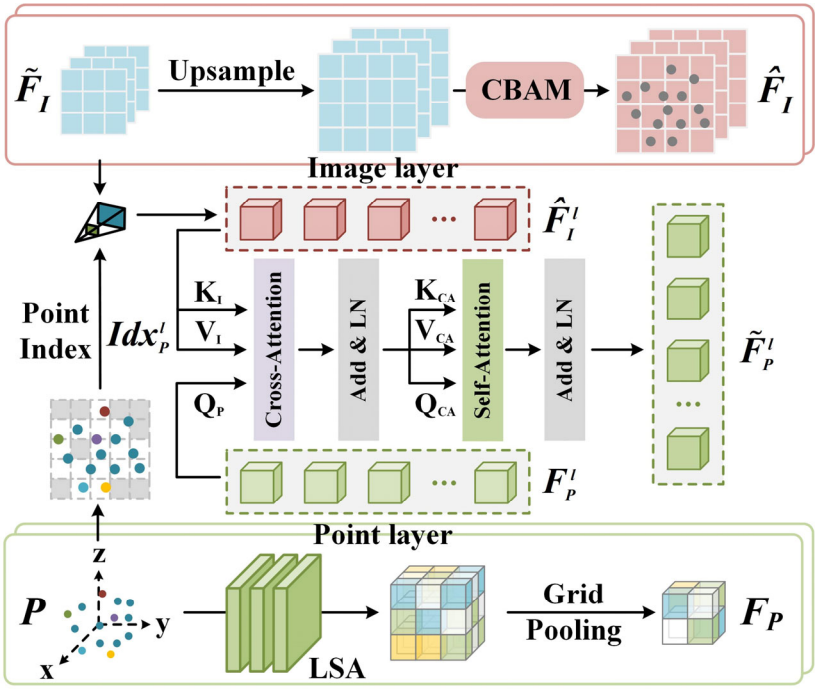


Figure 3. Semantic-aware cross-attention fusion module.

In parallel, the point cloud branch processes input features using local serialized attention (LSA) [42] and density-aware grid pooling (GP) to generate hierarchical geometric representations $F_P^l \in \mathbf{R}^{N_l \times C_l}$:

$$F_P^l = \mathrm{GP}\Big(\mathrm{LSA}(F_P^{l-1})\Big), \tag{7}$$

where N_l and C_l denote the number of points and feature dimensions at the l-th layer, respectively. Although grid pooling aggregates multiple points within the same voxel, it preserves a unique voxel-wise representative index to maintain a deterministic mapping between original points and downsampled representations, thereby establishing consistent correspondence for cross-modal feature retrieval.

To ensure accurate correspondence between point cloud and image features, a bidirectional indexing module is used between the two modalities. Together with F_P^l, the corresponding point indices Idx_P^l that align with the image feature maps are retained for cross-modal correspondence construction. After GP, the retained sampled point indices are denoted as

$$Idx_P^l = \{(u_i, v_i) | p_i \in P^l\}, \tag{8}$$

where P^l represents the set of sampled points after the l-th GP operation. An enhanced 2D image feature $\widehat{F}_I^l(u_i, v_i)$ is then gathered according to Idx_P^l and projected onto the point-

cloud coordinate system as $\widehat{F}_I^l(p_i) = \widehat{F}_I^l(u_i, v_i)$, resulting in the point-aligned semantic feature $\widehat{F}_I^l \in \mathbf{R}^{N_l \times C_l}$.

This alignment establishes a basic semantic correspondence between image and point cloud features, but it does not explicitly capture their interactions across modalities and scales. Cross-attention (CA) and self-attention (SA) mechanisms are further employed to integrate multi-scale semantic and geometric features between F_P^l and $\widehat{F}_I^l$:

$$F_{CA} = \mathrm{CA}(F_P^l, \widehat{F}_I^l) = \mathrm{LN}(\mathrm{Attention}(\mathbf{Q_P}, \mathbf{K_I}, \mathbf{V_I}) + F_P^l) \tag{9}$$

$$\tilde{F}_p^l = \mathrm{SA}(F_{CA}) = \mathrm{LN}(\mathrm{Attention}(\mathbf{Q_{CA}}, \mathbf{K_{CA}}, \mathbf{V_{CA}}) + F_{CA}) \tag{10}$$

where $\mathbf{Q_P}$, $\mathbf{K_I}$, and $\mathbf{V_I}$ are the query, key, and value projections, respectively, derived from the features F_P^l and $\widehat{F}_I^l$; $\mathbf{Q_{CA}}$, $\mathbf{K_{CA}}$, and $\mathbf{V_{CA}}$ are the query, key, and value projections, respectively, derived from feature F_{CA}; LN denotes layer normalization; and $\tilde{F}_p^l$ represents the semantic-aware 3D point features. The CA module captures complementary information across modalities, whereas the SA module reinforces intra-modal semantic consistency within the point cloud. Through this process, semantic knowledge from images is effectively embedded into the point cloud representation, enhancing semantic discrimination in 3D space.

3.4. BAFM

In the deeper layers of the network, inconsistencies may arise between the geometric priors in the image branch and the semantic cues in the point cloud branch. As feature representations become more abstract with network depth, differences between modality-specific cues tend to accumulate. Therefore, a late-stage fusion is introduced after sufficient independent feature extraction. Given that simple summation or concatenation struggles to balance contributions across modalities, we propose a BAFM in the late fusion stage to restore and strengthen cross-modal connections, as shown in Figure 4. The BAFM enhances the integration and refinement of cross-modal features through two key components: a gated adaptive weighting (GAW) block and a bidirectional cross-attention (BCA) fusion block.

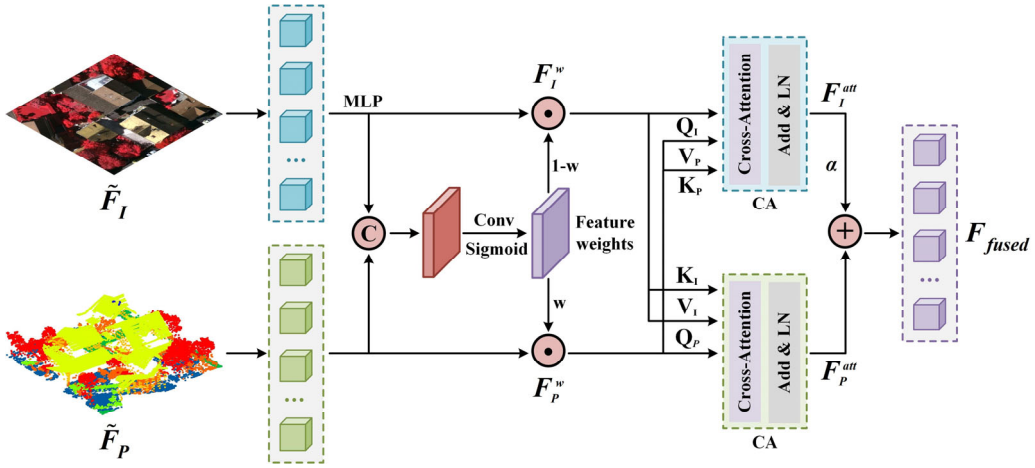


Figure 4. Bidirectional adaptive fusion module.

Following the 2D and 3D decoders, image features mainly emphasize semantic context, whereas point cloud features focus on geometric structure. A GAW mechanism is introduced to adaptively adjust the contributions of modality-specific channels that are

more informative for segmentation. Specifically, the GAW mechanism adaptively weights the point cloud features $\tilde{F}_P \in \mathbf{R}^{N\times C}$ and image features $\tilde{F}_I \in \mathbf{R}^{N\times C}$, as follows.

$$w = \text{Sigmoid}\left(\text{Conv}\left(\left[\tilde{F}_P, \text{MLP}\left(\tilde{F}_I\right)\right]\right)\right) \tag{11}$$

$$F_P^w = w \bigodot \tilde{F}_P \tag{12}$$

$$F_I^w = (1-w) \bigodot \tilde{F}_I \tag{13}$$

where $[\cdot]$ denotes channel-wise concatenation, MLP is multi-layer perceptron, $\odot$ is element-wise multiplication, and w represents a learnable gating weight that dynamically balances the contributions from both modalities.

The GAW mechanism focuses on adjusting the relative contributions of the two modalities at the feature level. Building on this, the BCA mechanism further captures inter-modal correlations by applying cross-attention in both directions, facilitating more effective semantic and geometric interaction.

$$F_P^{att} = \text{CA}(F_P^w, F_I^w) \tag{14}$$

$$F_I^{att} = \text{CA}(F_I^w, F_P^w) \tag{15}$$

$$F_{fused} = F_P^{att} + \alpha F_I^{att} \tag{16}$$

where F_{fused} is the final fused feature, and α is a learnable balancing parameter initialized to 0.5. Through bidirectional modeling, the BAFM aligns image features with geometric constraints from point clouds while integrating semantic context from images into the point-cloud representation. This process jointly improves cross-modal feature integration and preserves geometric consistency.

