

Article

A LiDAR-Based Multimodal 3D Object Detection Algorithm for Intelligent Driving in Open-Pit Mines

Shuqi Wang * and Xinyi Zhang

School of Communication and Information Engineering, Xi'an University of Science and Technology,
Xi'an 710600, China; 13378122538@163.com

* Correspondence: wangshuqi@xust.edu.cn

Abstract

This study addresses the challenges of sparse long-range point clouds, complex background interference, and inconsistent localization quality in 3D object detection for intelligent driving in open-pit mines. A camera–LiDAR multimodal detection method based on Voxel R-CNN is proposed. We introduce a Multimodal Focal Sparse Voxel Enhancement module that combines Focal Sparse Convolution with shallow visual features to guide voxel importance prediction and selective sparse propagation. An Intersection over Union (IoU)-Aware Quality and Geometry Refinement Head is further designed to improve the localization accuracy and ranking reliability of 3D proposals. Experimental results show that the proposed method achieves BEV mAP@0.40 and 3D mAP@0.40 values of 63.43% and 61.23%, respectively, outperforming the strongest comparison method, MambaFusion, by 5.37 and 7.61 percentage points. In the 60–80 m range, the average translation error is reduced to 0.41 m, while the inference speed reaches 27 FPS. These results demonstrate that the proposed method improves the detection and localization of distant sparse objects in complex open-pit mine environments while maintaining real-time inference capability.

Keywords: intelligent driving in open-pit mines; multimodal fusion; 3D object detection; Voxel R-CNN; sparse point cloud perception

1. Introduction

With the development of smart mining and unmanned transportation technologies in open-pit mines, increasingly stringent requirements have been imposed on the accuracy, stability, and real-time performance of forward environmental perception systems. Compared with urban roads, open-pit mine roads have irregular boundaries, while the ground, slopes, and large-scale equipment form extensive and complex backgrounds. These environments are also affected by strong glare, shadows, dust, and low contrast. Moreover, mining trucks, construction machinery, civilian vehicles, and pedestrians differ substantially in scale and geometric shape. In particular, within the 60–80 m range, LiDAR returns become increasingly sparse, and objects often contain only a small number of discrete points or partial surface information, resulting in incomplete foreground features and increased localization errors. Therefore, achieving robust representations of distant sparse objects and accurate 3D bounding box estimation remains an important challenge for intelligent driving perception in open-pit mines.

Academic Editor: Janos Botzheim

Received: 14 August 2026

Revised: 5 September 2026

Accepted: 10 September 2026

Published: 11 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

LiDAR directly provides the 3D positions and geometric structures of objects. In recent years, related research has gradually evolved from conventional point cloud encoding toward distance-aware modeling, sparse feature enhancement, and multiscale feature extraction. DO-SA&R [1] enhances distant-object features through distance-aware sampling and regression. SphereFormer [2] uses radial-window attention to accommodate the distance-dependent sparsity of point clouds, while RAANet [3] further combines range-aware convolution with point density estimation to alleviate feature imbalance between near- and long-range objects. For open-pit mining scenarios, Mine-SSD [4] enhances sparse-object representations through dual-threshold sampling and radius-adaptive grouping, whereas MHVTNet [5] uses multiscale hierarchical voxels and attention mechanisms to strengthen representations of mining objects at different scales. However, as the detection distance increases, the number of non-empty voxels associated with an object decreases rapidly, and limited foreground responses can easily be weakened by background features from the ground and slopes. Therefore, LiDAR alone remains insufficient to fully recover object structures.

Cameras provide denser edge, contour, and texture information and thus serve as an important complement to sparse point clouds. Early studies, including MVXNet [6] and Frustum PointNets [7], mainly employed point/voxel projection and frustum constraints to establish cross-modal associations. Subsequently, MSMD Fusion [8], SparseFusion [9], SupFusion [10], and CMT [11] improved fusion performance through multiscale interaction, sparse representation fusion, auxiliary supervision, and cross-modal Transformers, respectively. More recently, GAFusion [12], IS-Fusion [13], Lift-Attend-Splat [14], and DSTR [15] have further enhanced cross-modal representations through adaptive interaction, instance-scene collaboration, and attention mechanisms. EVT [16] and MambaFusion [17] focus on efficient view transformation and preservation of height information, whereas MoME [18] investigates multimodal robustness under sensor degradation. These approaches have advanced camera-LiDAR fusion; however, images captured in open-pit mines are susceptible to strong illumination, shadows, and dust. Repeated injection of visual features across multiple scales and stages may continuously propagate background noise through the point cloud backbone. Therefore, effectively using visual information to facilitate the identification of long-range foreground regions while restricting the propagation of unstable visual features remains an important issue.

In addition, long-range point cloud sparsity can lead to inconsistency between classification confidence and 3D localization quality. Previous studies have employed localization confidence or IoU prediction to improve proposal quality estimation and non-maximum suppression (NMS) ranking [19–21]. However, pedestrians, civilian vehicles, and large construction machinery in mining environments exhibit substantial differences in physical dimensions and geometric proportions, making it difficult for shared regression parameters to adequately model class-dependent geometric residual distributions. Focal Sparse Convolution (FSConv) [22] selectively expands sparse features based on voxel importance prediction, providing a promising mechanism for suppressing unnecessary background expansion. Meanwhile, Voxel R-CNN [23] employs voxel Region of Interest (RoI) pooling for efficient two-stage 3D bounding box refinement, providing a suitable foundation for balancing detection accuracy and real-time performance. Therefore, existing methods still lack a unified framework that can enhance long-range sparse foreground representations, constrain the propagation of unreliable visual information, and simultaneously improve 3D proposal localization and ranking reliability in complex open-pit mine environments.

