symmetry

MDPI

Article

Edge-Texture-Aware Semantic Dual-Query Fusion for Multimodal 3D Object Detection

Yuehan Wu [1,2], Zheng Zheng [1,3], Kai Liu [1,2], Leyan Chen [4] and Rihan Wu [5,*]

[1] School of Electronics and Information Engineering, Beihang University, Beijing 100191, China; buaawyhwr@126.com (Y.W.); zhengzheng@buaa.edu.cn (Z.Z.); liuk@buaa.edu.cn (K.L.)
[2] State Key Laboratory of CNS/ATM, Beihang University, Beijing 100191, China
[3] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
[4] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China; leyanchen9@mail.tsinghua.edu.cn
[5] School of Electronics and Information Engineering, Inner Mongolia University, Hohhot 010021, China
* Correspondence: rihanwu@imu.edu.cn

Abstract

Multimodal 3D object detection benefits from the complementary nature of camera images and LiDAR point clouds. However, existing voxel–pixel fusion methods typically rely on relatively coarse cross-modal interactions, which limit fine-grained structural modeling and degrade performance on small safety-critical objects. To address this issue, we propose ETA-SDQF, an edge-texture-aware semantic dual-query fusion framework designed to enhance 3D perception of vehicles, cyclists, and pedestrians. The proposed method first introduces an edge-texture-aware image backbone (ETAIB) based on the discrete wavelet transform (DWT), which improves the representation of multi-scale fine-grained image features. Then, we design a dual-query-guided attention fusion (DQGAF) module, which leverages deformable attention to adaptively aggregate voxel-aligned multi-scale image features under joint semantic and edge-texture guidance. Finally, we adopt a hybrid 3D feature learning strategy inspired by PV-RCNN, combining voxel-based feature learning with PointNet-style feature abstraction for processing fused features. This design improves the utilization of voxel features enriched with image semantics, thereby facilitating more reliable 3D object proposal generation. Experimental results on the KITTI dataset demonstrate that the proposed framework achieves better performance compared to existing baseline methods. It consistently improves pedestrian and cyclist detection, while maintaining competitive performance on car detection across different difficulty levels, showing potential benefits on challenging KITTI samples.

Keywords: autonomous driving; multimodal 3D object detection; dual-query fusion; edge-texture-aware; camera-LiDAR fusion

Check for updates

Received: 7 April 2026
Revised: 14 June 2026
Accepted: 1 July 2026
Published: 2 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.

1. Introduction

The advancement of perception technologies, particularly three-dimensional (3D) object detection, has laid an important foundation for the development of autonomous driving. Reliable 3D detection is crucial for safe navigation in complex environments [1,2], while vehicle-mounted sensors [3] (such as LiDAR and cameras) provide essential hardware support for accurate object recognition, localization, and environmental perception.

Traditional 3D object detection typically relies on single-modality sensors. Cameras can provide rich semantic information but struggle to capture accurate depth information [4–7]; LiDAR, on the other hand, can provide precise geometric structure information but offers

relatively limited semantic information [8–11]. Therefore, multimodal fusion methods can effectively combine information from both sensors, achieving complementary advantages.

Existing multimodal 3D object detection methods can be categorized into projection-based approaches [12–31] and non-projection-based approaches [32–43]. Projection-based methods fuse point cloud and image features via projection matrices, balancing geometric precision, semantic richness, and real-time performance, and are well-suited for autonomous driving. Despite some issues, such as projection errors and limited inter-modal alignment, they continue to provide efficient and reliable solutions. In contrast, non-projection-based methods (e.g., cross-attention and unified feature space fusion) usually achieve stronger detection performance, but at the cost of high computational overhead. Although lightweight variants have been developed to reduce complexity, they often sacrifice either multimodal alignment accuracy or multi-scale representation capability, which limits their robustness in complex driving scenarios. Therefore, this paper focuses on improving projection-based methods to achieve more efficient and accurate 3D object detection.

In recent years, projection-based methods have made remarkable progress in multi-modal representation learning, leading to notable improvements in 3D object detection performance. However, challenges still persist, particularly in detecting pedestrians and cyclists, which are critical for ensuring the safety of autonomous driving systems in dynamic real-world environments [20]. Pedestrians and cyclists present unique detection difficulties due to their small size, dynamic movement, and frequent occlusion in complex and cluttered environments [44].

To address these challenges, we propose a novel edge-texture-aware semantic dual-query fusion (ETA-SDQF) method, which effectively leverages edge and texture information from images to improve small object detection performance. On the KITTI dataset, the proposed method demonstrates improved performance on small object detection tasks such as pedestrians and cyclists.

The main contributions of this paper can be summarized as follows:

- We propose an edge-texture-aware image backbone (ETAIB) to better preserve edge and texture information. ETAIB applies DWT-based decomposition and reconstruction during feature extraction, enabling feature downsampling while retaining fine-grained structural details.
- We introduce a dual-query-guided attention fusion (DQGAF) module to enhance cross-modal alignment between image and LiDAR features. Based on a deformable attention mechanism, the module adaptively aggregates voxel-corresponding image features under the joint guidance of semantic and edge-texture information, thereby improving the stability of 3D detection.

2. Related Work

Multimodal 3D object detection methods can be categorized into projection-based and non-projection-based approaches according to their feature alignment strategies.

2.1. Projection-Based Methods

Projection-based methods perform multimodal fusion through geometric projection and can be further divided into four representative categories: point-projection-based methods, feature-projection-based methods, auto-projection-based methods, and decision-projection-based methods.

Point-projection-based methods [12–20] project raw images or image features onto raw point clouds, enriching each point with complementary semantic information such as RGB values [12], segmentation scores [13–15] or CNN-derived features [16–20]. Representative approaches, including PointPainting [13], FusionPainting [14], and MVP [16],

demonstrate that explicit point-level alignment can enhance point representations and sampling strategies.

Feature-projection-based methods [21–23] use calibration matrices to map voxel coordinates into the image plane during feature fusion, thereby aligning point cloud features with image space.

Auto-projection-based methods [24–26] improve robustness by learning projection offsets or leveraging local neighborhood relationships to compensate for inevitable calibration and projection errors, as exemplified by GraphAlign [24], AutoAlignV2 [25], and HMFI [26].

Decision-projection-based methods [27–31] perform fusion at higher semantic levels, such as Regions of Interest or final detection outputs, e.g., Graph-RCNN [28], F-PointNet [29], MV3D [30], and AVOD [31]. While computationally efficient, limited cross-modal interaction may constrain detection accuracy.

2.2. Non-Projection-Based Methods

Non-projection-based methods aim to mitigate misalignment issues by avoiding explicit geometric projection during fusion. They typically follow either query-learning-based [32–35] or unified-feature-based paradigms [36–43]. Query-learning-based approaches employ cross-attention mechanisms to implicitly model multimodal interactions without explicit spatial correspondence, enabling global feature aggregation, as demonstrated by TransFusion [32] and DeepInteraction [34]. Unified-feature-based approaches, such as UVTR [36], BEVFusion [38], CMT [40], FUTR3D [42], and VirConv [43], learn shared representation spaces across modalities, often through BEV transformation or transformer-based tokenization.

Overall, projection-based methods rely on spatial correspondences that preserve geometric consistency and enable efficient cross-modal fusion. In contrast, non-projection-based methods use more flexible feature interactions or unified feature spaces, and their modeling complexity can vary with different fusion designs.

2.3. Motivation for Our Method

Although existing multimodal 3D object detection methods have achieved notable progress, many projection-based approaches still underutilize fine-grained texture information and struggle to maintain robust cross-modal alignment in complex scenarios. To address these issues while preserving the advantages of projection-based fusion, we propose the ETA-SDQF framework. The framework consists of an edge-texture-aware image backbone (ETAIB), a dual-query-guided attention fusion module (DQGAF), and a hybrid 3D feature learning framework. Specifically, ETAIB employs discrete wavelet transform (DWT) to preserve fine-grained structural and texture details; the proposed fusion module enables adaptive and spatially consistent fusion between image and point cloud features; furthermore, inspired by PV-RCNN [10], the hybrid 3D feature learning framework further enhances the discriminability of point cloud features enriched with image semantics. Overall, these components collaboratively improve projection-based multimodal fusion and achieve more accurate and robust 3D object detection on challenging benchmarks.

3. Proposed ETA-SDQF Framework

3.1. Overview

We introduce the ETA-SDQF framework, an edge-texture-aware projection-based multimodal 3D object detection framework that enhances edge-texture utilization, thereby achieving more robust cross-modal alignment. As illustrated in Figure 1, the proposed framework in-

tegrates an edge-texture-enhanced image backbone, a dual-query-guided adaptive fusion module, and a hybrid 3D feature learning framework into a unified detection pipeline.

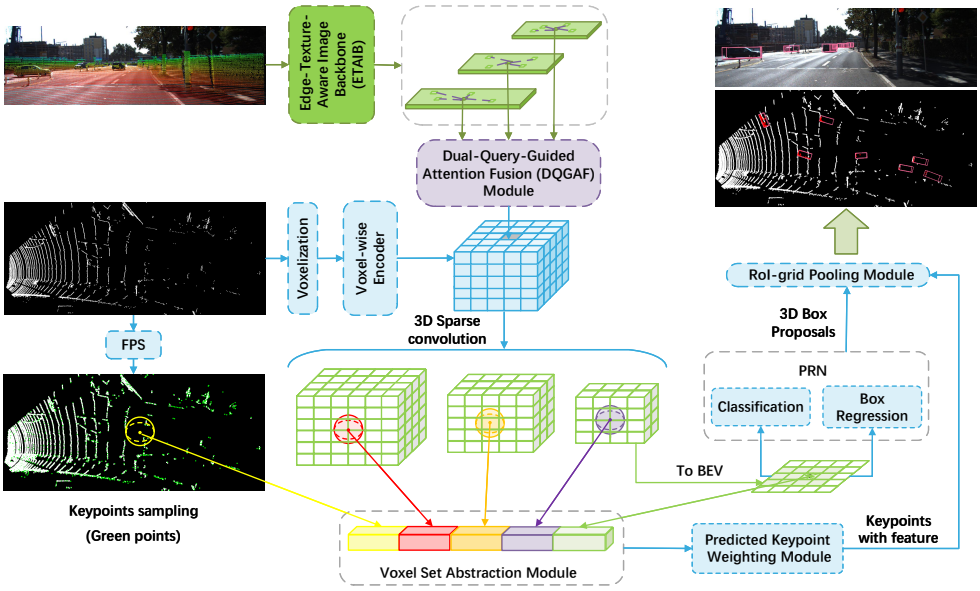


Figure 1. Overview of the ETA-SDQF framework. The image stream extracts multi-scale features, enriched with edge-texture details, using the ETAIB. The point cloud stream is voxelized and encoded, followed by the fusion of image semantic and point cloud voxel features through the DQGAF module, ensuring both adaptive and spatially consistent integration. The hybrid 3D feature learning framework processes the fused voxel features through a 3D sparse convolution-based encoder to extract semantic information and generate 3D object proposals. These features are then condensed into key points and mapped to RoI grid points for precise proposal refinement and confidence prediction.

Given synchronized RGB images and LiDAR point clouds, the image stream is processed by the edge-texture-enhanced image backbone. It replaces the conventional downsampling operations in ResNet [45] with the DWT, enabling the preservation of high-frequency edge and texture information across multi-scale feature maps (see Figure 2). These enhanced representations provide discriminative structural cues that facilitate the detection of small, thin, and partially occluded objects, such as pedestrians and cyclists.

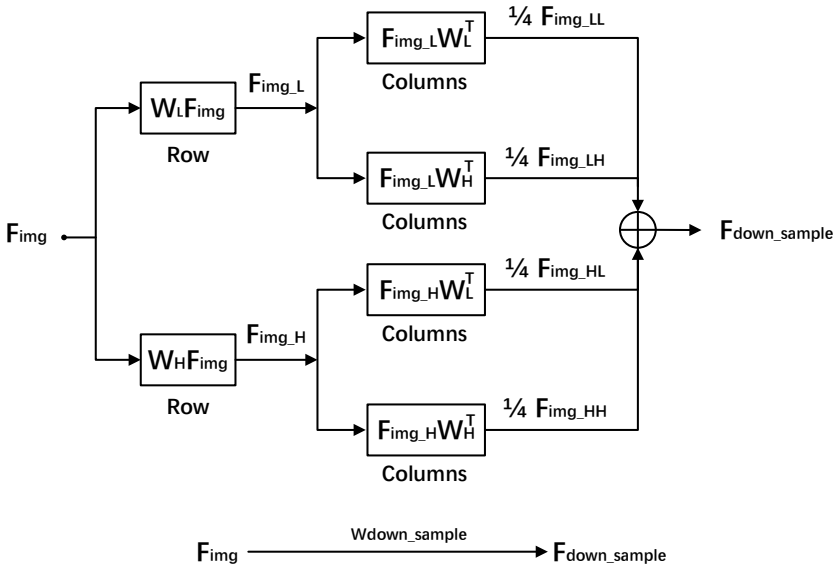


Figure 2. Schematic illustration of DWT-based downsampling.