3.5. Loss Function

To jointly optimize multimodal representations, the overall loss function L_{total} is defined as a weighted combination of four components: the point cloud loss L_{pc}, image loss L_{img}, fusion loss L_f, and cross-modal distillation loss L_{cd}. The total loss is formulated as follows:

$$L_{total} = L_{pc} + L_{img} + \lambda_1 L_f + \lambda_2 L_{cd}, \tag{17}$$

where λ_1 and λ_2 are weighting parameters, both set to 0.5 and 0.5. Specifically, L_{pc} is computed on the point cloud branch of the segmentation head, where the MLP-based classifier predicts per-point semantic labels under the supervision of ground-truth point annotations. L_{img} is defined on the image branch of the segmentation head to supervise pixel-wise semantic predictions using projected ground-truth labels via a 2D projection strategy. The fusion loss L_f is applied to the final fused prediction produced by the late-stage fusion module, where the integrated multimodal features from both the point cloud and image branches are decoded into the final semantic segmentation output.

For L_{pc}, L_{img}, and L_f, a combination of cross-entropy loss and Lovász loss is adopted to address class imbalance and enhance segmentation accuracy, particularly in boundary regions. To further align the semantic representations between the point cloud and image modalities, a Kullback–Leibler (KL) divergence-based distillation loss L_{cd} is introduced [43,44]:

$$L_{cd} = D_{kl}\left(P_{img} \parallel P_{pc}\right), \tag{18}$$

where P_{img} and P_{pc} denote the temperature-scaled logits from the image and point cloud branches, respectively.

4. Experiments and Analysis

4.1. Dataset Description

We evaluated the proposed method on three large-scale multimodal ALS datasets, as shown in Table 1.

Table 1. Multimodal datasets used in the study.

Dataset	Area (km^2)	Point Density (pts/m^2)	GSD (m/pixel)	Image Size (px)	Semantic Classes
ISPRS Vaihingen	1.36	4	0.09	512 × 512	5
N3C-California	725.72	8	0.60	512 × 512	4
UAVScenes	49.48	2	0.20	512 × 512	11

ISPRS Vaihingen dataset [45] is a standard multimodal remote sensing benchmark that comprises very high-resolution (VHR) aerial images and airborne LiDAR point clouds. The VHR imagery has a ground sample distance (GSD) of 0.09 m/pixel and includes near-infrared, red, and green spectral bands. The LiDAR data exhibit an average density of approximately 4 points/m^2. The data include datasets obtained from two areas: Site1 and Site2. The Site1 dataset was used for training and cropped into 46 image–point-cloud pairs with a resolution of 512 × 512 pixels, whereas the Site2 dataset was used as the test set and divided into 33 image–point-cloud pairs of the same size. In this dataset, LiDAR points and images are related through an affine geospatial transformation defined by the orthophoto metadata.

N3C-California dataset [1] covers residential areas in Santa Clara County, California, and provides LiDAR point clouds, digital surface models, and aerial imagery. The aerial imagery has a GSD of 0.6 m/pixel and consists of red, green, and blue spectral bands. The LiDAR data maintain an average density of approximately 8 points/m^2. This dataset contains 10,800 image tiles of 512 × 512 pixels, which were randomly divided into training, validation, and test sets in the ratio of 8:1:1. Similar to the ISPRS Vaihingen dataset, the geometric correspondence between LiDAR points and images is defined by an affine geospatial transformation.

UAVScenes dataset [46] comprises multimodal unmanned aerial vehicle (UAV) data collected from diverse environments, including towns, valleys, airports, and islands, totaling over 120,000 frames. The original aerial imagery has a resolution of 2048 × 2448 pixels with red, green, and blue spectral bands, whereas the LiDAR point clouds have an average density of approximately 2 points/m^2. From the town scenes, 2354 LiDAR-image pairs were selected, and all images were scaled and cropped to a resolution of 512 × 512 pixels. The dataset was randomly partitioned into training, validation, and test subsets in the ratio of 7:1.5:1.5. Since the LiDAR sensor and camera are mounted on the same UAV platform, a standard perspective projection is used to establish the geometric correspondence between LiDAR points and image pixels.

4.2. Experimental Setting and Implementation Details

Model Settings. The proposed network employed a dual-branch architecture. The point cloud branch used Point Transformer v3 (PTV3) [42] with a density-aware GP block as the 3D feature encoder, coupled with a decoder based on nearest-neighbor upsampling. The image branch utilized a ResNet-101 [47] backbone with a feature pyramid network decoder, initialized with ImageNet pre-trained weights. The initial embedding layer had 32 channels. In the point cloud branch, the encoder channel dimensions were set to [64, 128, 256, 512], with corresponding decoder channels of [64, 64, 128, 256]. A stride of

2 was employed for each downsampling stage. For the image branch, the encoder channels were [64, 128, 256, 512], and the final output feature dimension was 256.

Training Setup. All experiments were performed on a single NVIDIA RTX 6000 GPU (NVIDIA, Santa Clara, CA, USA) (48 GB). The models were trained for 300 epochs on the ISPRS Vaihingen and UAVScenes datasets and for 50 epochs on N3C-California, with a consistent batch size of 3 for all datasets. The AdamW optimizer was used with an initial learning rate of 0.002 and a weight decay of 0.005. For point cloud processing, grid sampling was applied with a voxel size of 0.10 m for ISPRS Vaihingen, 0.25 m for N3C-California, and 0.50 m for UAVScenes. The point intensity values were normalized to the range of [0, 1]. For images, normalization was performed using the mean and standard deviation obtained from the ImageNet dataset. In this study, several data augmentation strategies were applied during training. The point clouds were randomly shifted along the z-axis, rotated around the z-axis with a rotation angle sampled from $[-1, 1]$ and a probability of 0.5, and scaled using a random factor in the range of [0.95, 1.05]. In addition, random flipping was performed with a probability of 0.5, and random jittering was applied with a Gaussian noise of $\sigma = 0.005$ clipped to 0.02.

Evaluation Metric. The semantic-segmentation performance of the proposed model was evaluated in terms of overall accuracy (OA), mean accuracy (mAcc), mean intersection over union (mIoU), and per-class intersection over union (IoU). The calculation formula is as follows:

$$OA = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} (TP_c + FP_c)} \tag{19}$$

$$mAcc = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \tag{20}$$

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{21}$$

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} IoU_c \tag{22}$$

where C denotes the number of semantic classes. For class c, TP_c, FP_c, FN_c represent the numbers of true positives, false positives, and false negatives, respectively.

4.3. Semantic Segmentation Results

The proposed MCPFNet was evaluated on three multimodal datasets: ISPRS Vaihingen, N3C-California, and UAVScenes. Additionally, it was compared with several state-of-the-art (SOTA) multimodal fusion methods representing different fusion paradigms, namely 2D-to-3D fusion (PMFNet, VPANet), 3D-to-2D projection (IKDNet), unidirectional knowledge distillation (2DPASS), bidirectional knowledge distillation (CMDFusion), and late-stage fusion frameworks (DFAMNet, Mseg3D).

4.3.1. Results on ISPRS_Vaihingen

The quantitative results presented in Table 2 show that MCPFNet demonstrates the highest overall performance across all evaluation metrics, substantially outperforming existing SOTA multimodal methods. Specifically, it achieved 87.18% OA, 85.36% mAcc, and 74.51% mIoU, which are 2.28%, 4.19%, and 5.26% higher than those of the second-best method (i.e., VPANet), respectively. In particular, the IoU for the car category improved by 12.83%. These results demonstrate that the proposed EGFM and SCFMs effectively integrate geometric and semantic information to enhance the segmentation of small and structurally complex objects.

Table 2. Semantic segmentation results obtained on the ISPRS Vaihingen dataset. **Bold** represents the best results, and underline marks the second-best results.

Methods	OA (%)	mAcc (%)	mIoU (%)	IoU (%)				
				Low Veg.	Roads	Cars	Buildings	Trees
PMFNet	82.25	75.27	62.84	61.95	73.91	35.89	83.59	58.83
2DPASS	84.71	77.38	65.59	67.14	78.16	34.83	84.95	62.87
CMDFusion	81.96	71.80	60.86	62.26	72.93	27.04	81.15	60.92
DFAMNet	81.22	69.66	58.52	63.29	71.65	20.75	78.58	58.31
Mseg3D	83.26	76.46	66.23	64.34	77.13	48.09	81.07	60.54
IKDNet	83.38	75.88	64.23	65.21	78.14	37.95	81.48	58.35
VPANet	84.90	81.17	69.25	66.87	77.62	52.65	83.32	65.78
MCPFNet (Ours)	**87.18**	**85.36**	**74.51**	**70.57**	**82.39**	**65.48**	**86.37**	**67.74**

The visualization results of the ISPRS Vaihingen dataset are shown in Figure 5a–h. Compared with baseline methods, MCPFNet produces clearer boundaries between roads and near-ground vegetation and reduces misclassifications in occluded or structurally incomplete regions. This improvement is particularly evident in road segmentation, where MCPF-Net achieves more complete and continuous road surfaces with fewer gaps. This capability is further demonstrated by the complete extraction of cars on the road and in areas obscured by trees. By performing bidirectional fusion in both middle and late stages, the network fully leverages the complementary of geometry and semantics, resulting in more accurate and consistent segmentation across multiple spatial scales.