Based on the above analysis, this study develops a two-stage camera-LiDAR multimodal 3D object detection model based on Voxel R-CNN for multiclass object detection within the forward 0–80 m range of transport vehicles in open-pit mines. The model

primarily addresses insufficient foreground responses to distant sparse objects, interference from complex backgrounds, and inconsistencies between 3D proposal localization quality and ranking reliability. Different from directly fusing visual features into point/voxel representations, MFS uses shallow visual features to assist voxel importance estimation and thereby guide where sparse feature propagation should occur, while subsequent stages remain LiDAR-driven. Furthermore, compared with directly using localization confidence or IoU prediction for proposal scoring, IAQH jointly models class-dependent proposal quality, class-specific geometric residuals, and class-agnostic IoU quality. These two designs respectively address sparse foreground representation before proposal generation and localization-quality refinement after proposal generation. The main contributions are as follows:

- (1) A Multimodal Focal Sparse Voxel Enhancement (MFS) algorithm is proposed. Shallow visual features are used to assist voxel importance estimation and selectively guide sparse neighborhood propagation, rather than being repeatedly fused with LiDAR features throughout the backbone. Visual information is injected only once at a shallow scale, while the subsequent feature extraction stages remain LiDAR-driven. In this way, distant sparse foreground responses are enhanced while unnecessary background activation and visual-noise propagation are reduced.
- (2) An IoU-Aware Quality and Geometry Refinement Head (IAQH) is proposed. Class-dependent proposal quality, class-specific geometric residuals, and class-agnostic IoU quality are jointly modeled to improve proposal refinement and ranking. Without modifying the first-stage predicted classes, the two quality estimates are combined to recalibrate proposal ranking, thereby improving localization accuracy and ranking reliability.

2. Materials and Methods

2.1. Multimodal 3D Object Detection Model Architecture for Complex Open-Pit Mine Environments

To accurately detect forward 3D objects for transport vehicles in open-pit mines and address sparse point clouds of distant objects, interference from background voxels, and inconsistency between 3D proposal classification confidence and localization quality, a camera–LiDAR multimodal 3D object detection model is developed based on the two-stage Voxel R-CNN framework. The overall architecture is shown in Figure 1.

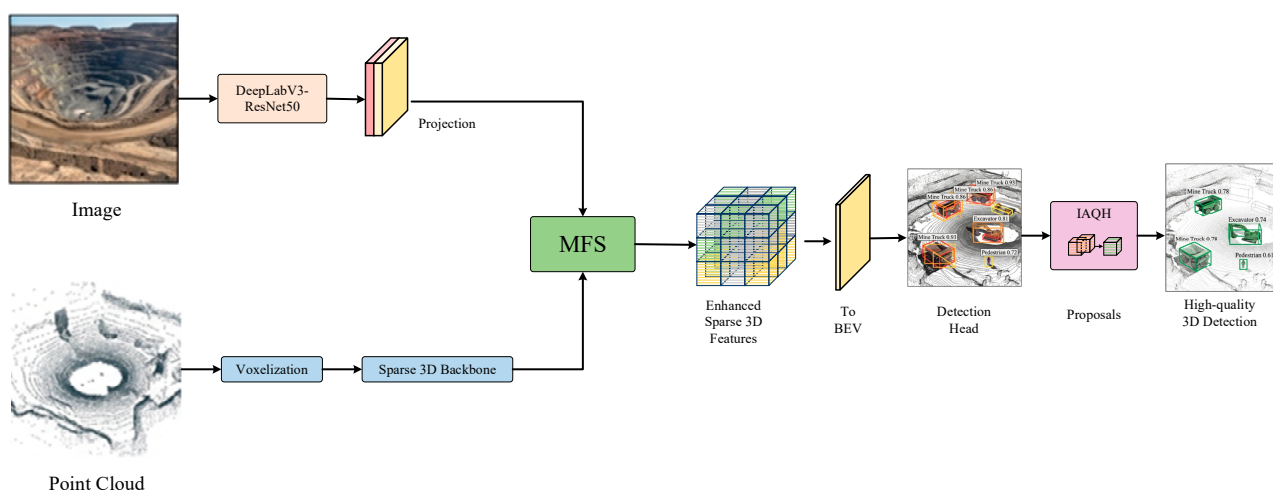


Figure 1. Model architecture.

The image branch employs a DeepLabV3-ResNet50 encoder. The output of layer1, the first residual stage of the ResNet-50 backbone, is extracted as the shallow visual structural feature. Based on the calibration relationship between the camera and LiDAR, the centers of visible voxels are projected onto the image plane, and the corresponding image edge, contour, and local texture features are sampled. For voxels outside the camera field of view or with invalid projection depths, the image features are set to zero while the complete LiDAR voxel features are retained.

In the point cloud branch, the original VoxelBackBone8x of Voxel R-CNN is replaced with a sparse backbone constructed using FSConv. First, LiDAR voxel features are used for importance prediction and focal sparse propagation at the conv1 stage. Image-conditioned FSConv is then introduced at the conv1 output scale, allowing the projected visual features and LiDAR voxel features to jointly participate in voxel importance prediction and dynamic selection of expansion locations. The conv2 and conv3 stages continue to employ LiDAR-driven FSConv to aggregate sparse foreground responses at different scales, whereas conv4 uses conventional sparse convolution for deep feature extraction. Image features are injected only once at the conv1 output scale to reduce the risk of persistent propagation of strong illumination, shadows, and complex background textures into deeper layers of the point cloud network.

The enhanced 3D sparse voxel features are mapped to a bird's-eye view, where the first-stage detection head performs object class prediction, 3D bounding box regression, and orientation estimation to generate 3D proposals. In the second stage, proposal features are obtained using RoI grid pooling. Three branches are then established in parallel: an eight-channel class-conditional quality branch, a class-specific bounding box regression branch, and a class-agnostic IoU quality branch. Importantly, the second-stage refinement procedure does not reclassify the 3D proposals. The class of each RoI is retained from the first-stage prediction. The class-conditional quality branch estimates proposal quality conditioned on that class, the class-specific regression branch selects the corresponding seven-dimensional geometric residuals, and the IoU branch provides a localization quality estimate shared across classes. During inference, the two quality outputs are independently normalized using sigmoid functions, and their probability product is used as the final 3D proposal ranking score.

The resulting model forms a complete detection pipeline comprising front-end sparse-object representation enhancement, first-stage 3D proposal generation, and back-end quality and geometry refinement. MFS primarily enhances feature visibility and 3D proposal generation for distant sparse objects, whereas IAQH mainly improves 3D bounding box regression accuracy and NMS ranking reliability. Together, these components improve 3D object detection and localization performance in open-pit mine environments.

2.2. Multimodal Focal Sparse Voxel Enhancement (MFS)

To address missed 3D detections caused by sparse point clouds of distant objects and insufficient foreground feature propagation in open-pit mine environments, the Multimodal Focal Sparse Voxel Enhancement (MFS) algorithm is developed based on the two-stage voxel detection framework of Voxel R-CNN.