Meanwhile, the point cloud stream undergoes voxelization and is processed through a voxel-wise encoder. To achieve more precise voxel-level multimodal alignment, we introduce the dual-query guided adaptive fusion module. This module first applies the DWT followed by its inverse transform (IDWT) to extract high-frequency edge-texture features from the enhanced image features, while simultaneously transforming the image features into semantic representations. It then projects voxel centers onto the image plane using calibrated camera-LiDAR transformation matrices. At the corresponding projected locations, multi-scale semantic and edge-texture features are sampled to construct voxel-projected image semantic representations and voxel-projected image edge-texture representations.

After obtaining the voxel-projected image semantic and edge-texture representations, the fusion module generates semantic and edge-texture query embeddings and employs a deformable-attention-based dual-branch architecture to separately aggregate multi-scale image semantic features. The semantic branch preserves global contextual consistency, while the edge-texture branch focuses on fine-grained structural alignment. This design is intended to improve feature alignment under potential projection uncertainty, illumination variation, and partial occlusion.

To further enhance 3D representation learning, inspired by PV-RCNN, the proposed framework introduces a hybrid 3D feature learning strategy that combines voxel-based feature learning with PointNet-style set abstraction for processing the fused voxel features. First, multi-modal fused voxel features are encoded by a 3D sparse convolutional network to extract multi-scale semantic representations and generate initial proposals. Then, a voxel set abstraction module aggregates voxel features into representative keypoints, which are further aggregated onto RoI grid points for proposal-based feature learning. This enables more refined proposal refinement and improves the reliability of confidence estimation.

3.2. Edge-Texture-Aware Image Backbone

To enhance edge and texture information in multi-scale image features and provide more effective guidance for aligning voxel features with image features, we propose an edge-texture-aware image backbone (ETAIB). Built upon ResNet, the proposed method replaces the conventional downsampling operation with DWT, thereby better preserving edge and texture information across different-scale features during feature extraction.

The two-dimensional discrete wavelet transform (2D DWT) extends the one-dimensional DWT to image signals by applying wavelet filtering along both horizontal and vertical directions. Let $\boldsymbol{W}_L$ and $\boldsymbol{W}_H$ denote the low-pass and high-pass wavelet filtering matrices, respectively. Given an input feature map $\boldsymbol{X}$, the 2D DWT decomposes it into four frequency subbands:

$$\begin{aligned} \boldsymbol{X}_{LL} &= \boldsymbol{W}_L \boldsymbol{X} \boldsymbol{W}_L^T, \quad \boldsymbol{X}_{LH} = \boldsymbol{W}_L \boldsymbol{X} \boldsymbol{W}_H^T, \\ \boldsymbol{X}_{HL} &= \boldsymbol{W}_H \boldsymbol{X} \boldsymbol{W}_L^T, \quad \boldsymbol{X}_{HH} = \boldsymbol{W}_H \boldsymbol{X} \boldsymbol{W}_H^T, \end{aligned} \tag{1}$$

where $\boldsymbol{X}_{LL}$ captures low-frequency semantic information, while $\boldsymbol{X}_{LH}$, $\boldsymbol{X}_{HL}$, and $\boldsymbol{X}_{HH}$ represent horizontal, vertical, and diagonal high-frequency details, respectively.

The original feature map can be reconstructed using the IDWT:

$$\begin{aligned} \boldsymbol{X} = {} & \boldsymbol{W}_L^T \boldsymbol{X}_{LL} \boldsymbol{W}_L + \boldsymbol{W}_H^T \boldsymbol{X}_{LH} \boldsymbol{W}_L \\ & + \boldsymbol{W}_L^T \boldsymbol{X}_{HL} \boldsymbol{W}_H + \boldsymbol{W}_H^T \boldsymbol{X}_{HH} \boldsymbol{W}_H. \end{aligned} \tag{2}$$

The proposed image backbone replaces the standard downsampling operation in ResNet with

$$\boldsymbol{F}_{\text{down_sample}} = \boldsymbol{W}_{\text{down_sample}}(\boldsymbol{F}_{\text{img}}), \tag{3}$$

where $\boldsymbol{W}_{\text{down_sample}}(\cdot)$ denotes the DWT-based downsampling operator. As illustrated in Figure 2, the input feature map is decomposed into four subbands: $\boldsymbol{F}_{\text{img_LL}}$, $\boldsymbol{F}_{\text{img_LH}}$, $\boldsymbol{F}_{\text{img_HL}}$, and $\boldsymbol{F}_{\text{img_HH}}$. These subbands are fused via element-wise averaging to achieve a two-fold reduction in spatial resolution while preserving both low-frequency semantic information and high-frequency edge-texture cues. Consequently, the backbone produces structurally richer image representations that facilitate more accurate multimodal feature fusion and improve the perception of small or partially occluded objects.

3.3. Dual-Query-Guided Attention Fusion (DQGAF)

As illustrated in Figure 3, we propose a DQGAF module for precise alignment between LiDAR point clouds and camera images. Edge-texture-enhanced image features are processed with DWT and IDWT to extract high-frequency edge-texture information (see Figure 4), and are further transformed into semantic features. Meanwhile, voxel centers are projected onto the image plane to sample multi-scale semantic and edge-texture features, thereby generating semantic and edge-texture query embeddings.

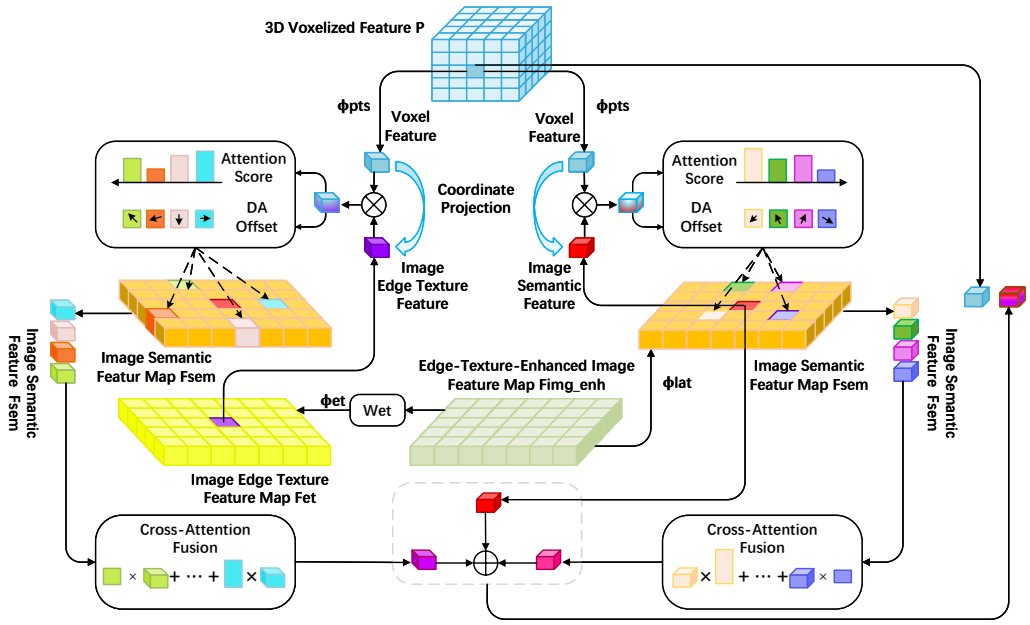


Figure 3. Illustration of the DQGAF module. The module first extracts semantic and edge-texture features from the edge-texture-enhanced image features. It then projects the voxel centers onto the image plane and samples multi-scale semantic and edge-texture features at the projected locations, thereby obtaining voxel-projected representations for both semantic and edge-texture features. The module generates semantic and edge-texture query embeddings, applying learnable convolutional offsets to guide the aggregation of image semantic features. The final fused feature combines the aggregated image semantic features (guided by the semantic and edge-texture query embeddings), the edge-texture-enhanced image features, and the original voxel features.

In the fusion module, semantic and edge-texture query embeddings are used in two separate branches to guide the aggregation of image semantic features, enabling adaptive and spatially consistent multimodal fusion. This module is designed to improve feature alignment under potential projection uncertainty, illumination variation, and partial occlusion.

The proposed fusion module can be divided into the following steps:

(1) Extraction of Semantic and Edge-Texture Features: Given the multi-scale edge-texture-enhanced image features $\boldsymbol{F}^{(l)}_{\text{img_enh}} \in \mathbb{R}^{1\times C\times H_l\times W_l}$ extracted from the ETAIB, the semantic and edge-texture features are extracted separately:

$$\boldsymbol{F}_{\text{sem}}^{(l)} = \phi_{\text{lat}}^{(l)}\left(\boldsymbol{F}_{\text{img_enh}}^{(l)}\right), \quad \boldsymbol{F}_{\text{et}}^{(l)} = \phi_{\text{et}}^{(l)}\left(\boldsymbol{W}_{\text{et}}\left(\boldsymbol{F}_{\text{img_enh}}^{(l)}\right)\right), \tag{4}$$

where $\boldsymbol{W}_{\text{et}}(\cdot)$ denotes the edge-texture feature extraction process, which consists of the 2D discrete wavelet transform (2D DWT) and its inverse transform (2D IDWT). Given the multi-scale edge-texture-enhanced image features $\boldsymbol{F}_{\text{img_enh}}^{(l)} \in \mathbb{R}^{1\times C\times H_l\times W_l}$, the 2D DWT is first applied to decompose them into four sub-band components: $\boldsymbol{F}_{\text{img_enh_LL}}^{(l)}$, $\boldsymbol{F}_{\text{img_enh_LH}}^{(l)}$, $\boldsymbol{F}_{\text{img_enh_HL}}^{(l)}$, and $\boldsymbol{F}_{\text{img_enh_HH}}^{(l)}$. Subsequently, only the horizontal and vertical high-frequency components, i.e., $\boldsymbol{F}_{\text{img_enh_LH}}^{(l)}$ and $\boldsymbol{F}_{\text{img_enh_HL}}^{(l)}$, are retained, while the low-frequency and diagonal components are discarded. The retained components are then reconstructed via 2D IDWT (see Figure 4), thereby enhancing the representation of object edges and textures.

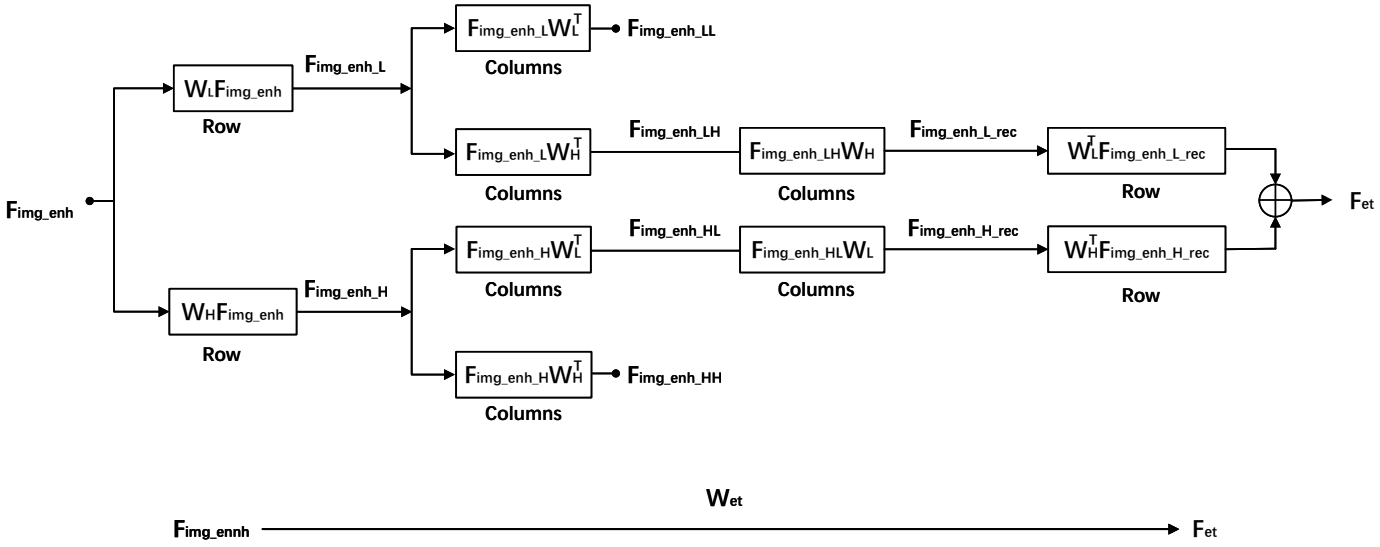


Figure 4. Schematic illustration of edge-texture extraction.