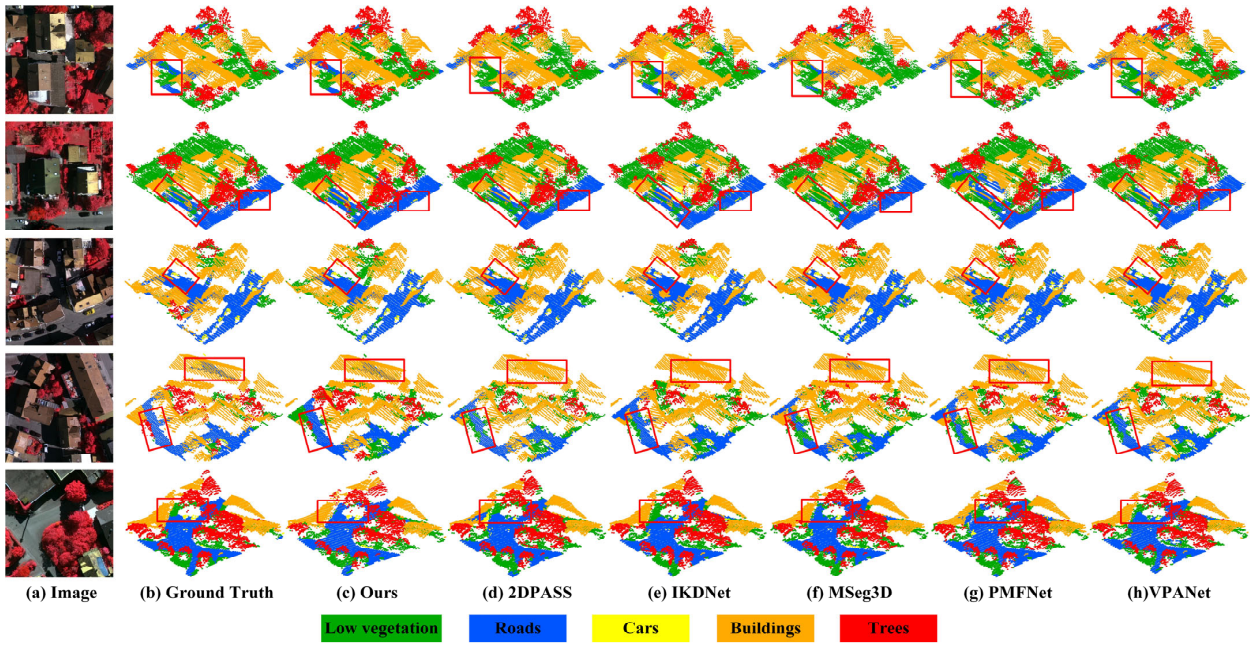


Figure 5. Results obtained on the ISPRS Vaihingen dataset. (**a**) Images, (**b**) ground truths, semantic segmentation results of (**c**) our method, (**d**) 2DPASS, (**e**) IKDNet, (**f**) MSeg3D, (**g**) PMFNet, (**h**) VPANet.

4.3.2. Results on N3C-California

Table 3 shows the quantitative results obtained on the large-scale multimodal benchmark N3C-California dataset. MCPFNet displays the highest overall performance in terms of OA and mIoU while obtaining a competitive mAcc that is slightly lower than that of VPANet. In per-class evaluation, MCPFNet exhibits the highest IoU for the ground and building categories and ranks second in the trees category, with a performance comparable

to that of VPANet. These results indicate that the proposed method effectively leverages multi-height information from LiDAR data to achieve finer-grained segmentation in areas with considerable elevation variations.

Table 3. Semantic segmentation results obtained on the N3C-California dataset. **Bold** represents the best results, and underline marks the second-best results.

Methods	OA (%)	mAcc (%)	mIoU (%)	IoU (%)			
				Others	Ground	Trees	Buildings
PMFNet	84.93	80.25	69.74	40.64	72.74	78.32	87.26
2DPASS	95.69	90.56	86.47	62.11	96.76	93.21	93.79
CMDFusion	95.68	90.60	86.45	62.68	96.46	93.39	93.27
DFAMNet	90.75	79.32	73.16	41.09	89.35	89.13	73.09
Mseg3D	98.44	96.67	94.61	<u>86.87</u>	<u>98.10</u>	98.64	<u>94.82</u>
IKDNet	92.35	86.68	79.37	45.08	95.54	87.22	89.63
VPANet	<u>98.46</u>	**97.13**	<u>94.77</u>	**87.58**	98.05	**98.73**	94.71
MCPFNet (Ours)	**98.63**	<u>97.10</u>	**95.15**	86.70	**98.44**	<u>98.65</u>	**96.80**

The qualitative advantages of MCPFNet are further confirmed by the visual comparisons shown in Figure 6a–h. Among the comparison methods, 2DPASS exhibits evident misclassifications and struggles to maintain the structural integrity of buildings. In comparison, MCPFNet produces more accurate recognition results, attributable to the refinement of geometric and semantic features by the BAFM in the late fusion stage. Additionally, the background class (labeled others) includes diverse structures such as bridges, whose image textures are often visually similar to buildings or roads. While this ambiguity frequently confuses other models, MCPFNet enhances semantic discrimination by integrating elevation information, enabling a more complete and accurate extraction of background information in building- and road-dense areas.

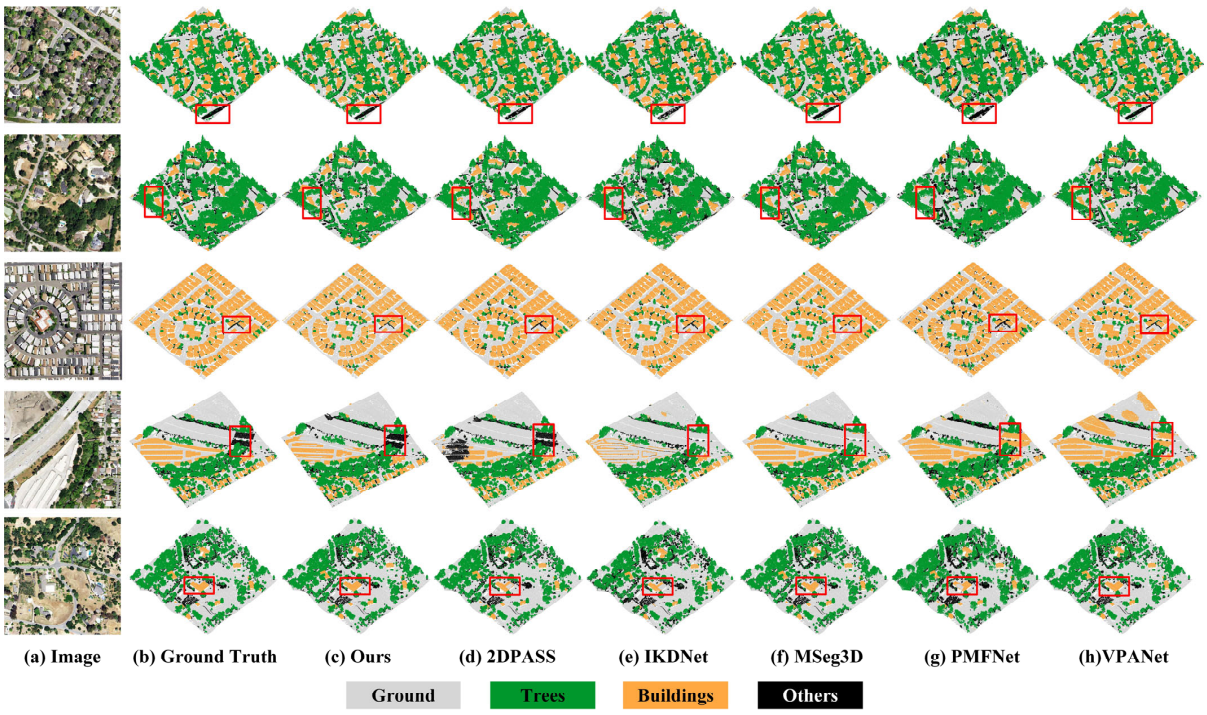


Figure 6. Results obtained on the N3C-California dataset. (**a**) Images, (**b**) ground truths, semantic segmentation results of (**c**) our method, (**d**) 2DPASS, (**e**) IKDNet, (**f**) MSeg3D, (**g**) PMFNet, (**h**) VPANet.

4.3.3. Results on UAVScenes

To evaluate the robustness of MCPFNet in single-frame point cloud scenes, experiments were conducted on the UAVScenes dataset. The quantitative results listed in Table 4 indicate that the proposed method achieves the highest mIoU, outperforming the second-best method, IKDNet, by 1.65%, while maintaining competitive overall accuracy due to the severe class imbalance in UAVScenes. While maintaining high segmentation accuracy for major categories (e.g., roofs, green fields, and various road types), MCPFNet also shows marked improvement in crucial but sparsely distributed categories such as transparent roofs and paved walks. These results validate the effectiveness of the multi-stage collaborative perception fusion strategy in complex and real-world environments.