MFS retains the voxelization, BEV mapping, first-stage detection head, and second-stage RoI refinement framework of Voxel R-CNN while primarily modifying VoxelBackBone8x. As shown in Figure 2, the improved backbone comprises the LiDAR-Driven Focal Sparse Propagation (LFS) module, the Voxel Foreground Estimation and Selection (VFE) module, and the Visual Semantic Injection (VSI) module. LFS performs selective neighborhood propagation based on FSConv; VFE predicts voxel foreground importance and determines expansion locations; and VSI introduces image features at a single shallow

scale, allowing visual information to participate in voxel selection and feature fusion. These modules jointly alleviate insufficient feature representation and missed detections of distant objects.

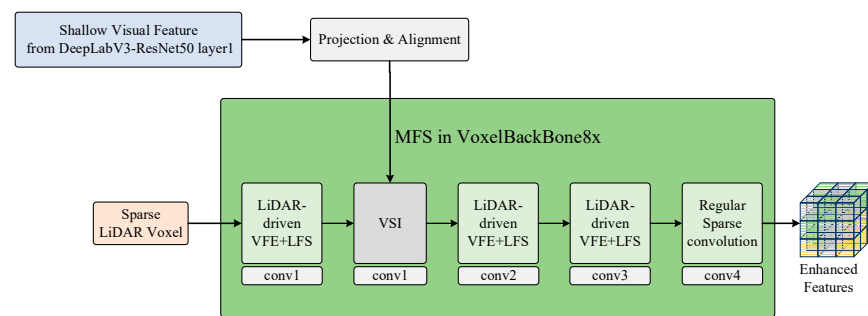


Figure 2. Block diagram of the MFS algorithm.

2.2.1. FSConv-Based Sparse Neighborhood Propagation and Voxel Importance Selection

For onboard forward perception in open-pit mines, mining vehicles within the 60–80 m range typically contain only a small number of discrete voxels, while small objects such as pedestrians contain even fewer valid voxels. The original Voxel R-CNN employs VoxelBackBone8x to extract multiscale sparse voxel features. As shown in Figure 3, submanifold sparse convolution updates features only at existing non-empty voxel locations, making it difficult to aggregate the small number of spatially separated responses from distant objects. Conventional sparse convolution can expand the feature propagation range but may simultaneously activate background regions such as the ground, slopes, and large equipment. Therefore, FSConv is introduced into the conv1, conv2, and conv3 stages of VoxelBackBone8x, where voxel importance controls the locations of neighborhood propagation. Conventional sparse convolution is retained in conv4 for deep feature extraction. This configuration selectively expands the feature propagation range of distant foreground voxels while suppressing irrelevant background expansion.

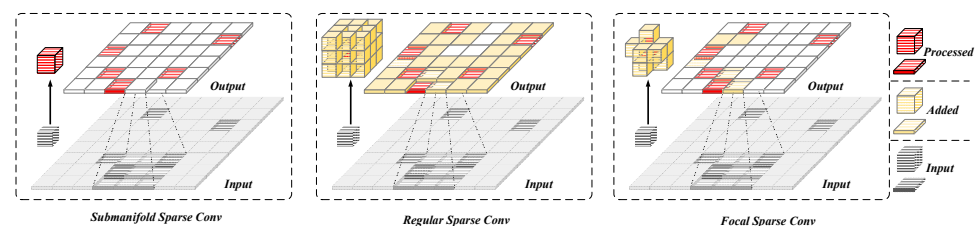


Figure 3. Comparison of different sparse convolution feature propagation mechanisms.

In the improved Voxel R-CNN backbone, VFE modules are deployed in the LiDAR branch of conv1, the visually conditioned branch at the conv1 output scale, and the LiDAR branches of conv2 and conv3, resulting in four voxel importance prediction branches.

FSConv first uses the importance prediction branches to estimate the latent foreground probability for each non-empty voxel. Let the predicted logit for voxel v in the k th branch be $s_v^{(k)}$, whose importance probability is expressed as:

$$\hat{m}_v^{(k)} = \sigma(s_v^{(k)}) \quad (1)$$

Non-empty voxels at the current scale are ranked according to their importance probabilities. Only the top 50% of voxels are selected as the expansion set, and LFS performs FSConv neighborhood propagation on the selected voxels, while the remaining voxels retain their original sparse states. This procedure converts the fixed sparse convolution

propagation in the original Voxel R-CNN into foreground-guided dynamic propagation, concentrating computation and feature expansion on potential object regions.

During training, foreground and background labels are generated according to whether a voxel center lies within any ground-truth 3D bounding box. Focal Loss is used to supervise the four importance prediction branches, and the auxiliary loss is defined as follows:

$$\mathcal{L}_{\text{voxel}} = \sum_{k=1}^4 \lambda_k^{\text{voxel}} \mathcal{L}_{\text{focal}}^{(k)} \quad (2)$$

Here, $\mathcal{L}_{\text{focal}}^{(k)}$ represents the voxel importance loss for the k th branch, and λ_k^{voxel} is the corresponding weight; this paper adopts equal weights. Therefore, VFE is responsible for filtering out potential foreground locations from the multiscale voxel features of Voxel R-CNN, while LFS uses FSConv to perform sparse neighborhood propagation on the selected locations, thereby enhancing features of distant objects and suppressing background expansion.

2.2.2. Visual Conditional Enhancement and 3D Proposal Generation Based on DeepLabV3-ResNet50

FSConv-based sparse propagation improves the spatial aggregation of distant voxels but cannot compensate for missing edge, contour, local texture, and class-semantic information in point clouds. When mining trucks, wide-body dump trucks, and water trucks exhibit similar local geometric distributions, or when pedestrians contain only a few LiDAR points, LiDAR features alone may be insufficient for reliable classification. Therefore, image features are introduced into the improved Voxel R-CNN backbone. Because open-pit mine images are susceptible to glare, shadows, and dust, shallow visual features are injected only once at the conv1 output scale to avoid persistent propagation of visual noise caused by repeated multiscale fusion.