The operators $\phi_{\text{lat}}^{(l)} : \mathbb{R}^C \to \mathbb{R}^{C_{\text{mid}}}$ and $\phi_{\text{et}}^{(l)} : \mathbb{R}^C \to \mathbb{R}^{C_{\text{mid}}}$ are learnable projection functions that map the multi-scale edge-texture-enhanced image features $\boldsymbol{F}_{\text{img_enh}}^{(l)}$ and their corresponding multi-scale edge-texture features $\boldsymbol{W}_{\text{et}}(\boldsymbol{F}_{\text{img_enh}}^{(l)})$ into a unified latent space of dimensionality C_{mid}, where $l = 1, \ldots, L$ indexes the feature pyramid levels.

This process results in the multi-scale image semantic and edge-texture features $\boldsymbol{F}_{\text{sem}}^{(l)} \in \mathbb{R}^{1\times C_{\text{mid}}\times H_l\times W_l}$ and $\boldsymbol{F}_{\text{et}}^{(l)} \in \mathbb{R}^{1\times C_{\text{mid}}\times H_l\times W_l}$. The reconstructed multi-scale edge-texture feature $\boldsymbol{F}_{\text{et}}^{(l)}$ thus emphasizes intensity transitions and object boundaries, providing a structural complement to the semantic features.

Notably, although no normalization or regularization is applied during the extraction of high-frequency sub-bands (LH, HL, HH), Batch Normalization in the network, together with attention-based adaptive reweighting of high-frequency responses, helps maintain stable gradients during training. This design avoids gradient instability or explosion, preserves edge and texture information, and shows good robustness and adaptability in sparse and small-object scenarios.

(2) Voxel-Aligned Feature Sampling: Given the multi-scale image semantic and edge-texture features $\boldsymbol{F}_{\text{sem}}^{(l)} \in \mathbb{R}^{1\times C_{\text{mid}}\times H_l\times W_l}$ and $\boldsymbol{F}_{\text{et}}^{(l)} \in \mathbb{R}^{1\times C_{\text{mid}}\times H_l\times W_l}$, and the set of non-empty voxel features $\boldsymbol{P} \in \mathbb{R}^{N\times d}$, each voxel center $\boldsymbol{V}_j = (v_x^j, v_y^j, v_z^j)^\top$, for $j = 1, \ldots, N$, is projected onto the image plane using the camera projection matrix $\boldsymbol{T}_{\text{cam}\leftarrow\text{lidar}}$ together with the rectifying rotation and camera calibration matrix $\boldsymbol{RC}$.

The 2D reference coordinates for the j-th voxel are computed as

$$\begin{bmatrix} g_x^j \\ g_y^j \end{bmatrix} = \pi(\boldsymbol{V}_j), \tag{5}$$

where $\pi(\cdot)$ denotes the projection from 3D LiDAR coordinates to 2D image coordinates. Specifically, the process involves the following steps:

The voxel center $\mathbf{V}_j = [X_j, Y_j, Z_j, 1]^\top$ is transformed from the LiDAR coordinate system to the camera coordinate system using the extrinsic matrix:

$$\mathbf{V}_j^{\text{cam}} = \mathbf{T}_{\text{cam}\leftarrow\text{lidar}}\mathbf{V}_j. \tag{6}$$

It is then projected onto the image plane using the camera intrinsic matrix $\mathbf{K}$:

$$\mathbf{p}_j = \mathbf{K}\mathbf{V}_j^{\text{cam}}. \tag{7}$$

The 2D pixel coordinates are obtained via perspective division:

$$g_x^j = \frac{x_j}{z_j}, \quad g_y^j = \frac{y_j}{z_j}. \tag{8}$$

Points falling outside the image boundary are filtered or masked to avoid invalid sampling.

After obtaining the reference points $\boldsymbol{G}^{\text{ref}} \in \mathbb{R}^{1\times N\times 1\times 2}$, the corresponding semantic and edge-texture features are bilinearly sampled as

$$F_{\text{sem_ref},j,c} = \sum_{l=1}^{L}\sum_{h=1}^{H_l}\sum_{w=1}^{W_l} F_{\text{sem},c,h,w}^{(l)} \cdot k_h(g_y^j) \cdot k_w(g_x^j), \tag{9}$$

$$F_{\text{et_ref},j,c} = \sum_{l=1}^{L}\sum_{h=1}^{H_l}\sum_{w=1}^{W_l} F_{\text{et},c,h,w}^{(l)} \cdot k_h(g_y^j) \cdot k_w(g_x^j), \tag{10}$$

where $k_h(\cdot)$ and $k_w(\cdot)$ denote the bilinear interpolation kernels along the vertical and horizontal directions, respectively. Here, $j = 1, \ldots, N$ denotes the voxel index, corresponding to the j-th voxel, and $c = 1, \ldots, C_{\text{mid}}$ indexes the feature channels.

The resulting voxel-aligned semantic and edge-texture features are denoted as

$$\boldsymbol{F}_{\text{sem_ref}} \in \mathbb{R}^{1\times N\times C_{\text{mid}}} \quad \text{and} \quad \boldsymbol{F}_{\text{et_ref}} \in \mathbb{R}^{1\times N\times C_{\text{mid}}}.$$

(3) Query Formation: To facilitate effective cross-modal interaction between the image and LiDAR modalities, voxel-specific query features are formed by integrating voxel features with the corresponding image-aligned semantic and edge-texture representations.

Given the bilinearly sampled semantic and edge-texture features $\boldsymbol{F}_{\text{sem_ref}} \in \mathbb{R}^{1\times N\times C_{\text{mid}}}$, $\boldsymbol{F}_{\text{et_ref}} \in \mathbb{R}^{1\times N\times C_{\text{mid}}}$, and the voxel feature set $\boldsymbol{P} \in \mathbb{R}^{N\times d}$, each voxel feature $\boldsymbol{P}_j$ is first linearly projected into the same latent space as the image features through a learnable mapping:

$$\tilde{\boldsymbol{P}}_j = \phi_{\text{pts}}(\boldsymbol{P}_j) = \boldsymbol{W}_{\text{pts}}\boldsymbol{P}_j + \boldsymbol{b}_{\text{pts}}, \quad \tilde{\boldsymbol{P}}_j \in \mathbb{R}^{C_{\text{mid}}}, \tag{11}$$

This projection ensures that the dimensionality of the LiDAR features aligns with that of the image feature embeddings, facilitating compatibility for subsequent fusion.

Following voxel-to-image alignment, the semantic and edge-texture query features are constructed through cross-modal feature interaction. Specifically, for the j-th voxel, its semantic query feature $\boldsymbol{Q}_j^{\text{sem}}$ and edge-texture-aware query feature $\boldsymbol{Q}_j^{\text{et}}$ are computed as

$$\boldsymbol{Q}_j^{\text{sem}} = \boldsymbol{F}_{\text{sem_ref},j} \odot \tilde{\boldsymbol{P}}_j, \quad \boldsymbol{Q}_j^{\text{et}} = \boldsymbol{F}_{\text{et_ref},j} \odot \tilde{\boldsymbol{P}}_j, \tag{12}$$

where $\odot$ denotes the Hadamard product, i.e., element-wise multiplication between tensors of identical shape. Here, $\boldsymbol{F}_{\text{sem_ref},j}$ and $\boldsymbol{F}_{\text{et_ref},j}$ represent the voxel-aligned semantic and edge-texture features for the j-th voxel, respectively, obtained by aggregating multi-scale image features. The normalized voxel feature $\tilde{\boldsymbol{P}}_j$ serves as a voxel-adaptive weighting term, ensuring that the resulting query features $\boldsymbol{Q}_j^{\text{sem}}$ and $\boldsymbol{Q}_j^{\text{et}}$ encode joint semantic-geometric and edge-texture-aware information for cross-modal fusion.

The resulting semantic query set $\boldsymbol{Q}^{\text{sem}} = [\boldsymbol{Q}_1^{\text{sem}}, \ldots, \boldsymbol{Q}_N^{\text{sem}}]$ and edge-texture-aware query set $\boldsymbol{Q}^{\text{et}} = [\boldsymbol{Q}_1^{\text{et}}, \ldots, \boldsymbol{Q}_N^{\text{et}}]$ are then fed into the dual-query deformable attention module, which predicts sampling offsets and attention weights for adaptive multi-scale feature aggregation.

In summary, each semantic query embedding $\boldsymbol{Q}_j^{\text{sem}}$ encodes a joint semantic-geometric representation of the voxel, while the edge-texture query embedding $\boldsymbol{Q}_j^{\text{et}}$ enhances sensitivity to fine-grained edge and texture structures. Together, they enable more accurate voxel-to-image correspondence during multimodal fusion.

(4) Deformable attention-based dual-branch fusion for multimodal feature aggregation: After constructing the semantic query embeddings $\boldsymbol{Q}_j^{\text{sem}}$ and edge-texture query embeddings $\boldsymbol{Q}_j^{\text{et}}$, the deformable attention mechanism [46] predicts sampling offsets and attention weights for each query. Specifically,

$$\boldsymbol{\Delta}_{Sj} = f_\theta^S\left(\boldsymbol{Q}_j^{\text{sem}}\right) \in \mathbb{R}^{H\times L\times P\times 2}, \tag{13}$$

$$\boldsymbol{W}_{Sj} = g_\theta^S\left(\boldsymbol{Q}_j^{\text{sem}}\right) \in \mathbb{R}^{H\times L\times P}, \tag{14}$$

$$\boldsymbol{\Delta}_{Ej} = f_\theta^E\left(\boldsymbol{Q}_j^{\text{et}}\right) \in \mathbb{R}^{H\times L\times P\times 2}, \tag{15}$$

$$\boldsymbol{W}_{Ej} = g_\theta^E\left(\boldsymbol{Q}_j^{\text{et}}\right) \in \mathbb{R}^{H\times L\times P}. \tag{16}$$

Here, f_θ^S and f_θ^E denote the learnable offset prediction functions for the semantic and edge-texture branches, respectively, while g_θ^S and g_θ^E are the corresponding attention weight prediction functions. The parameters H, L, and P represent the number of attention heads, feature levels, and sampling points per attention head.

The sampling locations for each voxel are computed using the predicted offsets normalized by the image resolution:

$$\boldsymbol{L}_{Sj} = \boldsymbol{G}_j + \frac{\boldsymbol{\Delta}_{Sj}}{\Delta_{\text{No}}}, \qquad \boldsymbol{L}_{Ej} = \boldsymbol{G}_j + \frac{\boldsymbol{\Delta}_{Ej}}{\Delta_{\text{No}}} \tag{17}$$

Here, $\boldsymbol{G}_j$ denotes the reference projection coordinate of the j-th voxel on the image plane. $\boldsymbol{L}_{Sj}$ and $\boldsymbol{L}_{Ej}$ represent the semantic-guided and edge-texture-guided sampling locations, respectively. $\boldsymbol{\Delta}_{Sj}$ and $\boldsymbol{\Delta}_{Ej}$ denote the learnable sampling offsets predicted from the semantic and edge-texture query embeddings, while Δ_{No} is the offset normalizer used to scale the offsets according to the spatial resolution of the image feature maps.

Once the sampling locations and attention weights are obtained, multi-scale deformable attention aggregates image semantic features from both semantic-guided and edge-texture-guided spatial neighborhoods:

$$\boldsymbol{O}_j = \text{MSDeformAttn}\left(\boldsymbol{F}_{\text{sem}}, \boldsymbol{L}_{Sj}, \boldsymbol{W}_{Sj}\right) + \text{MSDeformAttn}\left(\boldsymbol{F}_{\text{sem}}, \boldsymbol{L}_{Ej}, \boldsymbol{W}_{Ej}\right) + \boldsymbol{F}_{\text{sem_ref},j}, \tag{18}$$

Here, $\boldsymbol{F}_{\text{sem}}$ denotes the image semantic features obtained by summing the features across all scales, while $\boldsymbol{F}_{\text{sem_ref},j}$ represents the voxel-aligned semantic feature corresponding to voxel j.

To effectively integrate multimodal information from both image features and voxel features, the final aggregated image feature $\boldsymbol{O}_j$ and the voxel feature $\boldsymbol{P}_j$ are first projected into a shared latent embedding space via separate linear transformation layers. The projected features are then concatenated along the channel dimension to form the fused representation, which is subsequently passed through a non-linear activation function to enhance feature expressiveness.

$$\boldsymbol{F}_{\text{fuse}} = \sigma\Big(\big[T_{\text{img}}(\boldsymbol{O}_j) \parallel T_{\text{pts}}(\boldsymbol{P}_j) \big] \Big) \tag{19}$$

Here, $\parallel$ denotes channel-wise concatenation, while T_{img} and T_{pts} are learnable linear projection functions that map image and LiDAR features into a shared latent space for downstream 3D detection.