Table 4. Semantic segmentation results obtained on the UAVScenes dataset. **Bold** represents the best results, and underline marks the second-best results.

Methods	OA (%)	mAcc (%)	mIoU (%)	IoU (%)					
				Others	Roof	Dirt Road	Paved Road	Bridges	Green Field
				Wild Field	Solar Panel	Transp. Roof	Paved Walk	Sedan	Truck
PMFNet	94.27	60.12	55.28	69.05 **89.13**	81.81 00.00	83.07 18.37	91.40 00.00	00.00 **62.38**	95.52 **72.59**
2DPASS	91.21	69.98	56.60	58.12 77.43	80.61 00.00	75.82 58.57	81.32 11.35	47.64 47.75	92.61 47.94
CMDFusion	93.74	66.93	58.64	67.28 87.76	80.23 00.00	81.70 17.92	88.23 7.13	76.23 50.29	95.10 51.80
DFAMNet	90.27	39.20	36.47	59.10 65.30	74.97 00.00	67.12 4.93	18.70 00.00	0.56 31.89	92.10 23.00
Mseg3D	94.36	**72.53**	60.16	68.61 87.91	**84.40** 00.00	**84.58** 36.77	91.61 37.75	31.92 49.78	95.38 53.15
IKDNet	94.08	68.23	61.09	68.63 88.33	82.30 00.00	82.90 25.72	90.47 1.80	**82.53** 55.75	95.25 59.35
VPANet	**94.79**	69.64	60.18	**72.28** 90.22	83.67 00.00	85.61 24.34	**92.24** 1.50	65.45 59.28	**95.68** 51.92
MCPFNet (Ours)	93.17	71.21	**62.76**	63.87 85.21	79.94 00.00	79.90 **59.62**	87.92 **43.75**	75.66 38.94	94.45 43.92

The qualitative results from the town area are shown in Figure 7a–h. This region features diverse roof structures and ambiguous boundaries between dirt roads and wild fields. The high sparsity of single-frame point clouds often leads to fragmented object geometries. Other methods tend to produce mis-segmentation in severely sparse regions. For instance, bridges are misclassified as dirt roads, and sedans are confused in tree-obscured areas. In comparison, MCPFNet performs more accurate semantic segmentation, particularly along complex topographic boundaries, by strengthening feature representation through multi-level collaboration of geometric and semantic information.

4.4. Ablation Studies

4.4.1. Effect of the Fusion Module

As shown in Table 5, we evaluated the effects of different fusion modules and fusion stages on the overall semantic segmentation performance. Model A0 served as the baseline dual-branch framework, consisting of separate image and point cloud branches, where cross-modal feature transfer was guided only by the distillation loss. Based on this baseline, the middle-stage fusion progressively introduced the EGFM and SCFM to establish bidirectional interaction between geometric and semantic representations. Specifically, Model A1 incorporated the EGFM, which enhanced the spatial perception of the image branch by

introducing geometric structural information from the point cloud branch. This improved mIoU and mAcc by 0.62% and 1.27%, respectively.

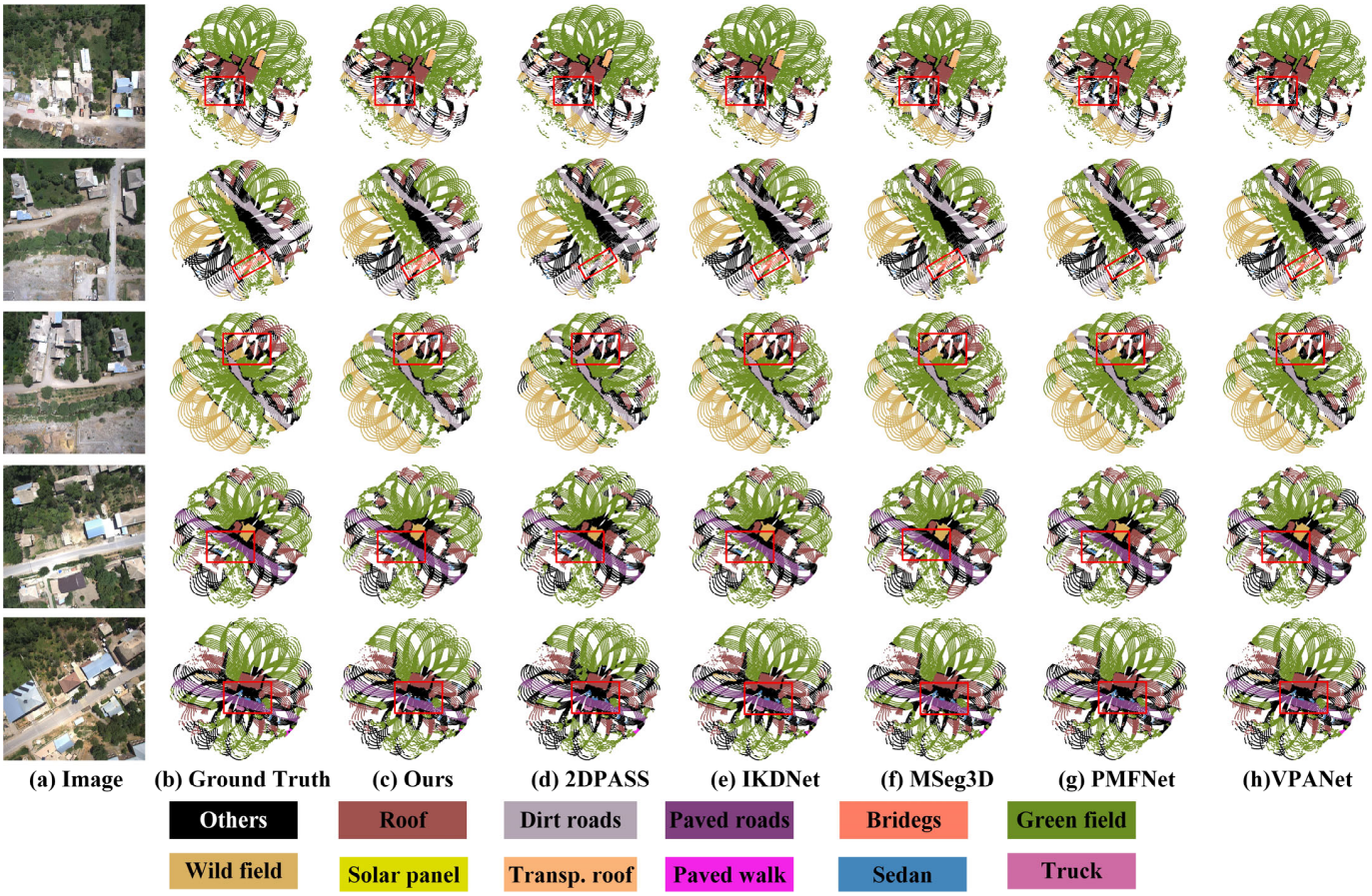


Figure 7. Results obtained on the UAVScenes dataset. (**a**) Images, (**b**) ground truths, semantic segmentation results of (**c**) our method, (**d**) 2DPASS, (**e**) IKDNet, (**f**) MSeg3D, (**g**) PMFNet, (**h**) VPANet.

Table 5. Impact of the fusion modules. √ denotes that the corresponding fusion module is used. **Bold** represents the best results.

Model	Middle Stage		Late Stage	mIoU (%)	mAcc (%)	OA (%)
	EGFM	SCFM	BAFM			
A0				73.10	82.17	87.05
A1	√			73.72	83.44	86.93
A2	√		√	73.85	83.17	**87.36**
A3		√	√	73.94	83.53	86.78
A4	√	√		73.99	82.72	**87.30**
A5			√	73.15	82.58	86.71
MCPFNet	√	√	√	**74.51**	**85.36**	87.18

To further analyze the effect of bidirectional interaction in the middle-stage fusion, the SCFM was additionally introduced in Model A4 based on Model A1. Compared with Model A1, Model A4 improved mIoU and OA by 0.27% and 0.37%, respectively. This indicated that the SCFM further strengthened geometric feature representation by introducing multi-scale semantic information from the image branch, thus forming more effective semantic-geometric interaction during middle-stage fusion. To verify the contribution of late-stage

fusion, the BAFM was individually introduced in Model A5 without the middle-stage fusion modules. Compared with the baseline Model A0, Model A5 achieved only marginal improvements of 0.05% in mIoU and 0.41% in mAcc, while OA decreased by 0.34%. This suggested that direct late-stage fusion without sufficient middle-stage semantic-geometric interaction had limited effectiveness.