This paper employs DeepLabV3-ResNet50 to extract shallow structural features from images and, based on the calibration relationship between the LiDAR and the camera, projects the voxel centers onto the image plane. Let the voxel center be $\mathbf{p}_v$, the camera's intrinsic matrix be $\mathbf{K}$, and the extrinsic transformation from LiDAR to camera be $[\mathbf{R}_{L \rightarrow C} \ t_{L \rightarrow C}]$; then, the projection relationship is expressed as:

$$\eta_v \begin{bmatrix} u_v \\ v_v \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}_{L \rightarrow C} \ t_{L \rightarrow C}] \begin{bmatrix} \mathbf{p}_v \\ 1 \end{bmatrix} \quad (3)$$

For voxels with valid projections, the corresponding visual features are sampled from the shallow image feature map. The feature map extracted from the layer1 stage of DeepLabV3-ResNet50 contains 256 channels and is reduced to 16 channels through a channel-reduction layer before visual feature sampling, thereby matching the dimensionality of the LiDAR voxel features at the conv1 output scale. For voxels outside the camera field of view or with invalid projection depths, the visual features are set to zero, while the LiDAR voxel features are retained.

At the conv1 output scale, the LiDAR voxel feature $x_v^L \in \mathbb{R}^{16}$ and the aligned visual feature $x_v^I \in \mathbb{R}^{16}$ are concatenated along the channel dimension to form a 32-dimensional multimodal feature $[x_v^L \parallel x_v^I] \in \mathbb{R}^{32}$. The concatenated feature is input to VFE for voxel importance estimation and selection of the voxel locations for sparse neighborhood propagation. Multimodal FSConv subsequently performs selective sparse propagation according to the predicted voxel importance, producing a 16-dimensional output feature $\hat{x}_v^L \in \mathbb{R}^{16}$. The FSConv output is then combined with the corresponding 16-dimensional visual feature by elementwise addition to obtain the enhanced voxel feature:

$$\mathbf{x}_v^{\text{fuse}} = \begin{cases} \hat{\mathbf{x}}_v^L + \mathbf{x}_v^I & v \in \mathcal{V}_{\text{vis}} \\ \hat{\mathbf{x}}_v^L & v \notin \mathcal{V}_{\text{vis}} \end{cases} \quad (4)$$

where $\hat{\mathbf{x}}_v^L \in \mathbb{R}^{16}$ denotes the output of multimodal FSConv, $\mathbf{x}_v^I \in \mathbb{R}^{16}$ denotes the aligned visual feature at the projected voxel location, and $\mathcal{V}_{\text{vis}}$ denotes the set of voxels with valid image projections. The elementwise addition preserves the feature dimensionality, yielding $\mathbf{x}_v^{\text{fuse}} \in \mathbb{R}^{16}$. Accordingly, channel-wise concatenation is used to construct the multimodal input for voxel importance estimation, whereas elementwise addition is applied after FSConv to obtain the enhanced voxel representation.

The fused features are sequentially processed through FSConv on the LiDAR side in conv2 and conv3, and through conventional sparse convolution in conv4. Image features are not re-injected at subsequent scales, thereby reducing the risk of strong light, shadows, and complex background textures propagating persistently through deep point cloud networks.

The fused features are further processed by the LiDAR-side VFE-LFS units in conv2 and conv3 and the conventional sparse convolution in conv4 to obtain the 3D sparse feature representation of the improved Voxel R-CNN:

$$\mathbf{F}_{3D} \in \mathbb{R}^{N \times 128} \quad (5)$$

Here, N represents the final number of non-empty voxels. Subsequently, following the single-stage detection workflow of Voxel R-CNN, $\mathbf{F}_{3D}$ is mapped to a bird's-eye view, where the detection head performs class prediction, 3D bounding box regression, and orientation estimation, and generates 3D proposals. Single-stage detection loss:

$$\mathcal{L}_{\text{rpn}} = \lambda_{\text{cls}}^{\text{rpn}} \mathcal{L}_{\text{cls}}^{\text{rpn}} + \lambda_{\text{loc}}^{\text{rpn}} \mathcal{L}_{\text{loc}}^{\text{rpn}} + \lambda_{\text{dir}}^{\text{rpn}} \mathcal{L}_{\text{dir}}^{\text{rpn}} \quad (6)$$

Specifically, $\mathcal{L}_{\text{cls}}^{\text{rpn}}$ represents the category anchor classification loss, $\mathcal{L}_{\text{loc}}^{\text{rpn}}$ represents the 3D bounding box regression loss, and $\mathcal{L}_{\text{dir}}^{\text{rpn}}$ represents the direction classification loss. The weights are set as follows: $\lambda_{\text{cls}}^{\text{rpn}} = 1.0$, $\lambda_{\text{loc}}^{\text{rpn}} = 2.0$, $\lambda_{\text{dir}}^{\text{rpn}} = 0.2$.

In summary, MFS integrates FSConv with shallow visual information in the first-stage voxel backbone of Voxel R-CNN. When the geometric information provided by long-range point clouds is insufficient, image edges, contours, and local textures assist voxel importance estimation, preventing some sparse foreground voxels from being prematurely suppressed. When images are degraded by strong illumination, dust, or low contrast, LiDAR geometric features continue to constrain the 3D positions and dimensions of objects, while regions invisible to the camera are modeled entirely by the point cloud branch. Consequently, the resulting feature representation combines single-pass visual assistance with multiscale LiDAR voxel importance refinement.

2.3. IoU-Aware Quality and Geometry Refinement Head (IAQH) Based on Voxel R-CNN

Although MFS improves the representation and 3D proposal recall of distant sparse objects, the generated 3D proposals may still exhibit errors in dimensions and heading. To further improve their 3D localization accuracy, the second-stage refinement network of Voxel R-CNN is specifically modified based on the high-recall 3D proposals generated by MFS. The resulting IoU-Aware Quality and Geometry Refinement Head (IAQH) performs proposal quality calibration, class-specific geometric refinement, and ranking reliability enhancement.

IAQH takes the high-recall 3D proposals generated by MFS as input and obtains shared RoI features through Voxel RoI Pooling. A class-conditional quality (CQ) branch, a class-specific regression (CSR) branch, and a class-agnostic IoU quality (IoUQ) branch are then established based on the shared RoI features. IAQH does not reclassify 3D proposals; instead, it retains the classes predicted by the first stage of the improved Voxel R-

CNN and performs quality estimation, geometric refinement, and ranking calibration while preserving class consistency. The IAQH architecture is shown in Figure 4.