3.4. Hybrid 3D Feature Learning Framework

To further enhance 3D representation learning, inspired by PV-RCNN, we propose a hybrid feature learning strategy that combines voxel-based feature learning with PointNet-style set abstraction for processing the fused voxel features. First, the multi-modal fused voxel features are encoded by a 3D sparse convolutional network to extract multi-scale semantic features and generate initial proposals. Then, a voxel set abstraction module aggregates voxel features into representative keypoints.

In particular, the keypoints are selected using 3D Farthest Point Sampling (FPS) to ensure a uniform spatial distribution, which is crucial for RoI-grid pooling and proposal refinement. FPS in physical space achieves a good balance between computational efficiency and spatial coverage, while maintaining stable coverage even in sparse regions, thereby improving the robustness of detection performance. Finally, the keypoint features are further aggregated onto RoI grid points to enhance proposal-level feature learning, enabling more refined bounding box refinement and improving the reliability of confidence estimation.

Notably, while set abstraction preserves geometry, its pooling may attenuate image-derived fine-grained texture cues for sparse small objects. The keypoint features are subsequently aggregated onto RoI grid points to form proposal-aware representations for bounding box refinement and confidence estimation.

4. Experimental Results

In this section, we first describe the experimental setup and introduce the KITTI dataset. We then compare the proposed multimodal fusion framework with several representative multimodal fusion approaches to evaluate its effectiveness. Finally, we conduct a series of ablation studies to analyze the contributions of the edge-texture-enhanced image backbone, the dual-query guided adaptive fusion module, and the hybrid 3D feature learning framework to overall performance.

4.1. Implementation Details and Strategies

(1) Data Processing: All experiments are conducted on the KITTI 3D Object Detection benchmark, a widely adopted dataset for autonomous driving research. The dataset was collected in Karlsruhe, Germany, and covers diverse driving environments, including urban, rural, and highway scenarios. It provides annotations for three object categories—Car, Pedestrian, and Cyclist—each further divided into Easy (E), Moderate (M), and Hard (H) levels according to object size, occlusion, and truncation.

The benchmark consists of 7481 training images and 7518 test images, each paired with corresponding point cloud data, totaling 80,256 annotated objects. Following the standard protocol, the training set is split into 3712 samples for training and 3769 samples for validation.

Average Precision (AP) is adopted as the primary evaluation metric for 3D object detection. It measures the area under the precision–recall curve by averaging interpolated precision values over a set of predefined recall levels.

The original KITTI benchmark employs an 11-point interpolation scheme, which samples recall at coarse intervals. However, this discretization has been shown to introduce evaluation bias due to insufficient sampling resolution. To address this issue, a refined evaluation protocol with denser recall sampling has been proposed in recent literature.

Specifically, we adopt the improved AP metric with 40 recall positions (AP_{40}), defined as

$$AP_{40} = \frac{1}{40} \sum_{r \in \{1/40, 2/40, \ldots, 1\}} \rho_{\text{interp}}(r), \tag{20}$$

where the interpolated precision is formulated as

$$\rho_{\text{interp}}(r) = \max_{\tilde{r} \geq r} \rho(\tilde{r}), \tag{21}$$

Here $\rho(r)$ denotes the precision at recall r. In addition to AP, the KITTI benchmark reports 3D Average Precision (3D AP), Bird's-Eye-View Average Precision (BEV AP), and Average Orientation Similarity (AOS). Both 3D AP and BEV AP are computed based on the Intersection-over-Union (IoU) between predicted and ground-truth bounding boxes, while AOS further evaluates orientation consistency for correctly matched detections.

Following the official KITTI evaluation protocol, IoU thresholds are set to 0.7 for the Car category and 0.5 for both Pedestrian and Cyclist categories.

To further assess the localization robustness and fine-grained detection capability of the proposed method under varying matching criteria, additional IoU sensitivity ablation experiments are conducted. Specifically, the IoU threshold for the Car category is varied from 0.4 to 0.7, whereas the thresholds for the Pedestrian and Cyclist categories are varied from 0.4 to 0.8.

(2) Network Configuration and Training: The proposed model is implemented using the MMDetection3D framework developed by OpenMMLab. The DynamicMVXFaster-RCNN architecture is employed with a voxel size of (0.05, 0.05, 0.1) and a predefined point cloud range of (0, 70.4), $(-40, 40)$, and $(-3, 1)$ meters along the x, y, and z axes, respectively. The overall framework consists of an image branch, a LiDAR branch, and a multimodal fusion module.

The image branch uses a ResNet-50 backbone enhanced with wavelet-based feature refinement (i.e., ETAIB), followed by an FPN that outputs five feature maps, each with 256 channels. To ensure stable multimodal optimization, the early convolutional layers of the backbone are partially frozen.

The LiDAR branch employs DynamicVFE to encode voxel-level geometric features. The fusion module integrates multi-scale image features into voxel representations using a deformable-attention-based dual-query fusion module combined with channel-wise modulation, producing 128-dimensional fused voxel features. These fused features are subsequently processed by a Sparse Encoder and SECOND backbone, followed by a SECONDFPN neck for multi-scale feature aggregation.

Class-specific anchor settings are used following KITTI conventions. For cars, the anchor range is defined as (0, 70.4), $(-40.0, 40.0)$, and $(-1.78, -1.78)$, while for pedestrians and cyclists, the range is (0, 70.4), $(-40.0, 40.0)$, and $(-0.6, -0.6)$. The anchor dimensions (w, l, h) are (1.6, 3.9, 1.56) for cars, (0.6, 0.8, 1.73) for pedestrians, and (0.6, 1.76, 1.73) for cyclists. IoU thresholds for positive/negative sample assignment are 0.6/0.45 for cars and 0.5/0.35 for pedestrians and cyclists. Training is performed using the AdamW optimizer with an initial learning rate of 5×10^{-4}. A cosine annealing schedule with warm-up is adopted

over 40 epochs: the learning rate increases to 5×10^{-3} during the first 15 epochs and then decays to 5×10^{-8}. Momentum is annealed from 0.85 to 1.0 following a cosine schedule. The model is trained on two NVIDIA RTX 4090 GPUs with a batch size of 2 per GPU.

(3) Data Preparation and Augmentation: Both LiDAR and RGB data from the KITTI dataset are used in our experiments. Each LiDAR point is represented as a 4D vector $(x, y, z, \text{intensity})$. For the image modality, we apply multi-scale resizing, Where the input resolution is randomly sampled from 640×192 to 2560×768, while preserving the original aspect ratio. This allows the model to better handle objects at different scales.

To improve generalization, we apply several geometric augmentations to the LiDAR data during training. Specifically, we perform random flipping along the x-axis (BEV flip), rotation around the z-axis with angles uniformly sampled from $[-\pi/4, \pi/4]$, global scaling within $[0.95, 1.05]$, and random translation with a standard deviation of 0.2 along each axis.

Importantly, these transformations are applied in a synchronized cross-modal manner. Although the augmentations are defined in 3D space, we consistently update the corresponding image annotations and LiDAR-to-image projection relationships to preserve geometric alignment between the two modalities. This consistency is essential for projection-based multimodal fusion methods, where accurate spatial correspondence is critical for effective feature learning.

For the image branch, we adopt a relatively simple augmentation strategy. Only multi-scale resizing and horizontal flipping are used, and both are strictly aligned with the transformations applied to LiDAR data.

Points and bounding boxes outside the predefined range are removed before training, and the remaining points are shuffled to improve data randomness and stabilize optimization.

During validation and testing, no stochastic augmentations are applied. Only deterministic resizing is used to standardize the input resolution, ensuring fair, stable, and reproducible evaluation results.

4.2. Comparison with Baseline Methods

In this section, we evaluate the proposed multimodal fusion framework ETA-SDQF on the KITTI benchmark and provide a fair comparison with representative state-of-the-art 3D object detection methods. Quantitative results in terms of AP_{R40} and Average Orientation Similarity (AOS) are reported in Tables 1 and 2.

Table 1. Comparison of 3D object detection performance between the proposed multimodal fusion framework ETA-SDQF and state-of-the-art networks on the KITTI test set.

Method	Mod	3D Car AP_{R40} @0.5			3D Ped. AP_{R40} @0.5			3D Cyc. AP_{R40} @0.5			FPS (Hz)
		E	M	H	E	M	H	E	M	H	
DSFNet [47]	L	80.65	70.44	65.03	43.19	34.60	31.85	67.80	54.26	48.25	33.33
MSFASA-3DNet [48]	L	86.06	77.21	71.89	42.89	36.32	34.33	75.60	60.30	53.83	33.33
HINTED [49]	L	84.00	74.13	67.03	47.33	37.75	34.10	76.21	63.01	55.85	25.00
IA-SSD (multi) [50]	L	88.34	80.13	75.04	46.51	39.03	35.61	**78.35**	61.94	55.70	71.43
LinkNet3D [51]	L	87.22	78.54	74.36	–	–	–	76.11	60.47	53.77	25.00
L-AUG [52]	L	**90.53**	**83.84**	**79.10**	46.76	39.07	35.74	73.32	59.30	53.74	10.00
PCNet3D++ [53]	L	84.00	74.19	69.65	45.22	37.47	34.95	76.16	58.67	52.02	20.00
Fade3D [54]	L	85.85	75.57	70.47	30.25	24.06	21.84	66.04	50.02	44.39	**100.00**
SeSame-point [55]	L	85.25	76.83	71.60	42.29	35.34	33.02	69.55	54.56	48.34	N/A
EPNet [18]	L&I	89.81	79.28	74.59	–	–	–	–	–	–	10.00
F-PointNet [29]	L&I	82.19	69.79	60.59	**50.53**	42.15	38.08	72.27	56.12	49.01	5.88
MV3D [30]	L&I	74.97	63.63	54.00	–	–	–	–	–	–	2.78
AVOD-FPN [31]	L&I	83.07	71.76	65.73	50.46	**42.27**	39.04	63.76	50.55	44.93	10.00
PL++: PV-RCNN++ [56]	L&S	86.60	75.23	70.34	41.53	33.89	31.42	62.80	48.97	42.80	2.92
PointViT [57]	L&I	89.22	79.97	72.64	–	–	–	–	–	–	20.00
Faraway-Frustum [58]	L&I	87.45	79.05	76.14	46.33	38.58	35.71	77.36	62.00	55.40	10.00
ETA-SDQF	L&I	87.69	80.88	76.41	49.20	42.26	**39.23**	77.65	**63.51**	**57.48**	15.75

Table 2. Comparison of AOS performance between the proposed multimodal fusion framework ETA-SDQF and state-of-the-art networks on the KITTI test set.

Method	Mod	AOS Car AP_{R40} @0.7			AOS Ped. AP_{R40} @0.5			AOS Cyc. AP_{R40} @0.5			FPS (Hz)
		E	M	H	E	M	H	E	M	H	
DSFNet	L	94.53	91.30	87.48	51.90	41.47	38.47	67.80	54.26	48.25	33.33
MSFASA-3DNet	L	95.19	93.84	90.71	42.89	36.32	34.33	75.60	60.30	53.83	33.33
HINTED	L	93.97	89.41	83.95	62.13	47.84	43.71	86.59	**79.73**	**73.13**	25.00
IA-SSD (multi)	L	96.07	93.47	90.51	62.51	49.58	47.17	84.82	70.13	65.13	71.43
LinkNet3D	L	95.43	93.95	91.15	–	–	–	**88.38**	74.52	67.89	25.00
L-AUG	L	95.80	94.76	**91.94**	58.42	47.59	44.64	83.69	73.07	67.72	10.00
PCNet3D++	L	94.93	91.72	88.32	55.89	45.03	42.29	82.16	69.59	63.42	20.00
Fade3D	L	**97.62**	92.23	87.09	30.04	23.24	21.82	73.50	57.98	52.62	**100.00**
SeSame-point	L	95.20	93.32	90.14	56.55	47.09	44.58	83.97	71.88	65.00	N/A
EPNet	L&I	96.13	94.22	89.68	–	–	–	–	–	–	10.00
AVOD-FPN	L&I	94.65	88.61	83.71	53.48	43.99	41.56	69.21	58.70	53.47	10.00
PL++: PV-RCNN++	L&S	94.76	91.42	88.30	53.77	43.90	40.77	73.67	59.07	52.52	2.94
PointViT	L&I	97.04	**96.47**	88.88	–	–	–	–	–	–	20.00
ETA-SDQF	L&I	95.94	94.39	91.73	**65.65**	**54.68**	**52.02**	87.41	76.26	69.98	15.75

Overall, ETA-SDQF achieves competitive and consistent performance across object categories and difficulty levels, demonstrating the effectiveness of the proposed Edge-Texture-Aware Semantic Dual-Query Fusion framework. The comparison covers recent approaches from 2024 to 2026, including Transformer-based methods (e.g., PointViT [57]), BEV-based architectures (e.g., DSFNet [47]), and LiDAR-centric detectors such as L-AUG [52], Fade3D [54], MSFASA-3DNet [48], and PCNet3D++ [53].