Based on Model A1, the BAFM was further introduced in Model A2 to perform adaptive late-stage fusion after geometric-aware enhancement. Compared with Model A1, Model A2 improved mIoU and OA by 0.13% and 0.43%, respectively. This demonstrated that the late-stage fusion further refined modality-specific features and improved the final feature aggregation. Similarly, Model A3 integrated the SCFM and BAFM while removing the EGFM and achieved a 0.09% higher mIoU than Model A2. This indicated that semantic-guided geometric enhancement also contributed positively to the subsequent adaptive fusion process.

When all fusion modules were jointly employed, the proposed MCPFNet achieved the best performance, with mIoU, mAcc, and OA reaching 74.51%, 85.36%, and 87.18%, respectively. Compared with Model A0, the overall performance improved by 1.41%, 3.19%, and 0.13%, respectively. These results demonstrated that the proposed method effectively enhanced cross-modal feature interaction and segmentation performance through multi-stage semantic-aware and geometric-aware fusion.

To provide a qualitative analysis of the effects of different fusion stages, the segmentation results are visualized in Figure 8. The middle-stage fusion (Figure 8a) produced clearer separations between semantically similar categories, such as trees and low-height vegetation. The late-stage fusion (Figure 8b) further refined road boundaries and preserved the structural integrity of small cars. When both fusion stages were jointly employed (Figure 8c), the results exhibited a clearer spatial hierarchy and more complete semantic structures. These results further confirmed that both fusion stages were important for achieving optimal segmentation performance.

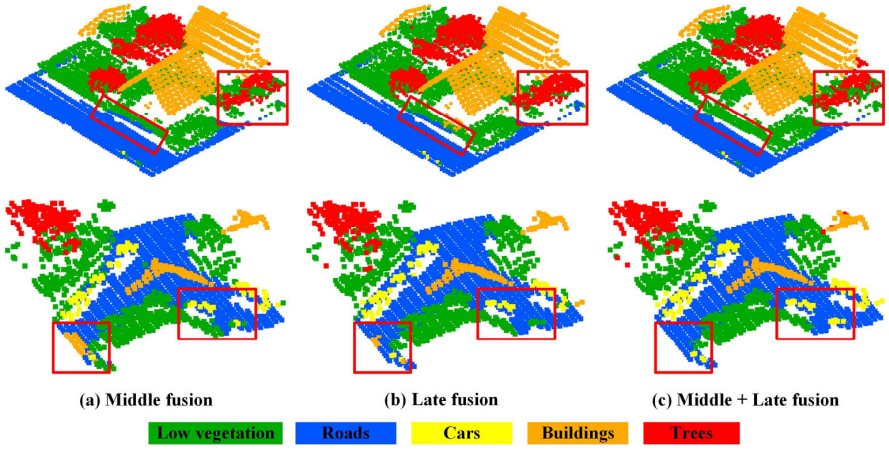


Figure 8. Comparative visualizations of the fusion stage. (**a**) Middle-stage fusion, (**b**) late-stage fusion, and (**c**) our complete framework combining both.

4.4.2. Effect of Geometric Features in the EGFM

As shown in Table 6, we evaluated the contribution of different geometric features in EGFM (see Equation (3)). By solely using the maximum height (Max_Z), the vertical structure of objects was captured, achieving an mIoU of 73.38%. Incorporating the mean intensity (Mean_R) improved mAcc by 2.04%, indicating that reflectance information helped distinguish objects with similar geometric structures. The subsequent addition of the minimum height (Min_Z) and mean height (Mean_Z) further improved mIoU by

0.20% and 0.07%, respectively. Although the performance gains were relatively limited, these height-based features shared similar geometric information and thus contributed incremental improvements. Their combination still provided more complete structural cues and resulted in more robust feature representation. These results demonstrated that the integration of multiple geometric features improved the overall feature representation.

Table 6. Effects of geometric features in the EGFM. √ denotes that the corresponding fusion module is used. **Bold** represents the best results.

Max_Z	Min_Z	Mean_Z	Mean_R	mIoU (%)	mAcc (%)	OA (%)
√				73.38	82.24	**86.95**
			√	73.36	84.28	86.00
√			√	73.45	**84.93**	86.33
√	√		√	73.65	83.49	85.99
√	√	√	√	**73.72**	83.44	86.93

We further visualized the spatial distributions of these geometric features in Figure 9. The height-based features (i.e., Max_Z, Min_Z, and Mean_Z) jointly characterized the spatial hierarchy through elevation variations, which improved the discrimination between buildings, roads, and other planar structures. In addition, the mean intensity feature refined the boundaries of vegetation and vehicles by exploiting differences in surface reflectance. These visual results further showed that multi-feature fusion provided a more complete geometric description of 3D objects and improved segmentation performance after fusion with image features.

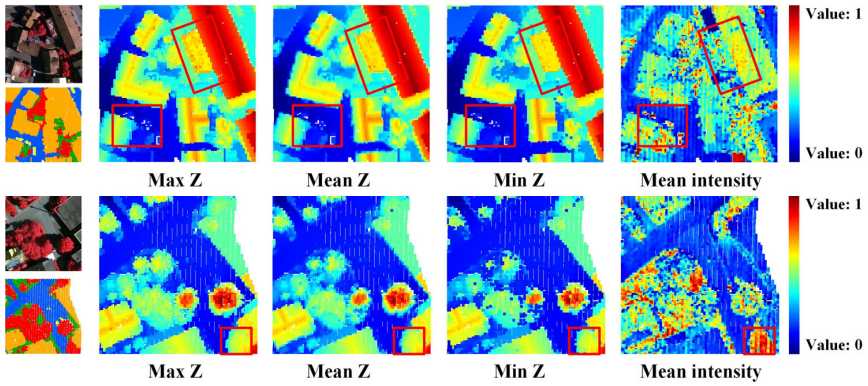


Figure 9. Visualization of geometric features in BGFM.

4.4.3. Effect of Fusion Layers in the SCFM

As shown in Table 7, we evaluated the effect of the number of fusion layers in the SCFM on the performance of the proposed model. The segmentation accuracy improved progressively as the number of fusion layers increased, with the best performance being achieved when all four layers were fused. Specifically, mIoU increased from 74.07% to 74.51%, while mAcc improved from 83.34% to 85.36%, indicating that deeper multi-layer fusion enhanced feature interaction across modalities.

From the computational perspective, increasing the number of fusion layers led to a gradual increase in model complexity, including parameters, inference latency, and memory consumption. However, the overall overhead remained relatively small, with only 0.75 M additional parameters and 19.71 ms extra latency when extending the fusion from one layer to four layers.

This multi-layer design integrates global semantic representations from deeper features and local structural details from shallower features. The results indicated that multi-

layer fusion effectively leveraged information from different receptive fields, thereby enhancing the consistency and robustness of cross-modal feature representations.

Table 7. Effects of fusion layers in the SCFM. **Bold** represents the best results.

Fusion Layers	mIoU (%)	mAcc (%)	OA (%)	Parameters (M)	Latency (ms)	Memory (MB)
1	74.07	83.34	86.53	**86.76**	**312.28**	**1355.83**
2	74.21	83.51	**87.33**	86.81	321.93	1358.73
3	74.32	84.68	87.00	86.98	327.76	1360.04
4	**74.51**	**85.36**	87.18	87.51	331.99	1373.98

4.4.4. Effect of Fusion Block in the BAFM

We evaluated the effectiveness of the two key components within the BAFM, as shown in Table 8. The concatenation operation directly combines point cloud and image features, followed by a convolution layer that adjusts the output channel dimensions. Compared with this simple operation, the GAW block improved mIoU by 0.21% and mAcc by 0.76%, while slightly reducing performance on individual classes such as road and building. Incorporating the BCA block further increased mIoU, mAcc, and OA by 0.10%, 0.66%, and 0.69%, respectively, and also achieved more balanced improvements across road and building classes. These results demonstrated that GAW and BCA blocks enhanced cross-modal interaction and improved both overall and category-level segmentation consistency.

Table 8. Effects of GAW and BCA block in the BAFM. **Bold** represents the best results.

Block	mIoU (%)	mAcc (%)	OA (%)	IoU (%)	
				Road	Building
Concat	74.20	83.94	87.06	80.57	**86.51**
GAW	74.41	84.70	86.49	78.65	84.41
GAW + BCA	**74.51**	**85.36**	**87.18**	**82.39**	86.37

Furthermore, the BCA module was quantitatively analyzed. As shown in Figure 10, after introducing the BCA module, the segmentation results for objects such as roads and buildings became more complete and exhibited clearer structural boundaries. These visual improvements demonstrated that the BCA block effectively integrated the high-confidence features selected by the GAW block, further strengthening the joint representation of geometric and semantic information.