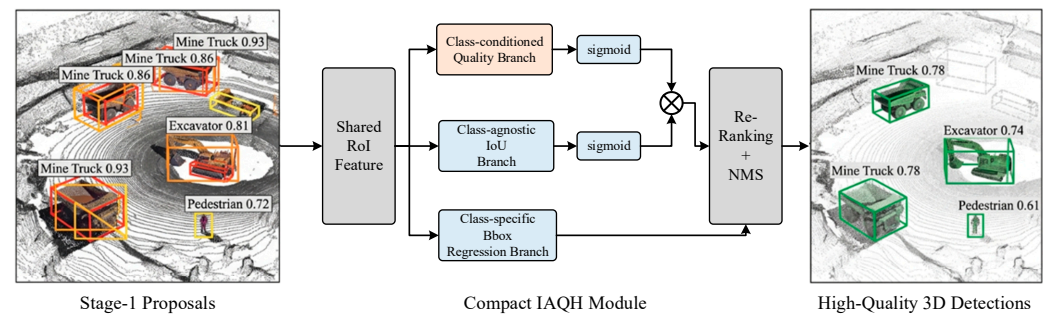


Figure 4. IAQH network architecture.

2.3.1. Continuous Quality Modeling and Class-Specific Regression Based on Voxel RoI Features

Traditional binary positive/negative labels can only determine whether a candidate bounding box meets a given IoU threshold; they cannot describe the localization differences between two candidate bounding boxes of similar quality, which hinders the model's ability to learn fine-grained ranking relationships. Therefore, this paper constructs continuous soft quality labels using the maximum 3D IoU between Voxel R-CNN candidate bounding boxes and ground-truth bounding boxes of the same class. Let u_i denote the maximum IoU of the i -th RoI with ground-truth bounding boxes of the same class, then

$$y_i = \begin{cases} 0, & u_i < u_{bg} \\ \frac{u_i - u_{bg}}{u_{fg} - u_{bg}}, & u_{bg} \leq u_i \leq u_{fg} \\ 1, & u_i > u_{fg} \end{cases} \quad (7)$$

In this case, $u_{bg} = 0.25$, $u_{fg} = 0.75$.

The category-conditional quality branch outputs an eight-channel quality logit based on shared voxel RoI features. For the i -th RoI, the soft quality label is written only to the channel corresponding to its stage-1 category:

$$t_{i,c}^{cls} = y_i \mathbb{I}(c = c_i^{roi}) \quad (8)$$

The indicator function selects the channel corresponding to the first-stage RoI class. This channel is supervised using the continuous quality label, while the targets of the remaining seven channels are set to zero. During training, all eight channels participate in binary cross-entropy loss computation rather than optimizing only the channel corresponding to the current class. This design enables the model to learn class-specific quality mappings for objects with substantially different scales, such as pedestrians, civilian vehicles, and large construction machinery.

This branch does not perform secondary classification; instead, it learns 3D proposal localization quality conditioned on the first-stage class. Meanwhile, the class-specific regression branch outputs $C \times 7$ -dimensional 3D bounding box residuals and selects the corresponding seven-dimensional regression channel according to the first-stage class. Smooth L1 loss is calculated only for foreground RoIs with a 3D IoU greater than 0.55, and a corner constraint is additionally applied to improve the stability of center, dimension, and heading refinement across different object classes.

2.3.2. Dual-Quality Joint Ranking and End-to-End Optimization of the Improved Voxel R-CNN

To further improve 3D proposal ranking reliability, IAQH establishes an independent IoU branch based on the shared Voxel RoI features to predict class-agnostic localization quality logits. The same continuous soft quality labels are used for supervision. To address the large number of low-quality background RoIs, a sigmoid focal binary cross-entropy (BCE) loss supporting continuous soft labels is adopted, with a balancing factor of $\alpha = 0.25$ and a focusing parameter of $\gamma = 2$. This loss reduces the dominance of numerous easily classified low-quality samples in gradient optimization, allowing the network to focus more strongly on difficult RoIs whose localization quality lies near the decision boundary.

The class-conditional quality branch characterizes class-dependent proposal quality, whereas the independent IoU branch characterizes localization reliability shared across classes. During inference, the outputs of the two quality branches are independently normalized using sigmoid functions and combined through their probability product to obtain the final ranking score:

$$s_i = \sigma(z_{i,c}^{\text{cls}}) \sigma(z_i^{\text{iou}}) \quad (9)$$

This score replaces the first-stage classification confidence used alone for NMS ranking. A 3D proposal obtains a high ranking score only when both its class-conditional quality and independent IoU quality are high, thereby reducing the probability that 3D proposals with localization errors are preferentially retained. The second-stage detection loss is defined as follows:

$$\mathcal{L}_{\text{rcnn}} = \mathcal{L}_{\text{cls}}^{\text{8cls}} + \mathcal{L}_{\text{reg}}^{\text{8cls}} + \mathcal{L}_{\text{iou}} \quad (10)$$

The overall loss of the improved Voxel R-CNN model is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rpn}} + \mathcal{L}_{\text{voxel}} + \mathcal{L}_{\text{rcnn}} \quad (11)$$

In this context, $\mathcal{L}_{\text{voxel}}$ is the sum of the Focal Loss values across the four voxel importance branches.

The complete training and inference procedure of the proposed method is summarized in Algorithm 1.

Algorithm 1. Training and Inference Procedure of the Proposed MFS-IAQH Detector

Input: LiDAR point cloud $\mathbf{P}$, RGB image $\mathbf{I}$, camera intrinsic matrix $\mathbf{K}$, LiDAR-to-camera transformation $[\mathbf{R}_{L \rightarrow C} \ \mathbf{t}_{L \rightarrow C}]$, and ground-truth 3D boxes $\mathbf{G}$ (training only).

Output: training loss L_{total} , or final 3D detections $\mathcal{D}$ during inference.