For pedestrian detection, ETA-SDQF achieves APs of (49.20%/42.26%/39.23%) under the Easy, Moderate, and Hard settings, respectively, showing performance comparable to or better than recent LiDAR-based methods such as HINTED [49] and IA-SSD(multi) [50]. For cyclists, it achieves APs of (77.65%/63.51%/57.48%), with stronger performance on the Moderate and Hard splits. For vehicles, ETA-SDQF obtains APs of (87.69%/80.88%/76.41%), maintaining stable accuracy across all difficulty levels. These results suggest that the proposed multimodal fusion strategy effectively integrates LiDAR geometric structure with image-based edge and texture cues, which is particularly beneficial for sparse and small-object perception.

For orientation estimation, ETA-SDQF achieves pedestrian AOSs of (65.65%/54.68%/52.02%), cyclist AOSs of (87.41%/76.26%/69.98%), and vehicle AOSs of (95.94%/94.39%/91.73%). These results demonstrate robust orientation prediction across object scales, especially for challenging small and partially occluded instances. Notably, the ETAIB module enhances edge- and texture-aware representations via discrete wavelet transform, producing more discriminative multimodal features that improve cross-modal alignment.

Regarding efficiency, ETA-SDQF runs at 15.75 FPS on two NVIDIA RTX 4090 GPUs. This result is reported under a controlled experimental setting and serves as an efficiency indicator of the proposed framework. Together with consistent improvements under fair evaluation settings, it demonstrates that the proposed method achieves a favorable balance between accuracy and computational cost, supporting its applicability in multi-modal 3D perception tasks, while the actual deployment speed may vary across different hardware platforms.

4.3. Ablation Study

In this section, we conduct ablation studies to analyze the effectiveness of each component in our proposed method. All experiments are performed on the KITTI validation set unless otherwise specified.

(1) **Ablation Studies of Different Modules in ETA-SDQF for 3D, BEV, and AOS Tasks.** The ablation results in Tables 3–5 systematically evaluate the contributions of each component to 3D, BEV, and AOS performance within the ETA-SDQF framework. Overall, the progressive integration of modules enhances multimodal feature representation, confirming the benefits of combining semantic–texture enhancement with geometric aggregation. The edge–texture-aware image backbone provides notable improvements for small-object detection in both 3D and BEV tasks. Specifically, pedestrian AP increases from (58.80%/53.80%/50.08%) to (65.58%/58.30%/53.81%) in 3D and from (66.72%/60.91%/58.10%) to (71.91%/65.70%/61.20%) in BEV across Easy, Moderate, and Hard settings, while cyclist performance also improves consistently. Vehicle detection remains largely unchanged, suggesting that LiDAR already provides strong geometric priors for large rigid objects. Meanwhile, DWT-based feature downsampling in our implementation contributes to reduced GPU memory usage (from 10.69 GB to 3.00 GB) and slightly increased inference speed (from 41.67 FPS to 43.86 FPS), although these improvements may also depend on other factors in the feature-processing pipeline.

Table 3. Comparison of 3D object detection performance between improved and original networks on the KITTI validation set.

Method	Mod	3D Car AP_{R40} @0.7			3D Ped. AP_{R40} @0.5			3D Cyc. AP_{R40} @0.5			FPS (Hz)	Mem (GB)
		E	M	H	E	M	H	E	M	H		
MVX-Net [20]	L&I	88.48	78.64	74.45	58.80	53.80	50.08	73.37	51.77	49.09	41.67	10.69
MVX-Net + ETAIB	L&I	88.95	78.90	74.62	65.58	58.30	53.81	72.08	55.33	52.11	**43.86**	**3.00**
MVX-Net + ETAIB + DQGAF	L&I	90.08	78.85	76.00	66.47	**60.67**	**57.74**	80.00	59.60	56.92	42.00	3.20
PV-RCNN [10]	L	91.74	84.68	**82.66**	**67.41**	59.75	54.57	91.09	72.93	68.54	21.05	5.46
ETA-SDQF	L&I	**91.93**	**84.80**	82.46	66.25	59.82	55.14	**91.48**	**73.45**	**69.14**	15.75	6.99

Table 4. Comparison of BEV object detection performance between improved and original networks on the KITTI validation set.

Method	Mod	BEV Car AP_{R40} @0.7			BEV Ped. AP_{R40} @0.5			BEV Cyc. AP_{R40} @0.5			FPS (Hz)	Mem (GB)
		E	M	H	E	M	H	E	M	H		
MVX-Net	L&I	92.36	87.94	85.80	66.72	60.91	58.10	76.16	54.48	51.65	41.67	10.69
MVX-Net + ETAIB	L&I	94.05	88.24	85.81	71.91	65.70	61.20	72.83	57.30	53.99	**43.86**	**3.00**
MVX-Net + ETAIB + DQGAF	L&I	94.02	88.02	85.61	**73.62**	**68.40**	**64.39**	81.37	61.94	59.06	42.00	3.20
PV-RCNN	L	94.78	90.66	**88.67**	70.56	62.57	57.83	91.99	76.71	72.40	21.05	5.46
ETA-SDQF	L&I	**95.11**	**90.72**	88.56	70.39	63.42	58.83	**93.70**	**77.77**	**73.35**	15.75	6.99

Table 5. Comparison of AOS object detection performance between improved and original networks on the KITTI validation set.

Method	Mod	AOS Car AP_{R40} @0.7			AOS Ped. AP_{R40} @0.5			AOS Cyc. AP_{R40} @0.5			FPS (Hz)	Mem (GB)
		E	M	H	E	M	H	E	M	H		
MVX-Net	L&I	95.97	92.01	89.07	49.68	47.23	44.33	76.15	56.36	53.00	41.67	10.69
MVX-Net + ETAIB	L&I	95.98	91.78	89.05	42.04	40.91	38.38	72.84	56.62	53.90	**43.86**	**3.00**
MVX-Net + ETAIB + DQGAF	L&I	97.86	92.19	89.16	59.93	55.20	52.09	83.53	63.83	62.06	42.00	3.20
PV-RCNN	L	97.96	94.20	93.86	68.05	62.58	58.93	96.50	**84.48**	**81.23**	21.05	5.46
ETA-SDQF	L&I	**98.52**	**94.45**	**93.96**	**70.92**	**66.69**	**62.96**	**97.20**	83.22	80.01	15.75	6.99

However, edge–texture enhancement alone does not immediately translate into improved AOS performance. Pedestrian AOS decreases from (49.68%/47.23%/44.33%) to (42.04%/40.91%/38.38%), indicating that orientation-sensitive prediction can become less stable when texture cues are introduced without sufficiently adaptive cross-modal alignment. Incorporating the dual-query guided fusion module restores orientation estimation, with pedestrian AOS increasing to (59.93%/55.20%/52.09%) and cyclist AOS reaching

(83.53%/63.83%/62.06%). These results suggest that adaptive fusion effectively aligns edge–texture features with voxel representations, stabilizing orientation-aware learning.

The hybrid 3D feature learning framework introduces a trade-off. While pedestrian performance shows a slight decrease relative to the preceding fusion stage, the framework maintains strong results for both pedestrians and cyclists and further improves vehicle detection. Specifically, the full model achieves (91.93%/84.80%/82.46%) vehicle AP in 3D, (95.11%/90.72%/88.56%) in BEV, and (98.52%/94.45%/93.96%) in AOS. Pedestrian performance remains competitive with (66.25%/59.82%/55.14%) in 3D, (70.39%/63.42%/58.83%) in BEV, and (70.92%/66.69%/62.96%) in AOS. Cyclist detection continues to be strong, achieving (91.48%/73.45%/69.14%) in 3D, (93.70%/77.77%/73.35%) in BEV, and (97.20%/83.22%/80.01%) in AOS.

Although the hybrid 3D feature learning framework improves proposal refinement and geometric abstraction, its pooling-based aggregation may partially suppress fine-grained image-derived cues for very sparse objects, such as pedestrians. This may explain the minor decrease in Pedestrian Easy AP from 66.47% to 66.25%, despite the overall benefits of the hybrid 3D feature learning framework for multimodal feature integration.

Finally, a performance gap is observed between the validation and test sets, with Cyclist Moderate AP decreasing from 73.45% to 63.51%. This is primarily attributed to the more challenging KITTI test distribution and class imbalance, and may also reflect slight overfitting induced by validation-based model selection.

(2) **Ablation Studies on Different Wavelets for MVX-Net + ETAIB + DQGAF in 3D, BEV, and AOS Tasks.** The ablation results in Tables 6–8 evaluate the influence of different Wave bases under the MVX-Net + ETAIB + DQGAF framework. Overall, all evaluated variants (Bior3.5, Db4, and Sym6) achieve competitive performance, while the Haar Wave consistently yields superior or more stable results.

For 3D detection, Haar achieves the best or near-best performance across categories, particularly under Moderate and Hard settings, with car AP reaching 78.85% and 76.00%, pedestrian AP reaching 60.76%/57.74%, and cyclist AP reaching 59.60%/56.92%. These results indicate that its simple orthogonal structure better preserves fine-grained spatial details, which are critical under sparse LiDAR observations.

Table 6. Comparison of 3D object detection performance of MVX-Net + ETAIB + DQGAF using different wavelet bases.

Wavelet	Mod	3D Car AP_{R40} @0.7			3D Ped. AP_{R40} @0.5			3D Cyc. AP_{R40} @0.5			FPS (Hz)	Mem (GB)
		E	M	H	E	M	H	E	M	H		
Bior3.5	L&I	**90.73**	78.80	74.33	61.30	56.41	52.23	76.07	54.96	51.48	40.98	3.20
Db4	L&I	88.83	77.57	73.38	62.79	56.87	53.97	72.78	57.66	54.65	40.80	3.20
Sym6	L&I	90.46	78.53	75.68	60.94	55.35	51.73	**81.71**	59.28	56.25	40.16	3.20
Haar	L&I	90.08	**78.85**	**76.00**	**66.47**	**60.76**	**57.74**	80.00	**59.60**	**56.92**	**42.00**	3.20

Table 7. Comparison of BEV object detection performance of MVX-Net + ETAIB + DQGAF using different Wave bases.

Wavelet	Mod	BEV Car AP_{R40} @0.7			BEV Ped. AP_{R40} @0.5			BEV Cyc. AP_{R40} @0.5			FPS (Hz)	Mem (GB)
		E	M	H	E	M	H	E	M	H		
Bior3.5	L&I	**94.58**	87.97	85.60	69.42	64.65	61.41	79.76	59.65	56.13	40.98	3.20
Db4	L&I	93.64	87.43	85.20	70.38	64.93	61.00	78.02	**62.58**	58.88	40.80	3.20
Sym6	L&I	94.36	85.93	85.32	69.13	63.68	60.85	**83.08**	62.29	**59.51**	40.16	3.20
Haar	L&I	94.02	**88.02**	**85.61**	**73.62**	**68.40**	**64.39**	81.37	61.94	59.06	**42.00**	3.20

Table 8. Comparison of AOS object detection performance of MVX-Net + ETAIB + DQGAF using different Wave bases.

Wavelet	Mod	AOS Car AP_{R40} @0.7			AOS Ped. AP_{R40} @0.5			AOS Cyc. AP_{R40} @0.5			FPS (Hz)	Mem (GB)
		E	M	H	E	M	H	E	M	H		
Bior3.5	L&I	97.82	91.91	88.90	57.54	55.16	52.21	80.99	60.15	57.78	40.98	3.20
Db4	L&I	97.72	92.01	88.91	58.39	**56.54**	**53.12**	79.86	63.10	60.41	40.80	3.20
Sym6	L&I	95.72	91.55	88.35	55.03	51.71	49.25	**84.31**	63.11	60.51	40.16	3.20
Haar	L&I	**97.86**	**92.19**	**89.16**	**59.93**	55.20	52.09	83.53	**63.83**	**62.06**	**42.00**	3.20

A similar trend is observed in BEV evaluation, where Haar provides the most stable cross-category performance, especially for pedestrians (73.62%/68.40%/64.39%). Although Bior3.5 and Db4 remain competitive, their performance is slightly inferior, suggesting that more complex frequency decompositions do not necessarily improve BEV discrimination. Sym6, however, shows advantages in cyclist detection under Easy and Hard settings (83.08% and 59.51%), reflecting its stronger high-frequency modeling ability for elongated structures.

For AOS, Wave bases exhibit clear category-dependent behavior. Haar achieves the most consistent overall performance, leading in car orientation estimation across all difficulty levels (97.86%/92.19%/89.16%) and showing strong cyclist performance (63.83% and 62.06%). For pedestrians, it remains competitive (59.93%/55.20%/52.09%), but is not uniformly dominant, as Db4 and Sym6 show marginal advantages in specific settings, indicating the absence of a universally optimal Wave across all categories.