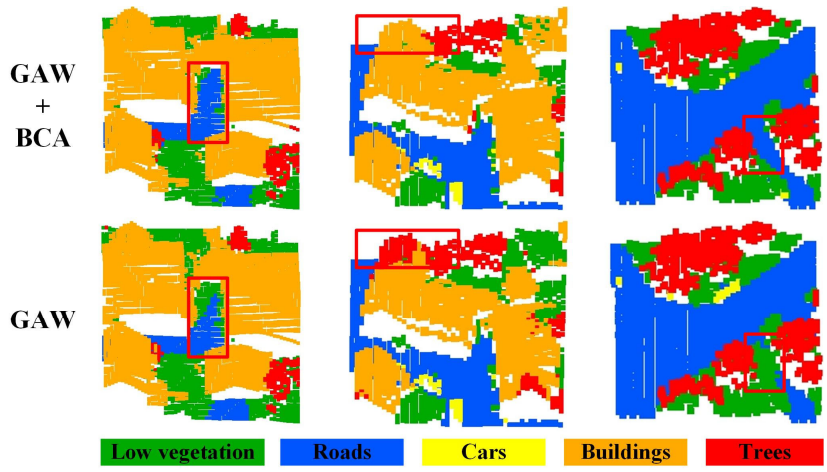


Figure 10. Visualization results with and without the BCA block in BAFM.

4.4.5. Effect of Gaussian Kernel in Mask-Guided Interpolation

To analyze the influence of the Gaussian kernel parameter σ in Equation (4), we visualized the projected geometric feature maps and corresponding masks under different σ settings, as shown in Figure 11. When σ = 0.5, the interpolation range was too limited, leaving many empty grid cells unfilled. Increasing σ to 1.0 improved the feature propagation, while some sparse regions remain incomplete. When σ = 1.5, the projected features became more complete and preserved clearer structural boundaries. In contrast, σ = 2.0 introduced excessive smoothing, which weakened local structural details. These results indicated that an appropriate Gaussian kernel can better balance feature completion and structural preservation in sparse projection regions.

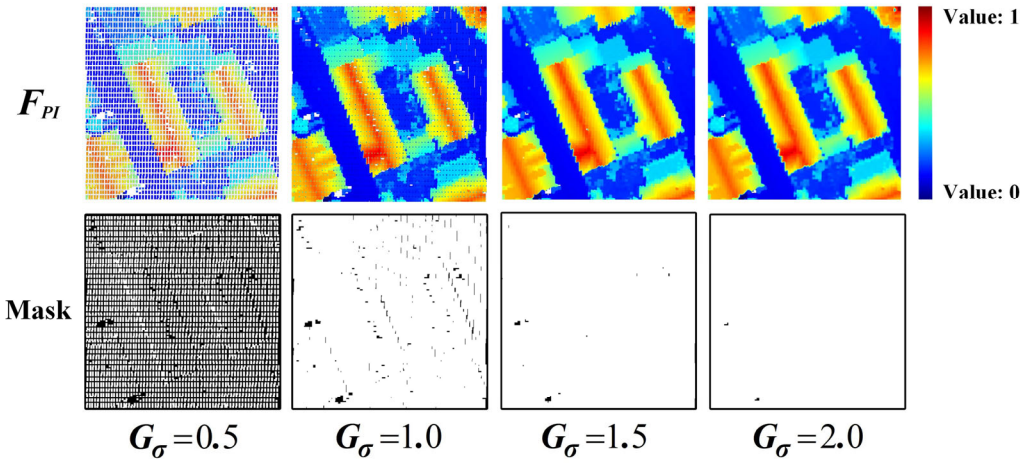


Figure 11. Visualization results of mask-guided Gaussian interpolation under different σ settings.

4.4.6. Effect of Loss Weight Parameters

As shown in Table 9, we evaluated the effect of different weight settings for the fusion loss L_f and distillation loss L_{cd}. When the loss weights were imbalanced, the segmentation performance generally decreased. Specifically, assigning a larger weight to the distillation loss improved the overall accuracy, while the category-level consistency was weakened, resulting in lower mIoU and mAcc. In contrast, increasing the fusion loss weight enhanced feature aggregation but reduced the supervision from cross-modal knowledge transfer. The best performance was achieved when both losses were equally weighted at 0.5, which indicated that balanced optimization between feature fusion and distillation contributed to more effective cross-modal representation learning.

Table 9. Effects of different loss weight settings for fusion loss and distillation loss. **Bold** represents the best results.

λ_1	λ_2	mIoU (%)	mAcc (%)	OA (%)
0.1	0.9	74.39	84.85	**87.42**
0.3	0.7	74.19	83.77	86.96
0.5	0.5	**74.51**	**85.36**	87.18
0.7	0.3	74.26	84.39	86.15
0.9	0.1	73.91	83.03	87.25

5. Discussion

In this section, we further analyze the proposed method from multiple perspectives, including fusion effectiveness, computational efficiency, and model limitations.

5.1. Effect of Multimodal Fusion

To further evaluate the effectiveness of multimodal fusion, we compared the proposed MCPFNet with several representative single-modal point cloud and image-based segmentation methods, as shown in Table 10. Among the single-modal point cloud methods, PTv3 and PSMamba achieved relatively strong performance, with mIoU of 68.22% and 69.79%, respectively, indicating their effectiveness in geometric feature learning. However, the proposed MCPFNet achieved 74.51% mIoU, outperforming PTv3 by 6.29% and PSMamba by 4.72%. Compared with single-modal image-based methods, the improvement was more pronounced. These results demonstrated that multimodal fusion effectively integrated complementary semantic and geometric information, leading to more discriminative feature representations and improved segmentation performance.

Table 10. Comparison between single-modal and multimodal segmentation methods. **Bold** represents the best results.

Modality	Method	mIoU (%)	mAcc (%)	OA (%)
Single-modal (Point cloud)	Cylinder3D	59.90	70.37	82.82
	RandLA-Net	60.16	72.80	80.36
	PTv3	68.22	78.31	83.64
	PSMamba	69.79	78.90	85.52
	Point SSM	45.26	62.35	68.43
Single-modal (Image)	ResNet	51.20	62.94	77.29
	Swin Transformer	50.54	62.62	76.48
	SegFormer	51.08	63.24	76.48
Mutimodal	MCPFNet	**74.51**	**85.36**	**87.18**

5.2. Effect of Multi-Stage Fusion

To further verify the effectiveness of the proposed multi-stage fusion strategy, we conducted a comparison with several representative multimodal methods under the same backbone, as shown in Table 11. Compared with existing methods, MCPFNet consistently achieved superior performance. Specifically, it outperformed 2DPASS and CMDFusion by a clear margin in mIoU, while also maintaining the highest mAcc and OA. In particular, DFAMNet showed limited performance under the same backbone setting, indicating that single-stage fusion strategies were insufficient to fully exploit cross-modal information. In contrast, the proposed multi-stage fusion progressively integrated cross-modal features at different levels, leading to more effective feature interaction and more robust semantic representation.

Table 11. Comparison of different multimodal fusion methods under the same backbone settings. **Bold** represents the best results.

Method	2D Backbone	3D Backbone	mIoU (%)	mAcc (%)	OA (%)
2DPASS	ResNet-101	SPVCNN	65.59	77.38	84.71
	ResNet-101	PTv3	69.01	78.11	86.56
CMDFusion	ResNet-101	SPVCNN	60.86	71.80	81.96
	ResNet-101	PTv3	68.87	78.91	85.77
DFAMNet	ResNet-101	SPUNet	58.52	69.66	81.22
	ResNet-101	PTv3	62.64	73.04	85.77
MCPFNet	ResNet-101	PTv3	**74.51**	**85.36**	**87.18**

5.3. Effect of Different Backbones

The semantic-segmentation performance of MCPFNet was evaluated for various 2D and 3D backbones. Additionally, the number of model parameters, training latency, and memory usage of the model for different backbones were compared.

As shown in Figures 12 and 13, using ResNet101 as the 2D backbone led to a more efficient balance between accuracy and efficiency. Specifically, the combination of ResNet101 and PSMamba [48] resulted in the highest mIoU (75.15%), whereas the combination of ResNet101 with PTv3 resulted in the highest OA (78.18%) and mAcc (85.36%). Although ResNet152 substantially increased the number of parameters and memory consumption, it did not yield further improvements in accuracy. This implies that excessively deep 2D backbones may introduce feature redundancy in multimodal fusion. By contrast, ResNet50 exhibits competitive performance while maintaining lower latency, rendering it more suitable for real-time or resource-constrained applications.

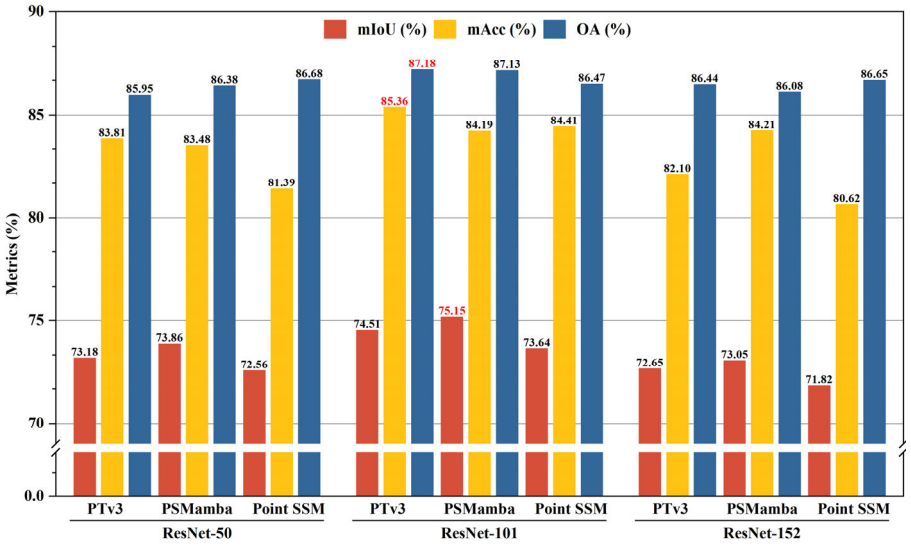


Figure 12. Accuracy comparison of different models. Values in red indicate the best performance.