- 1 Voxelize $\mathbf{P}$ and obtain initial voxel features using MeanVFE.
 - 2 Apply LiDAR-driven VFE-LFS at conv1 and retain the top 50% voxels according to voxel importance.
 - 3 Extract the shallow image feature map from DeepLabV3-ResNet50 layer1 and reduce its channel dimension from 256 to 16 using a 1×1 channel-reduction layer, and align it to visible voxels using Equation (3).
 - 4 At the conv1 output, concatenate x_v^l and x_v^l for VFE selection; apply multimodal FSConv and fuse the outputs by Equation (4).
 - 5 Apply LiDAR-driven VFE-LFS at conv2 and conv3, followed by conventional sparse convolution at conv4.
 - 6 Obtain the final sparse 3D representation $\mathbf{F}_{3D} \in \mathbb{R}^{N \times 128}$ and map it to BEV.
 - 7 Generate 3D proposals by proposal NMS and retain the first-stage class c_i^{roi} for each RoI. compute L_{rpn} during training.
 - 8 Pool multiscale conv2–conv4 RoI features with a $6 \times 6 \times 6$ grid.
 - 9 IAQH predicts CQ, CSR, and IoUQ outputs.
 - 10 **if training then**
 - 11 Sample 128 RoIs (foreground ratio 0.5), compute u_i , and construct y_i with $u_{bg} = 0.25$ and $u_{fg} = 0.75$.
-

```

12   Optimize CQ, CSR for  $u_i > 0.55$ , and IoUQ using focal BCE ( $\alpha = 0.25, \gamma = 2$ ).
13   Compute  $L_{total} = L_{rpn} + L_{voxel} + L_{rcnn}$ ; update the network and return  $L_{total}$ .
14   else refine each RoI with the CSR channel indexed by  $c_i^{roi}$ , and compute  $s_i$  according to Equation (9).
15   Apply final NMS using  $s_i$  and return the detections  $\mathcal{D}$ .
16   end if

```

The key module-specific hyperparameters are specified in Algorithm 1, while the dataset-related and training hyperparameters are provided in Section 2.5.

2.4. Dataset and Evaluation Metrics

Experiments are conducted using the detection_v2.1 subset of the open-source Automine dataset, which is specifically designed for intelligent vehicle localization and perception tasks in unstructured environments such as open-pit mines. The dataset was acquired using BFS-U3-31S4C-C industrial cameras (Teledyne FLIR Integrated Imaging Solutions, Richmond, BC, Canada), an OS2-64 LiDAR (Ouster, Inc., San Francisco, CA, USA) on the SUV platform, and an HDL-32E LiDAR (Velodyne Lidar, Inc., San Jose, CA, USA) on the wide-body truck and mining truck platforms. The dataset covers multiple mining sites, different periods during daytime operation, and various environmental conditions, including dust and glare. Data cleaning and class merging are performed during preprocessing. Point cloud frames lacking intensity fields and samples belonging to the irrelevant bus class are removed. The Excavator/bottom and Excavator/top classes are merged into a single Excavator class. To reduce temporal redundancy between consecutive frames, uniform temporal subsampling with a stride of 2 is applied, i.e., one frame is retained from every two consecutive frames. To avoid further reducing the representation of minority classes, all frames containing Pedestrian, Excavator, or Tracked Transport Vehicle instances are retained without temporal subsampling. After preprocessing, a total of 2207 frames are retained. The cleaned dataset contains eight classes: wide-body dump trucks, mining trucks, pushdozers, civilian vehicles, water trucks, tracked transport vehicles, excavators, and pedestrians. The data are divided by scene into 1767 training frames, 220 validation frames, and 220 test frames to reduce information leakage caused by high similarity between temporally adjacent frames assigned to different subsets. The class-wise distribution of annotated object instances in the training, validation, and test sets is summarized in Table 1.

Table 1. Distribution of annotated object instances in the cleaned Automine dataset.

	Training	Validation	Test	Total
Wide-Truck	747	98	88	933
Mine-Truck	544	72	63	679
Pushdozer	288	33	38	359
Civil-Car	264	29	36	329
Water car	163	22	19	204
Track	136	14	20	170
Excavator	99	14	11	124
Pedestrian	40	4	6	50

Average Precision (AP) is used to evaluate object detection performance in both the bird's-eye-view (BEV) and 3D spaces. Considering the substantial differences in physical scale and point-cloud density among the eight mining object categories, IoU = 0.40 is retained as the primary threshold for multiclass evaluation. To further assess localization accuracy and avoid conclusions based on a single IoU criterion, BEV and 3D mAP are additionally evaluated at stricter IoU thresholds of 0.50, 0.60, and 0.70. The inference

frame rate, measured in frames per second (FPS), is defined as the reciprocal of the average processing time per frame and directly reflects the computational efficiency and real-time performance of the model in practical mining environments. Average Translation Error (ATE) is also employed to measure the 3D Euclidean distance between the centers of predicted and ground-truth bounding boxes. ATE is further evaluated within different distance intervals according to the distance from the ground-truth object center to the origin of the LiDAR coordinate system.

2.5. Experimental Setup

The experimental platform configuration is presented in Table 2.

Table 2. Experimental platform configuration.

Component	Specification
CPU	Intel Xeon Platinum 8470Q
Operating System	Ubuntu 20.04
GPU	NVIDIA vGPU-32GB × 1
CUDA	11.8
PyTorch	2.0.0
Python	3.8
GPU Memory	32 G

Experiments are conducted on the cleaned Automine dataset. The detection ranges are set to [0, 80] m along the X-axis, [−40, 40] m along the Y-axis, and [−3, 4] m along the Z-axis. A voxel size of [0.2, 0.2, 0.175] m is used along the three axes. To improve model generalization, several data augmentation operations are applied, including random cropping, rotation, flipping, color jittering, and Gaussian noise addition. Augmentations involving spatial geometric transformations are synchronously applied to images and point clouds, and the corresponding projection relationships are updated accordingly. Color jittering is applied only to the image branch.

For a fair comparison, all baseline methods were reproduced in our experimental environment. MSMD Fusion, MVXNet, Frustum PointNets, IS-Fusion, and MambaFusion were reimplemented according to their original architectures and recommended configurations, while CenterPoint-Fusion was constructed based on the CenterPoint framework by incorporating image information into the detection pipeline. The backbone configurations and pretrained weights of each baseline followed the default settings of the corresponding original method. To minimize the influence of differences in experimental settings, all baseline methods and the proposed method were trained and evaluated using the same cleaned Automine dataset, scene-level training/validation/test split, eight-class mapping, 0–80 m forward detection range, and evaluation scripts. The training hyperparameters were also kept consistent across all methods. Specifically, all models were trained for 100 epochs using the Adam optimizer with an initial learning rate of 0.003 and a batch size of 2. The same evaluation protocol and IoU thresholds were used for all methods when reporting the comparative results.