From the perspective of image processing, the effectiveness of the Haar wavelet mainly stems from its compact support and strong spatiotemporal frequency localization capability. Compared with higher-order wavelets (e.g., Db4 and Sym6), the Haar wavelet is more effective in preserving edge discontinuities and sparse geometric structures, while introducing only mild smoothing effects. This property makes it particularly suitable for retaining high-frequency details that are critical for small and deformable objects.

Overall, the Haar wavelet achieves a good balance between structural preservation and computational efficiency, thereby leading to more stable performance across evaluation metrics such as AP, BEV, and AOS.

(3) **Ablation Analysis on the Effect of Diagonal High-Frequency Components (HH).** The ablation results in Tables 9–11 evaluate the influence of diagonal high-frequency components (HH) in the edge–texture reconstruction stage of the DQGAF module across 3D, BEV, and AOS tasks. Overall, removing HH consistently leads to more stable and robust performance, particularly for pedestrian and cyclist categories, while maintaining comparable or slightly improved performance for car detection.

For the 3D detection task, removing HH yields clear improvements on challenging categories. Specifically, pedestrian detection improves from 63.90%/58.36%/53.99% (with HH) to 66.47%/60.76%/57.74% (without HH), and cyclist detection increases from 78.30%/58.82%/56.08% to 80.00%/59.60%/56.92%. These results indicate that diagonal high-frequency responses may introduce redundant or noisy information, which is particularly detrimental for sparse and small-scale objects in LiDAR-based 3D perception.

A similar trend is observed in BEV evaluation. Removing HH consistently improves pedestrian detection performance across all difficulty levels, from 71.69%, 66.64%, and 62.44% to 73.62%, 68.40%, and 64.39% under the Easy, Moderate, and Hard settings, respectively. Cyclist detection exhibits a similar tendency, whereas car detection remains largely unaffected. These observations suggest that large rigid objects are less sensitive to diagonal high-frequency perturbations than sparse and small-scale objects.

Table 9. Ablation study on the effect of diagonal high-frequency (HH) components in edge-texture reconstruction for 3D object detection.

Setting	Car AP_{R40} @0.7			Pedestrian AP_{R40} @0.5			Cyclist AP_{R40} @0.5			FPS (Hz)	Mem (GB)
	E	M	H	E	M	H	E	M	H		
Without HH	90.08	**78.85**	**76.00**	**66.47**	**60.76**	**57.74**	80.00	**59.60**	**56.92**	42.0	3.20
With HH	**90.49**	78.60	74.03	63.90	58.36	53.99	78.30	58.82	56.08	42.0	3.20

Table 10. Ablation study on the effect of diagonal high-frequency (HH) components in edge-texture reconstruction for BEV object detection.

Setting	Car AP_{R40} @0.7			Pedestrian AP_{R40} @0.5			Cyclist AP_{R40} @0.5			FPS (Hz)	Mem (GB)
	E	M	H	E	M	H	E	M	H		
Without HH	94.02	88.02	85.61	**73.62**	**68.40**	**64.39**	**81.37**	**61.94**	**59.06**	42.0	3.20
With HH	**94.55**	**88.08**	**85.64**	71.69	66.64	62.44	78.85	61.41	58.68	42.0	3.20

Table 11. Ablation study on the effect of diagonal high-frequency (HH) components in edge-texture reconstruction for orientation-aware object detection (AOS).

Setting	Car AP_{R40} @0.7			Pedestrian AP_{R40} @0.5			Cyclist AP_{R40} @0.5			FPS (Hz)	Mem (GB)
	E	M	H	E	M	H	E	M	H		
Without HH	**97.86**	92.19	89.16	**59.93**	**55.20**	**52.09**	**83.53**	**63.83**	**62.06**	42.0	3.20
With HH	95.99	**92.40**	**89.30**	57.05	54.34	51.35	81.31	61.17	58.65	42.0	3.20

For AOS evaluation, removing HH further improves orientation-aware detection performance, particularly for pedestrians, where AP improves from 57.05%/54.34%/51.35% (with HH) to 59.93%/55.20%/52.09% (without HH). This suggests that diagonal high-frequency components may interfere with orientation-sensitive feature encoding, leading to less stable angle estimation in complex scenes.

The underlying reason for this behavior is that the diagonal high-frequency components mainly contain responses that are weakly correlated with object boundaries. In contrast, the horizontal and vertical high-frequency components are more effective at capturing object edges and texture information, providing more reliable structural cues for image–point cloud feature alignment, whereas diagonal responses are more likely to introduce noise, projection artifacts, and sensor instability.

Since the DQGAF module relies on deformable attention for cross-modal alignment, such noisy responses may interfere with the attention focusing process, thereby weakening the feature correspondence between modalities.

In addition, the HH component increases the model's sensitivity to local texture variations, which is particularly detrimental for distant and sparsely sampled objects such as pedestrians and cyclists. In contrast, suppressing the HH component can effectively improve the signal-to-noise ratio of edge–texture features, leading to more reliable cross-modal interaction and more stable proposal generation performance.

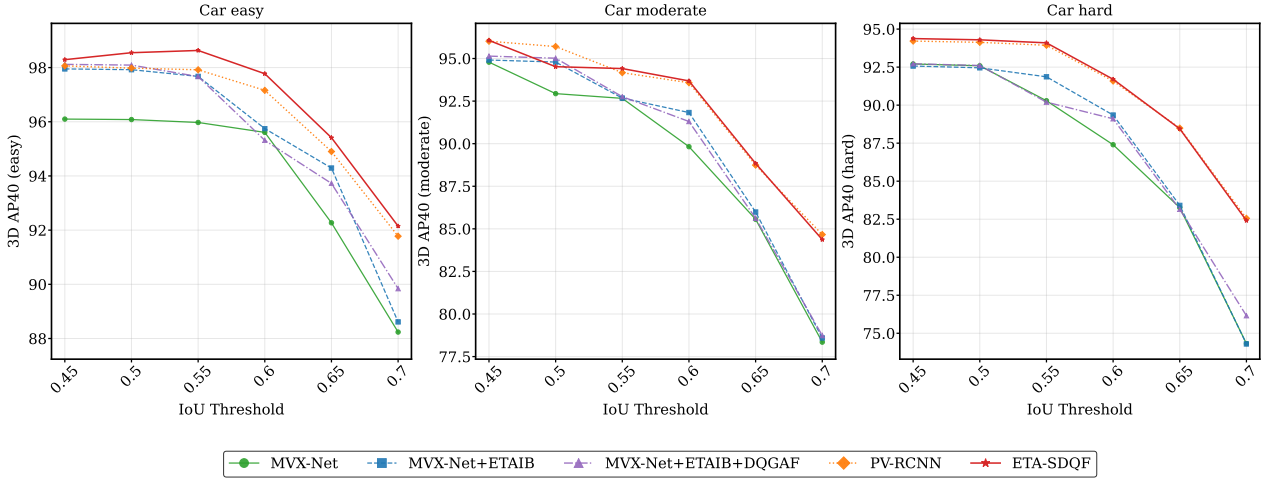
Overall, these results demonstrate that suppressing diagonal high-frequency components (HH) is beneficial for multimodal 3D object detection, as it improves feature robustness, enhances small-object discrimination, and strengthens cross-modal alignment within the DQGAF framework.

(4) **IoU Sensitivity Analysis.** To further evaluate the localization robustness of the proposed method, we conduct an IoU sensitivity analysis by varying the IoU threshold from 0.40 to 0.80 for pedestrians and cyclists, and from 0.40 to 0.70 for cars on the KITTI validation set.

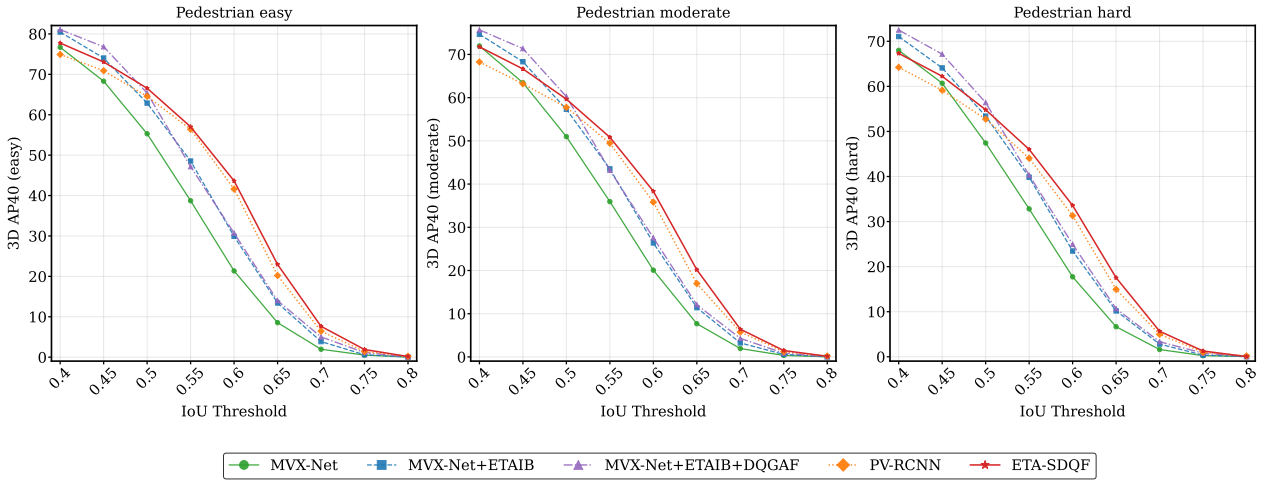
As shown in Figure 5, the proposed ETA-SDQF consistently demonstrates stable performance across different IoU thresholds. In particular, it maintains competitive results even under higher IoU thresholds, indicating its robustness in accurate 3D object localization.

For pedestrian detection, we observe that under lower IoU thresholds (0.40–0.50), methods incorporating the ETAIB and DQGAF modules achieve more significant performance improvements. This can be attributed to the joint effect of edge–texture enhancement and cross-modal alignment, which improves fine-grained feature representation as well as cross-modal feature consistency, thereby generating more candidate proposals close to ground-truth locations. However, these proposals still exhibit certain deviations in precise localization and orientation estimation.

As the IoU threshold increases, these less accurately localized proposals are gradually filtered out, leading to a diminishing performance gain or even a relative drop. In contrast, the hybrid 3D feature learning framework based on PV-RCNN significantly improves localization accuracy, enabling ETA-SDQF to achieve more stable performance under stricter IoU thresholds (0.60–0.80).

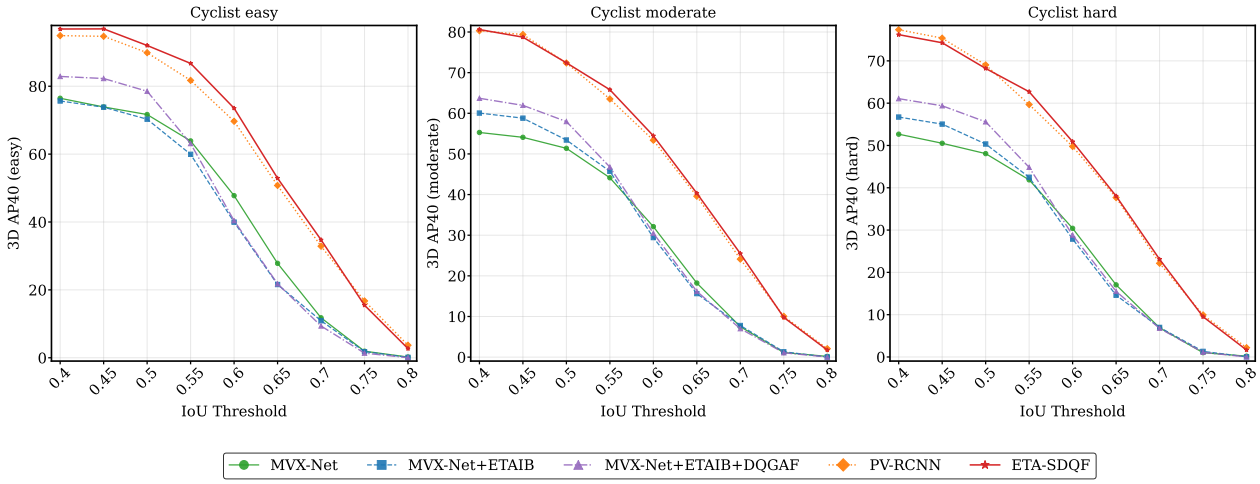


(**a**) 3D detection performance (3D Car AP_{R40}) under varying IoU thresholds (0.4–0.7) on KITTI val set.



(**b**) 3D detection performance (3D Pedestrian AP_{R40}) under varying IoU thresholds (0.4–0.8) on KITTI val set.

Figure 5. *Cont.*



(**c**) 3D detection performance (3D Cyclist AP_{R40}) under varying IoU thresholds (0.4–0.8) on KITTI val set.

Figure 5. The 3D object detection performance of each category under various IoU thresholds on the KITTI validation set.