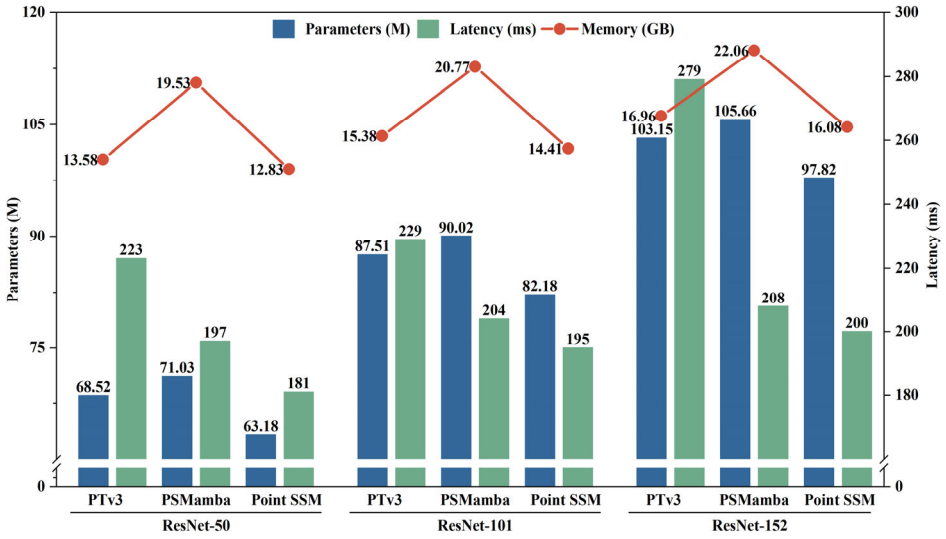


Figure 13. Computational complexity comparison of different models.

When paired with the same 2D backbone, PSMamba outperformed both the PTV3 and PointSSM backbones [49] in terms of most metrics. These results suggest that the 3D spatial serialization and state-space model more effectively capture long-range geometric

dependencies for point cloud semantic segmentation. However, the performance gain of PSMamba was achieved at the cost of increased parameter size and memory usage compared to those of PTV3 and PointSSM. Overall, PTV3 provides an excellent balance of efficiency and capability in our framework, while PSMamba is more suited for high-performance multimodal fusion.

5.4. Computational Efficiency Analysis of Fusion Modules

As shown in Table 12, we analyzed the computational cost of the proposed fusion modules. The baseline model involved relatively high computational complexity due to the strong image and point cloud backbones. After progressively introducing the EGFM, SCFM, and BAFM, the overall parameters increased by only 0.83 M, and the inference latency increased by 56.74 ms. Meanwhile, the memory consumption increased by only 11.40 MB. These results indicated that the proposed fusion modules introduced limited additional overhead while maintaining efficient multimodal feature interaction.

Table 12. Computational cost analysis of different fusion modules.

Model	Parameters (M)	Latency (ms)	Memory (MB)
Baseline	86.68	275.25	1362.58
+EGFM	86.69	295.50	1356.84
+EGFM and SCFM	87.46	319.18	1370.42
+EGFM and SCFM & BAFM (Ours)	87.51	331.99	1373.98

5.5. Complexity Analysis Across Different Datasets

Table 13 presents the computational complexity comparison of different multimodal semantic segmentation methods across three datasets. All methods were tested on a single NVIDIA RTX A6000 GPU (48 GB) with a batch size of 1. MCPFNet achieved competitive segmentation performance while maintaining stable computational cost across different datasets. Compared with lightweight methods such as IKDNet, the proposed network required more parameters and memory consumption but achieved substantially better segmentation accuracy. Although VPANet introduced higher parameter complexity and memory overhead, its segmentation performance remained lower than MCPFNet. These results indicate that the proposed multi-stage fusion strategy achieves a favorable balance between segmentation performance and computational cost across different multimodal scene understanding benchmarks.

Table 13. Comparison results of computational complexity of different methods across three datasets. **Bold** represents the best results.

Method	Parameters (M)	ISPRS_ Vaihingen		N3C-California		UAVScenes	
		Latency (ms)	Memory (MB)	Latency (ms)	Memory (MB)	Latency (ms)	Memory (MB)
PMFNet	65.52	209.57	1045.18	243.32	1062.47	136.39	1046.70
2DPASS	56.73	141.73	961.60	**232.84**	1482.25	215.79	951.82
CMFusion	58.75	**118.89**	1070.50	250.39	2026.62	**92.53**	1032.98
DFAMNet	56.78	148.64	992.35	237.42	1502.10	110.95	952.15
Mseg3D	49.69	191.84	922.17	327.89	1059.48	130.79	917.24
IKDNet	**18.51**	294.60	**316.60**	2133.07	**334.60**	306.46	**315.48**
VPANet	102.21	303.11	1625.81	455.37	1668.52	169.96	1624.77
MCPFNet	87.51	331.99	1373.98	680.64	1399.59	397.47	1370.46

5.6. Limitations

The proposed MCPFNet achieves strong segmentation performance, but its computational cost is relatively high. The multi-stage fusion design introduces additional overhead due to cross-modal interactions at different feature levels. Although the fusion modules are lightweight, the overall model complexity remains relatively high because of the strong image and point cloud backbones. Future work will explore more efficient multimodal architectures by adopting lightweight backbone networks and designing more compact fusion modules to achieve a better balance between segmentation accuracy and computational efficiency.

6. Conclusions

This study proposes an innovative network model, MCPFNet, to enhance semantic segmentation of urban scenes. The network introduces an EGFM in the middle stage to transfer precise 3D structural priors from point clouds to images, strengthening spatial perception representation in the 2D branch. This is complemented by an SCFM, which integrates multi-scale semantic and geometric features to achieve cross-modal complementary fusion. In the late fusion stage, a bidirectional adaptive fusion module further refines geometric and semantic features, effectively reducing semantic ambiguity and improving segmentation accuracy. Extensive experiments on three multimodal datasets demonstrate the effectiveness of MCPFNet and its remarkable performance in fine-grained semantic scene understanding. In particular, evaluations on the UAVScenes single-frame dataset confirm the robustness of our method in low-altitude UAV scenes, which present perceptual challenges analogous to those in autonomous driving. In the future, we plan to investigate the feasibility of using our approach in more complex low-altitude 3D perception tasks to further contribute to the field of low-altitude observation and perception.

Author Contributions: Conceptualization, H.L. (Huchen Li) and F.D.; methodology, H.L. (Huchen Li) and X.L.; formal analysis, H.L. (Huchen Li); investigation, H.L. (Huchen Li); visualization, H.L. (Huchen Li) and B.L.; resources, W.H. and F.D.; validation, W.H. and X.L.; supervision, F.D.; funding acquisition, F.D.; writing—original draft, H.L. (Huchen Li); writing—review and editing, H.L. (Huchen Li), W.H., X.L., B.L., H.L. (Haibing Liu), S.C. and F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the "Innovation Yongjiang 2035" Key R&D Programme of Ningbo City, grant number 2025Z191, and the Ningbo Science and Technology Innovation Yongjiang 2035 Key Technology Project, grant number 2024Z298.

Data Availability Statement: The ISPRS Vaihingen dataset can be obtained from https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/Default.aspx (accessed on 9 November 2024). The N3C-California dataset can be obtained from https://github.com/wymqqq/IKDNet-pytorch (accessed on 28 March 2025). The UAVScenes dataset can be obtained from https://github.com/sijieaaa/UAVScenes (accessed on 10 September 2025).