3. Results

3.1. Model Comparison Experiments

To evaluate the overall performance of the proposed method for intelligent driving perception in open-pit mines, comparisons are conducted with six representative multi-modal 3D object detection methods using the same dataset split, detection range, class mapping, and evaluation scripts. MSMD Fusion employs multi-depth seed back-

projection and multiscale cross-modal fusion. MVXNet performs early fusion by projecting image features onto point/voxel representations. CenterPoint-Fusion incorporates image information into a center-based 3D detection framework, whereas Frustum PointNets performs point-cloud detection within image-guided 3D frustums. IS-Fusion jointly models instance-level and scene-level multimodal information through instance–scene collaborative fusion. MambaFusion employs height-fidelity LiDAR encoding and Mamba-based dense global fusion to preserve geometric information while modeling long-range cross-modal dependencies. Together, these methods cover early voxel-level fusion, multiscale fusion, frustum-based fusion, center-based fusion, instance–scene collaborative fusion, and Mamba-based dense global fusion strategies.

As shown in Tables 3 and 4, the proposed method achieves BEV mAP@0.40 and 3D mAP@0.40 values of 63.43% and 61.23%, respectively, outperforming all comparison methods. Among the competing methods, MambaFusion achieves the highest overall mAP, with BEV and 3D mAP@0.40 values of 58.06% and 53.62%, respectively. Compared with MambaFusion, the proposed method improves BEV and 3D mAP by 5.37 and 7.61 percentage points, respectively. The larger improvement in 3D mAP indicates that the proposed method provides more accurate three-dimensional localization and bounding-box estimation in complex open-pit mine environments. To further examine whether the performance advantage depends on the selected IoU criterion, additional evaluations were conducted under increasingly strict IoU thresholds ranging from 0.40 to 0.70, as shown in Table 5.

Table 3. Model comparison results in terms of BEV AP@0.40.

Method	Wide-Truck	Mine-Truck	Pushdozer	Civil-Car	Water Car	Track	Excavator	Pedestrian
MSMDFusion [8]	62.57	72.31	61.95	28.68	24.84	9.48	16.41	1.00
MVXNet [6]	84.36	90.18	55.71	63.74	22.02	7.23	0	0
CenterPoint-Fusion [24]	80.85	88.90	79.85	36.19	32.43	12.58	22.16	1.00
Frustum PointNets [7]	99.82	89.73	76.32	84.84	49.32	28.32	0	0
IS-Fusion [13]	95.20	90.26	82.44	84.29	50.35	28.39	15.26	8.35
MambaFusion [17]	96.63	91.03	83.27	85.86	51.73	29.14	16.45	10.38
This work	97.45	91.56	87.21	88.83	54.67	35.20	28.84	23.71

Note: Bold values indicate the best performance in each column.

Table 4. Model comparison results in terms of 3D AP@0.40.

Method	Wide-Truck	Mine-Truck	Pushdozer	Civil-Car	Water Car	Track	Excavator	Pedestrian
MSMDFusion [8]	50.68	64.53	59.33	24.72	16.56	9.48	10.97	1.00
MVXNet [6]	83.20	88.67	53.24	61.31	21.57	7.23	0	0
CenterPoint-Fusion [24]	75.24	77.63	73.20	19.06	31.4	9.56	11.84	1.74
Frustum PointNets [7]	96.54	77.82	38.71	76.26	41.64	13.23	0	0
IS-Fusion [13]	93.35	86.27	76.75	77.54	43.22	21.00	12.87	7.00
MambaFusion [17]	94.07	87.07	78.59	79.04	44.52	22.51	14.58	8.55
This work	93.71	89.78	91.04	83.85	49.26	33.49	25.73	22.95

Note: Bold values indicate the best performance in each column.

Table 5. Comparison of BEV and 3D mAP under different IoU thresholds.

Method	BEV mAP				3D mAP			
	@0.40	@0.50	@0.60	@0.70	@0.40	@0.50	@0.60	@0.70
MSMDFusion [8]	34.66	32.72	28.86	17.42	29.66	29.27	16.87	7.85
MVXNet [6]	40.41	38.55	35.57	32.64	39.40	37.54	31.66	20.00
CenterPoint-Fusion [24]	44.25	44.07	36.28	25.07	37.46	37.42	25.38	10.27

Frustum PointNets [7]	53.54	51.35	46.77	32.99	43.03	40.63	30.85	15.10
IS-Fusion [13]	56.82	54.87	48.23	37.81	52.27	47.37	38.83	26.76
MambaFusion [17]	58.06	55.48	49.62	43.59	53.62	48.84	40.94	31.53
This work	63.43	58.01	54.45	48.74	61.23	54.06	49.47	41.36

Note: Bold values indicate the best performance in each column.

As shown in Table 5, the proposed method consistently outperforms all comparison methods across IoU thresholds ranging from 0.40 to 0.70. As the IoU criterion becomes stricter, the BEV mAP of the proposed method decreases from 63.43% at IoU = 0.40 to 48.74% at IoU = 0.70, while the 3D mAP decreases from 61.23% to 41.36%. Nevertheless, the proposed method maintains the highest BEV and 3D mAP at all evaluated thresholds. These results demonstrate that the performance advantage of the proposed method is maintained under increasingly strict localization criteria and is not dependent on the specific choice of IoU = 0.40.

As shown in Table 6, the ATE values of the proposed method in the 0–30 m, 30–60 m, and 60–80 m ranges are 0.15 m, 0.27 m, and 0.41 m, respectively, all lower than those of the comparison methods. Within the 60–80 m range, the proposed method achieves a Recall of 86.47%, exceeding the best competing Recall obtained by Frustum PointNets by 3.14 percentage points. Meanwhile, its ATE is reduced to 0.41 m, which is 0.12 m lower than that of MambaFusion, the strongest competing method in terms of long-range translation accuracy. These results indicate that the proposed method improves not only the center localization accuracy of matched objects at long ranges but also object detection coverage. This work reaches 27 FPS, which is second only to MVXNet (31 FPS) among the methods with available speed measurements and is substantially faster than IS-Fusion (12 FPS) and MambaFusion (14 FPS), demonstrating a favorable balance among detection accuracy, long-range stability, and real-time performance.

Table 6. Comparison of inference speed, distance-wise Recall, and ATE across different models.

Method	FPS	Recall (0–30 m)	ATE (0–30 m)	Recall (30–60 m)	ATE (30–60 m)	Recall (60–80 m)	ATE (60–80 m)
MSMDFusion [8]	26	67.86	0.43	64.32	0.51	53.33	0.77
MVXNet [6]	31	81.67	0.42	80.28	0.44	67.86	0.69
CenterPoint-Fusion [24]	25	77.83	0.41	75.00	0.48	64.29	0.79
Frustum PointNets [7]	N/A	92.86	0.29	86.38	0.28	83.33	0.71
IS-Fusion [13]	12	88.29	0.22	84.73	0.31	78.84	0.57
MambaFusion [17]	14	89.93	0.26	85.62	0.29	80.37	0.53
This work	27	94.76	0.15	91.25	0.27	86.47	0.41

Note: Bold values indicate the best performance in each column.