4.4. Qualitative Results

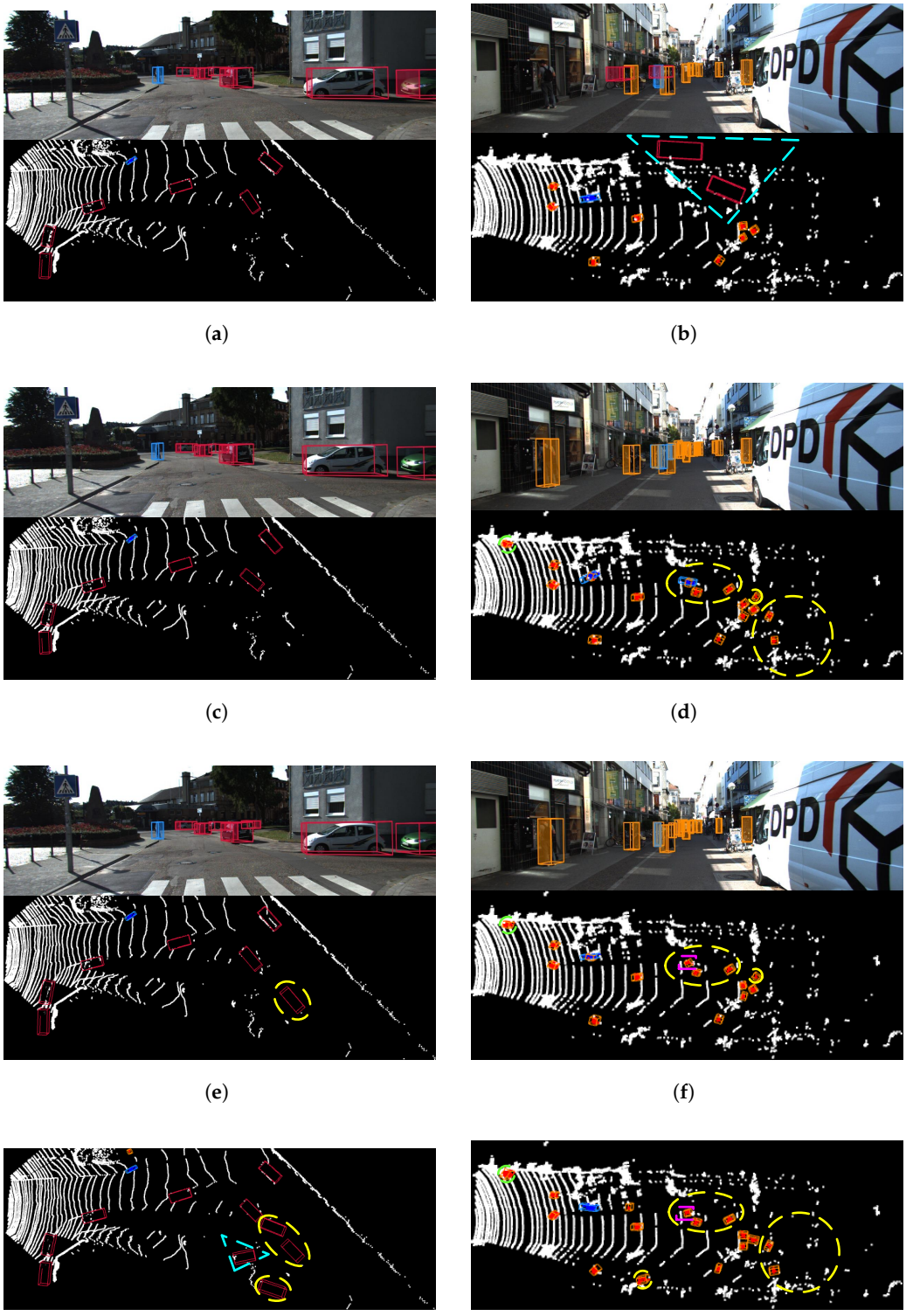
We present a qualitative comparison of 3D detection performance on representative samples from the KITTI validation set (i.e., 000382.png and 001267.png), covering different MVX-Net variants, the original MVX-Net, and the PV-RCNN baseline, as illustrated in Figure 6. Predicted cars, pedestrians, and cyclists are marked with red, orange, and blue bounding boxes, respectively. Each result pair presents the image with projected 3D detections alongside the corresponding raw LiDAR point cloud with 3D detections.

In Scene 1 (occluded vehicles under low-light conditions), the original MVX-Net suffers from multiple missed detections. This is mainly because parts of the scene are occluded by buildings, resulting in poor illumination, and the region is relatively far from the ego vehicle, leading to sparse LiDAR point returns. These results are highlighted by dashed yellow circles. Although the ETAIB module enhances visual feature representation, its effectiveness remains limited under severe occlusion. With the introduction of the DQGAF module, several previously missed vehicles are successfully recovered. PV-RCNN demonstrates stronger detection capability but still produces certain false positives, as indicated by the cyan dashed triangles. By further integrating ETAIB, DQGAF, and H3DFLF, the complete ETA-SDQF achieves improved robustness, successfully detecting distant and heavily occluded vehicles, and yielding the most comprehensive detection results among all compared methods.

In Scene 2 (dense, distant pedestrians), the original MVX-Net fails to detect multiple targets due to insufficient multimodal feature representation. Pedestrians wearing background-like clothing that remain undetected are highlighted by dashed lime-green circles. Although ETAIB enhances edge and texture information, distinguishing these challenging pedestrians remains difficult. This limitation is gradually alleviated with the incorporation of subsequent modules. The proposed fusion module further improves performance by modeling inter-feature dependencies and strengthening multimodal alignment, thereby reducing pedestrian misclassification errors, which are indicated by dashed magenta boxes. PV-RCNN also exhibits strong detection performance; however, several targets remain undetected. With the incorporation of H3DFLF, the complete ETA-SDQF model achieves the best overall performance, accurately detecting most pedestrians and cyclists while producing well-aligned 3D bounding boxes.

We further present qualitative comparisons of the proposed ETA-SDQF in near-range and far-range scenarios using representative validation samples (i.e., 000448.png and

000453.png), as illustrated in Figure 7. Predicted cars, pedestrians, and cyclists are marked with red, orange, and blue bounding boxes, respectively. Each result pair presents the image with projected 3D detections alongside the corresponding raw LiDAR point cloud with 3D detections.



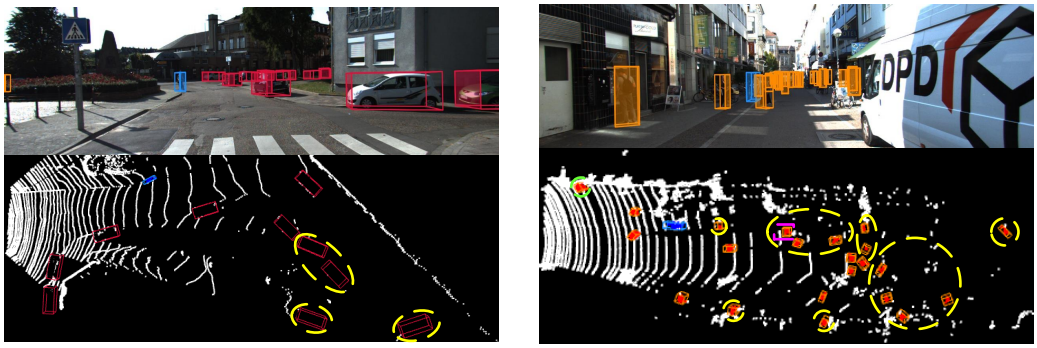
(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

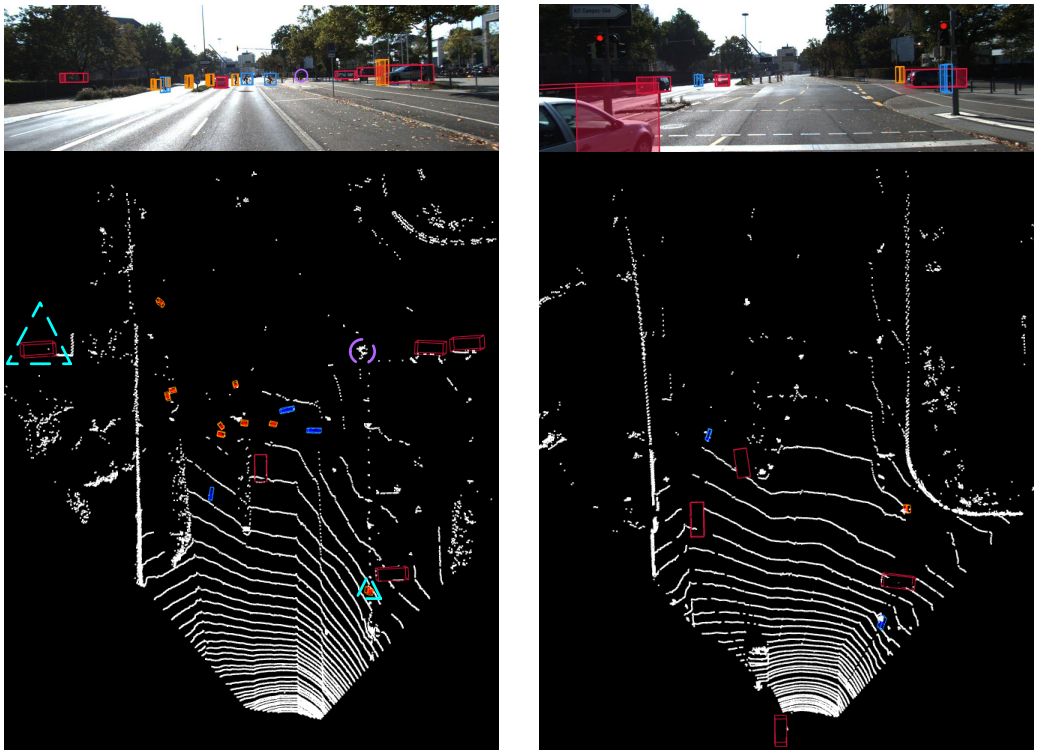
(**g**) (**h**)

Figure 6. *Cont.*



(**i**) (**j**)

Figure 6. Qualitative results on the KITTI validation set. Predicted cars, pedestrians, and cyclists are marked with red, orange, and blue bounding boxes, respectively. Missed targets by MVX-Net [20] are indicated by dashed yellow circles, while dashed cyan triangles denote false positives. Dashed lime-green circles highlight pedestrians with background-like clothing that are not detected, and dashed magenta boxes indicate misclassified pedestrians. Each result pair presents the image with projected 3D detections alongside the corresponding raw LiDAR point cloud with 3D detections. (**a**) Scene 1 of MVX-Net [20]. (**b**) Scene 2 of MVX-Net. (**c**) Scene 1 of MVX-Net + ETAIB. (**d**) Scene 2 of MVX-Net + ETAIB. (**e**) Scene 1 of MVX-Net + ETAIB + DQGAF. (**f**) Scene 2 of MVX-Net + ETAIB + DQGAF. (**g**) Scene 1 of PV-RCNN [10]. (**h**) Scene 2 of PV-RCNN. (**i**) Scene 1 of ETA-SDQF. (**j**) Scene 2 of ETA-SDQF.



(**a**) Scene 3 Objects far from the ego vehicle (**b**) Scene 4 Objects close to the ego vehicle

Figure 7. Comparison of object detection results in near and far scenes on the KITTI validation set using the ETA-SDQF. Predicted cars, pedestrians, and cyclists are marked with red, orange, and blue bounding boxes, respectively. Dashed cyan triangles denote false positives. Light purple circles highlight missed distant cyclists. Each result pair presents the image with projected 3D detections alongside the corresponding raw LiDAR point cloud with 3D detections.

In Scene 3 (long-range object detection), objects are located at relatively large distances from the ego vehicle. As a result, LiDAR points returned from these objects are sparse, and

their projected sizes in image space are small, making both localization and classification more challenging. Consequently, several false positives are observed, which are marked by dashed cyan triangles. In addition, a missed distant cyclist is highlighted by a light purple circle, further indicating that long-range object perception remains a persistent challenge.

In Scene 4 (near-range object detection), as the ego vehicle approaches the targets, most false positives and missed detections are effectively eliminated. The proposed ETA-SDQF achieves more reliable localization and classification performance in this scenario, demonstrating greater robustness for near-range object detection. These observations also indicate that far-range object detection remains a challenging problem and an important direction for future research.

5. Conclusions

This paper presents ETA-SDQF, a projection-based multimodal 3D object detection framework that leverages an edge–texture-aware image backbone and a dual-query adaptive fusion module to enable fine-grained fusion of image and LiDAR features, while combining voxel- and point-based representations to enhance 3D proposal generation. On the KITTI dataset, ETA-SDQF shows promising results in pedestrian and cyclist detection, with comparable performance on cars, indicating the potential of the multimodal fusion approach for improving 3D perception in moderately dense scenes and at medium ranges.