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

References

1. Wang, Y.; Wan, Y.; Zhang, Y.; Zhang, B.; Gao, Z. Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and LiDAR point clouds. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 385–404. [CrossRef]
2. Li, J.; Li, Y.; Zhang, S.; Chen, Y. Image-point cloud embedding network for simultaneous image-based farmland instance extraction and point cloud-based semantic segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104361. [CrossRef]
3. Huang, C.; Fang, S.; Wu, H.; Wang, Y.; Yang, Y. Low-altitude intelligent transportation: System architecture, infrastructure, and key technologies. *J. Ind. Inf. Integr.* **2024**, *42*, 100694. [CrossRef]

4. Xu, H.; Chen, J.; Meng, S.; Wang, Y.; Chau, L.-P. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Inf. Fusion* **2025**, *114*, 102671. [CrossRef]
5. Kong, L.; Xu, X.; Ren, J.; Zhang, W.; Pan, L.; Chen, K.; Ooi, W.T.; Liu, Z. Multi-Modal Data-Efficient 3D Scene Understanding for Autonomous Driving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 3748–3765. [CrossRef] [PubMed]
6. Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; Chanussot, J. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102926. [CrossRef]
7. Su, C.; Hu, X.; Meng, Q.; Zhang, L.; Shi, W.; Zhao, M. A multimodal fusion framework for urban scene understanding and functional identification using geospatial data. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *127*, 103696. [CrossRef]
8. Li, H.; Guan, H.; Ma, L.; Lei, X.; Yu, Y.; Wang, H.; Delavar, M.R.; Li, J. MVPNet: A multi-scale voxel-point adaptive fusion network for point cloud semantic segmentation in urban scenes. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *122*, 103391. [CrossRef]
9. Yu, H.; Yan, Q.; Xiao, T.; Deng, F. MeshSegNet: A local to global feature fusion-based semantic segmentation approach for urban meshes. *Geo-Spat. Inf. Sci.* 2025, *in press*. [CrossRef]
10. Samadzadegan, F.; Toosi, A.; Dadrass Javan, F. A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: Current status and prospects. *Int. J. Remote Sens.* **2024**, *46*, 1327–1402. [CrossRef]
11. Liu, W.; Wang, H.; Qiao, Y.; Zhang, H.; Yang, J. DLAFNet: Direct LiDAR-Aerial Fusion Network for Semantic Segmentation of 2-D Aerial Image and 3-D LiDAR Point Cloud. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 1864–1875. [CrossRef]
12. Guo, Z.; Xu, R.; Feng, C.-C.; Zeng, Z. PIF-Net: A Deep Point-Image Fusion Network for Multimodality Semantic Segmentation of Very High-Resolution Imagery and Aerial Point Cloud. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5700615. [CrossRef]
13. Rizaldy, A.; Gloaguen, R.; Fassnacht, F.E.; Ghamisi, P. HyperPointFormer: Multimodal Fusion in 3-D Space With Dual-Branch Cross-Attention Transformers. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 21254–21274. [CrossRef]
14. Ballouch, Z.; Hajji, R.; Poux, F.; Kharroubi, A.; Billen, R. A Prior Level Fusion Approach for the Semantic Segmentation of 3D Point Clouds Using Deep Learning. *Remote Sens.* **2022**, *14*, 3415. [CrossRef]
15. Cui, Y.; Zhu, H.; Li, L.; Wu, Y.; Zuo, Y.; Wu, Q. Improving the segmentation of confusable indoor structures using point convolution and sparse vector attention. *Geo-Spat. Inf. Sci.* 2025, *in press*. [CrossRef]
16. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
17. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11108–11117.
18. Cen, J.; Zhang, S.; Pei, Y.; Li, K.; Zheng, H.; Luo, M.; Zhang, Y.; Chen, Q. CMDFusion: Bidirectional Fusion Network With Cross-Modality Knowledge Distillation for LiDAR Semantic Segmentation. *IEEE Rob. Autom. Lett.* **2024**, *9*, 771–778. [CrossRef]
19. Rong, M.; Shen, S. 3D Semantic Segmentation of Aerial Photogrammetry Models Based on Orthographic Projection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7425–7437. [CrossRef]
20. Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; Li, Z. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 677–695.
21. Luo, Y.; Han, T.; Liu, Y.; Su, J.; Chen, Y.; Li, J.; Wu, Y.; Cai, G. CSFNet: Cross-Modal Semantic Focus Network for Semantic Segmentation of Large-Scale Point Clouds. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5701415. [CrossRef]
22. Hang, R.; Li, Z.; Ghamisi, P.; Hong, D.; Xia, G.; Liu, Q. Classification of Hyperspectral and LiDAR Data Using Coupled CNNs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4939–4950. [CrossRef]
23. Li, P.; Chen, Z.; Long, C.; Zhang, H.; Lv, Y.; Huang, R.; Dong, Z.; Yang, B. A novel projection map driven multimodal fusion framework for ALS point cloud semantic segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104907. [CrossRef]
24. Zhang, S.; Meng, X.; Liu, Q.; Yang, G.; Sun, W. Feature-Decision Level Collaborative Fusion Network for Hyperspectral and LiDAR Classification. *Remote Sens.* **2023**, *15*, 4148. [CrossRef]
25. Xu, J.; Zhang, R.; Dou, J.; Zhu, Y.; Sun, J.; Pu, S. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16024–16033.
26. Liu, Y.; Chen, R.; Li, X.; Kong, L.; Yang, Y.; Xia, Z.; Bai, Y.; Zhu, X.; Ma, Y.; Li, Y. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 21662–21673.
27. Wang, Z.; Li, C.; Ma, J.; Feng, Z.; Xiao, L. PVI-Net: Point–Voxel–Image Fusion for Semantic Segmentation of Point Clouds in Large-Scale Autonomous Driving Scenarios. *Information* **2024**, *15*, 148. [CrossRef]
28. Wang, C.; Li, J.; Chen, X.; Song, D.; Wang, B. Align and Blend: A Unified Multi-Modal LiDAR Segmentation Network. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 6651–6660. [CrossRef]
29. Zou, Z.; Li, Y. Efficient urban-scale point clouds segmentation with bev projection. *arXiv* **2021**, arXiv:2109.09074.

30. Wei, W.; Oswald, M.R.; Nejadasl, F.K.; Gevers, T. Apnet: Urban-level scene segmentation of aerial images and point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 1755–1764.
31. Song, Z.; Guo, M.; Wu, L.; Liu, H.; Huang, Y.; Liu, Z. DAPSS: A Novel Network for DOM Assisted Oblique Photography Point Cloud Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 18264–18280. [CrossRef]
32. Dai, H.; Hu, X.; Zhang, J.; Shu, Z.; Xu, J.; Du, J. Large-Scale ALS Point Cloud Segmentation via Projection-Based Context Embedding. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5704216. [CrossRef]
33. Zhuang, Z.; Li, R.; Jia, K.; Wang, Q.; Li, Y.; Tan, M. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 16280–16290.
34. Tan, M.; Zhuang, Z.; Chen, S.; Li, R.; Jia, K.; Wang, Q.; Li, Y. EPMF: Efficient perception-aware multi-sensor fusion for 3D semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 8258–8273. [CrossRef] [PubMed]
35. Li, M.; Wang, G.; Zhu, M.; Li, C.; Liu, H.; Pan, X.; Long, Q. DFAMNet: Dual fusion attention multi-modal network for semantic segmentation on LiDAR point clouds. *Appl. Intell.* **2024**, *54*, 3169–3180. [CrossRef]
36. Li, J.; Lu, J.; Wei, M.; Kang, H.; Zhang, Q. TPV-IGKD: Image-Guided Knowledge Distillation for 3D Semantic Segmentation With Tri-Plane-View. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 10405–10416. [CrossRef]
37. Lin, F.; Lin, T.; Yao, Y.; Ren, H.; Wu, J.; Cai, Q. VPA-Net: A visual perception assistance network for 3d lidar semantic segmentation. *Pattern Recognit.* **2025**, *158*, 111014. [CrossRef]
38. Sun, T.; Zhang, Z.; Tan, X.; Peng, Y.; Qu, Y.; Xie, Y. Uni-to-Multi Modal Knowledge Distillation for Bidirectional LiDAR-Camera Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 11059–11072. [CrossRef] [PubMed]
39. Li, J.; Dai, H.; Han, H.; Ding, Y. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 21694–21704.
40. Liu, Y.; Liu, Y.; Duan, Y. MVG-Net: LiDAR Point Cloud Semantic Segmentation Network Integrating Multi-View Images. *Remote Sens.* **2024**, *16*, 2821. [CrossRef]
41. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
42. Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; Zhao, H. Point transformer v3: Simpler faster stronger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 4840–4851.
43. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2604–2613.
44. Tian, Z.; Cui, J.; Jiang, L.; Qi, X.; Lai, X.; Chen, Y.; Liu, S.; Jia, J. Learning context-aware classifier for semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 2438–2446.
45. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS benchmark on urban object Classification and 3D Building Reconstruction. In Proceedings of the XXII ISPRS Congress, Melbourne, Australia, 25 August–1 September 2012.
46. Wang, S.; Li, S.; Zhang, Y.; Yu, S.; Yuan, S.; She, R.; Guo, Q.; Zheng, J.; Howe, O.K.; Chandra, L. UAVScenes: A Multi-Modal Dataset for UAVs. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–25 October 2025; pp. 28946–28958.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Li, H.; Huang, W.; Liu, J.; Chen, K.; Deng, F. Grid Point Serialized Transformer for LiDAR Point Cloud Semantic Segmentation in Various Densities and Heights Scenes. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5705815. [CrossRef]
49. Li, D.; Guan, J.; Chen, Z.; Liao, J.; Du, J. PointSSM: State space model for large-scale LiDAR point cloud semantic segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104830. [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.