However, in long-range scenarios, model performance may still be limited due to the sparsity of LiDAR point clouds generated from distant objects and the relatively small object scale in image space. For objects with appearance patterns similar to the background or under challenging illumination conditions, existing RGB–LiDAR fusion methods still struggle to handle them effectively. These observations are based on the KITTI dataset, and further validation on other datasets and more complex, diverse environments is required to more comprehensively evaluate the generalization capability of the proposed method.

Future work will explore the incorporation of additional sensing modalities, such as infrared or thermal imaging, to improve robustness under occlusion, low-visibility, and camouflage-like scenarios. Meanwhile, further efforts will focus on enhancing feature representation for long-range objects and developing adaptive fusion strategies to improve localization accuracy and detection reliability. Cross-dataset evaluation will also be an important direction for assessing the generalization ability of the proposed framework.

Overall, this work demonstrates the potential of image edge–texture-guided multimodal fusion for 3D object detection, providing insights for building more robust multi-sensor perception systems, while also highlighting current limitations and directions for future improvement.

Author Contributions: Conceptualization, Y.W., Z.Z., K.L., L.C. and R.W.; methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, visualization, Y.W.; resources, funding acquisition, Z.Z., K.L. and R.W.; writing—review and editing, supervision, Z.Z., K.L., L.C. and R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China under grant No. U2570202 and by the Natural Science Foundation of the Inner Mongolia Autonomous Region under grant No. 2025QN05132.

Data Availability Statement: The data used in this study are publicly available from the KITTI dataset at: http://www.cvlibs.net/datasets/kitti/ (accessed on 6 April 2026).

Acknowledgments: During the preparation of this manuscript, the author(s) used ChatGPT (GPT-5.3-mini) for the purposes of English academic writing assistance, including translation, grammar

refinement, and text polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Liu, J.; Wang, H.; Peng, L.; Cao, Z.; Yang, D.; Li, J. PNNUAD: Perception Neural Networks Uncertainty Aware Decision-Making for Autonomous Vehicle. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 24355–24368. [CrossRef]
2. Kumar, G.A.; Lee, J.H.; Hwang, J.; Park, J.; Youn, S.H.; Kwon, S. LiDAR and Camera Fusion Approach for Object Distance Estimation in Self-Driving Vehicles. *Symmetry* **2020**, *12*, 324. [CrossRef]
3. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A Survey on 3D Object Detection Methods for Autonomous Driving Applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [CrossRef]
4. Guo, X.; Shi, S.; Wang, X.; Li, H. LIGA-Stereo: Learning LiDAR Geometry Aware Representations for Stereo-based 3D Detector. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 3133–3143. [CrossRef]
5. Liu, Z.; Zhou, D.; Lu, F.; Fang, J.; Zhang, L. AutoShape: Real-Time Shape-Aware Monocular 3D Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 15621–15630. [CrossRef]
6. Peng, X.; Zhu, X.; Wang, T.; Ma, Y. SIDE: Center-based Stereo 3D Detector with Structure-aware Instance Depth Estimation. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 225–234. [CrossRef]
7. Yan, W.; Gu, M.; Ren, J.; Yue, G.; Liu, Z.; Xu, J.; Lin, W. Collaborative structure and feature learning for multi-view clustering. *Inf. Fusion* **2023**, *98*, 101832. [CrossRef]
8. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4490–4499. [CrossRef]
9. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection from Point Clouds. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12689–12697. [CrossRef]
10. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10526–10535. [CrossRef]
11. Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; Jia, J. LargeKernel3D: Scaling up Kernels in 3D Sparse CNNs. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 13488–13498. [CrossRef]
12. Wang, L.; Zhang, P.; Li, M.; Zhang, F. Sparse Sensor Fusion for 3D Object Detection with Symmetry-Aware Colored Point Clouds. *Symmetry* **2024**, *16*, 1690. [CrossRef]
13. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. PointPainting: Sequential Fusion for 3D Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4603–4611. [CrossRef]
14. Xu, S.; Zhou, D.; Fang, J.; Yin, J.; Bin, Z.; Zhang, L. FusionPainting: Multimodal Fusion with Adaptive Attention for 3D Object Detection. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 3047–3054. [CrossRef]
15. Xu, S.; Li, F.; Song, Z.; Fang, J.; Wang, S.; Yang, Z.X. Multi-Sem Fusion: Multimodal Semantic Fusion for 3-D Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
16. Meyer, G.P.; Charland, J.; Hegde, D.; Laddha, A.; Vallespi-Gonzalez, C. Sensor Fusion for Joint 3D Object Detection and Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 6–20 June 2019; pp. 1230–1237. [CrossRef]
17. Wang, C.; Ma, C.; Zhu, M.; Yang, X. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 11789–11798. [CrossRef]
18. Huang, T.; Liu, Z.; Chen, X.; Bai, X. EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection. In *Proceedings of the Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 35–52.
19. Liu, Z.; Huang, T.; Li, B.; Chen, X.; Wang, X.; Bai, X. EPNet++: Cascade Bi-Directional Fusion for multimodal 3D Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8324–8341. [CrossRef] [PubMed]

20. Sindagi, V.A.; Zhou, Y.; Tuzel, O. MVX-Net: Multimodal VoxelNet for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 5–20 June 2019; pp. 7276–7285.
21. Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; Jia, J. Focal Sparse Convolutional Networks for 3D Object Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 5418–5427. [CrossRef]
22. Wang, Z.; Zhan, W.; Tomizuka, M. Fusing Bird's Eye View LIDAR Point Cloud and Front View Camera Image for 3D Object Detection. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Suzhou, China, 26–30 June 2018 2018; pp. 1–6. [CrossRef]
23. Liang, M.; Yang, B.; Wang, S.; Urtasun, R. Deep Continuous Fusion for Multi-sensor 3D Object Detection. In *Proceedings of the Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 663–678.
24. Song, Z.; Wei, H.; Bai, L.; Yang, L.; Jia, C. GraphAlign: Enhancing Accurate Feature Alignment by Graph matching for multimodal 3D Object Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 3335–3346. [CrossRef]
25. Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; Zhao, F. AutoAlignV2: Deformable Feature Aggregation for Dynamic multimodal 3D Object Detection. *arXiv* **2022**, arXiv:2207.10316.
26. Li, X.; Shi, B.; Hou, Y.; Wu, X.; Ma, T.; Li, Y.; He, L. Homogeneous multimodal Feature Fusion and Interaction for 3D Object Detection. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 691–707.
27. Wang, Z.; Jia, K. Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal 3D Object Detection. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1742–1749. [CrossRef]
28. Yang, H.; Liu, Z.; Wu, X.; Wang, W.; Qian, W.; He, X.; Cai, D. Graph R-CNN: Towards Accurate 3D Object Detection with Semantic-Decorated Local Graph. In *Proceedings of the Computer Vision—ECCV 2022* Springer: Cham, Switzerland, 2022; pp. 662–679.
29. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 918–927. [CrossRef]
30. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3D Object Detection Network for Autonomous Driving. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6526–6534. [CrossRef]
31. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3D Proposal Generation and Object Detection from View Aggregation. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–8. [CrossRef]
32. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.L. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 1080–1089. [CrossRef]
33. Li, Y.; Yu, A.W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q.V.; et al. DeepFusion: Lidar-Camera Deep Fusion for multimodal 3D Object Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 17161–17170. [CrossRef]
34. Yang, Z.; Chen, J.; Miao, Z.; Li, W.; Zhu, X.; Zhang, L. DeepInteraction: 3D Object Detection via Modality Interaction. *arXiv* **2022**, arXiv:2208.11112.
35. Zhang, C.; Wang, H.; Chen, L.; Li, Y.; Cai, Y. MixedFusion: An Efficient Multimodal Data Fusion Framework for 3-D Object Detection and Tracking. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 1842–1856. [CrossRef] [PubMed]
36. Li, Y.; Chen, Y.; Qi, X.; Li, Z.; Sun, J.; Jia, J. Unifying Voxel-based Representation with Transformer for 3D Object Detection. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 18442–18455. [CrossRef]
37. Xie, Y.; Xu, C.; Rakotosaona, M.J.; Rim, P.; Tombari, F.; Keutzer, K.; Tomizuka, M.; Zhan, W. SparseFusion: Fusing multimodal Sparse Representations for Multi-Sensor 3D Object Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 17545–17556. [CrossRef]
38. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 2774–2781. [CrossRef]
39. Wang, H.; Tang, H.; Shi, S.; Li, A.; Li, Z.; Schiele, B.; Wang, L. UniTR: A Unified and Efficient multimodal Transformer for Bird's-Eye-View Representation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 6769–6779. [CrossRef]

40. Yan, J.; Liu, Y.; Sun, J.; Jia, F.; Li, S.; Wang, T.; Zhang, X. Cross Modal Transformer: Towards Fast and Robust 3D Object Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 18222–18232. [CrossRef]
41. Chen, Y.; Yu, Z.; Chen, Y.; Lan, S.; Anandkumar, A.; Jia, J.; Alvarez, J.M. FocalFormer3D : Focusing on Hard Instance for 3D Object Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 8360–8371. [CrossRef]
42. Chen, X.; Zhang, T.; Wang, Y.; Wang, Y.; Zhao, H. FUTR3D: A Unified Sensor Fusion Framework for 3D Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 172–181. [CrossRef]
43. Wu, H.; Wen, C.; Shi, S.; Li, X.; Wang, C. Virtual Sparse Convolution for Multimodal 3D Object Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 21653–21662. [CrossRef]
44. Tang, Y.; Liu, M.; Li, B.; Wang, Y.; Ouyang, W. NAS-PED: Neural Architecture Search for Pedestrian Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1800–1817. [CrossRef] [PubMed]
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
46. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
47. Prasanth, T.; Padhy, R.P.; Sivaselvan, B. LiDAR Sensor-Based Dual-Scale Fusion Approach for Bird's-Eye View Sensing in Autonomous Vehicles. *IEEE Sens. Lett.* **2026**, *10*, 1–4. [CrossRef]
48. Prasanth, T.; Padhy, R.P.; Sivaselvan, B. MSFASA-3DNet: Multi-Scale Feature Aggregation and Spatial Attention for 3D Object Detection. *IEEE Trans. Artif. Intell.* **2026**, 1–13. [CrossRef]
49. Xia, Q.; Ye, W.; Wu, H.; Zhao, S.; Xing, L.; Huang, X.; Deng, J.; Li, X.; Wen, C.; Wang, C. HINTED: Hard Instance Enhanced Detector with Mixed-Density Feature Fusion for Sparsely-Supervised 3D Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 15321–15330. [CrossRef]
50. Zhang, Y.; Hu, Q.; Xu, G.; Ma, Y.; Wan, J.; Guo, Y. Not All Points Are Equal: Learning Highly Efficient Point-based Detectors for 3D LiDAR Point Clouds. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18931–18940. [CrossRef]
51. Thurimerla, P.; Padhy, R.P.; Sivaselvan, B.; Bakshi, S. LinkNet3D: An efficient end-to-end network for 3D object detection. *Comput. Electr. Eng.* **2026**, *129*, 110821. [CrossRef]
52. Cortinhal, T.; Gouigah, I.; Aksoy, E.E. Semantics-aware LiDAR-Only Pseudo Point Cloud Generation for 3D Object Detection. In Proceedings of the 2024 IEEE Intelligent Vehicles Symposium (IV), Shenzhen, China, 22–24 November 2024; pp. 3220–3226. [CrossRef]
53. Thurimerla, P.; Padhy, R.P.; Sivaselvan, B. PCNet3D++: A pillar-based cascaded 3D object detection model with an enhanced 2D backbone. *Image Vis. Comput.* **2026**, *166*, 105854. [CrossRef]
54. Ye, W.; Xia, Q.; Wu, H.; Dong, Z.; Zhong, R.; Wang, C.; Wen, C. Fade3D: Fast and Deployable 3D Object Detection for Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 12934–12946. [CrossRef]
55. Hayeon, O.; Yang, C.; Huh, K. SeSame: Simple, Easy 3D Object Detection with Point-Wise Semantics. In *Proceedings of the Computer Vision—ACCV 2024*; Springer: Singapore, 2025; pp. 211–227.
56. Gong, X.; Huang, X.; Chen, S.; Zhang, B. Enhancing 3D Detection Accuracy in Autonomous Driving through Pseudo-LiDAR Augmentation and Downsampling. In Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Shenzhen, China, 22–24 November 2024; pp. 1105–1110. [CrossRef]
57. Sood, V.; Bnalin; Pandey, G. Point Virtual Transformer. *arXiv* **2026**, arXiv:2602.06406.
58. Zhang, H.; Yang, D.; Yurtsever, E.; Redmill, K.A.; Özgüner, Ü. Faraway-Frustum: Dealing with Lidar Sparsity for 3D Object Detection using Fusion. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 2646–2652. [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.