To further evaluate the robustness of the proposed method under adverse environmental conditions, test samples were manually grouped according to the presence of strong illumination with high-contrast shadows and dust. Strong illumination and shadows were considered jointly because they frequently co-occur in open-pit mine scenes. As shown in Table 7, the proposed method achieves the highest BEV and 3D mAP under both challenging conditions. Compared with MambaFusion, the strongest multimodal baseline in this experiment, the proposed method improves BEV/3D mAP by 0.35/0.08 percentage points under strong illumination and shadows and by 2.03/2.49 percentage points under dusty conditions. The larger gains under dust suggest that the proposed shallow visual injection and LiDAR-driven subsequent processing are particularly beneficial when visual information is degraded.

Table 7. Robustness comparison under challenging environmental conditions.

Method	Strong Illumination & Shadows		Dust	
	BEV mAP@0.40	3D mAP@0.40	BEV mAP@0.40	3D mAP@0.40
Voxel R-CNN [23]	42.57	41.87	37.19	35.22
Frustum PointNets [7]	45.18	43.95	35.54	34.74
IS-Fusion [13]	45.55	44.20	39.20	38.05
MambaFusion [17]	45.90	44.35	40.35	39.24
This work	46.25	44.43	42.38	41.73

Note: Bold values indicate the best performance in each column.

The detection performance of the proposed model on normal transportation roads in open-pit mines is illustrated in Figure 5.



Figure 5. Visualization of the detection results of the model described in this paper. (a) Original Frame. (b) This work. In (b), green boxes denote the ground-truth bounding boxes, while red boxes denote the predicted bounding boxes.

The medium-range (30–60 m) and long-range (60–80 m) object detection results are shown in Figure 6. As the object distance increases, the point cloud becomes progressively sparser, and the comparison model exhibits missed detections and localization errors. Through visual information supplementation and voxel importance enhancement, the proposed model preserves stronger responses to distant objects in these samples and produces more complete 3D bounding boxes.

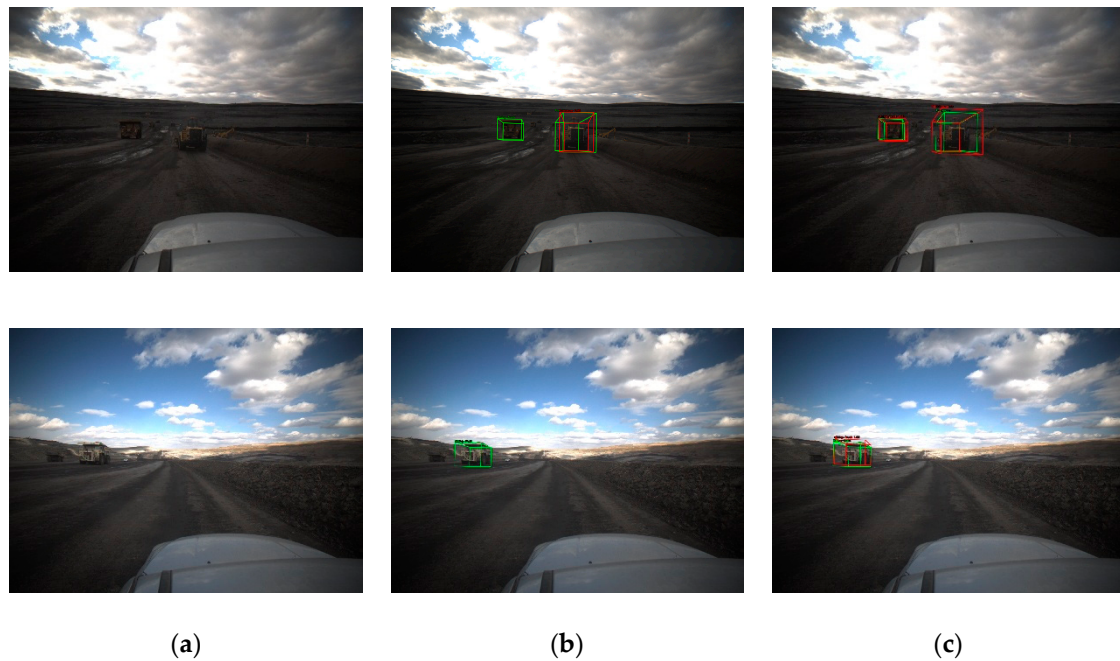


Figure 6. Comparison of Detection Results from Different Models in Medium- and Long-Range Scenarios. (a) Original Frame. (b) MambaFusion. (c) This work. In (b,c), green boxes denote the ground-truth bounding boxes, while red boxes denote the predicted bounding boxes.

3.2. Model Ablation Experiments

To evaluate the effectiveness and statistical reliability of the proposed algorithms and their constituent mechanisms, Voxel R-CNN is used as the baseline. The main ablation experiments for MFS and IAQH are independently repeated three times using different random seeds, while the dataset split, training parameters, and evaluation settings are kept unchanged. The results are reported as the mean $\pm$ standard deviation. Since the main comparison experiments report results under a fixed-seed setting, slight numerical differences in the complete-model results are expected between the main comparison and the multi-run ablation experiments. Fine-grained ablation experiments are further performed on the LFS propagation mechanism, VSI enhancement strategy, and VFE selection mechanism within MFS. In these experiments, when VFE is removed, LFS performs neighborhood propagation for all non-empty voxels at the current scale. After VFE is introduced, selective expansion is performed only for voxels with relatively high importance scores. The ablation results are presented in Tables 8 and 9.

Table 8. Model ablation results over three independent runs (mean $\pm$ standard deviation).

Baseline	MFS	IAQH	BEV mAP@0.40	3D mAP@0.40	FPS	ATE (0–60 m)	ATE (60–80 m)
✓			54.15 $\pm$ 0.24	54.12 $\pm$ 0.22	34	0.31 $\pm$ 0.08	0.67 $\pm$ 0.15
✓	✓		59.71 $\pm$ 0.33	55.39 $\pm$ 0.32	30	0.30 $\pm$ 0.11	0.61 $\pm$ 0.13
✓		✓	58.65 $\pm$ 0.29	60.32 $\pm$ 0.18	28	0.26 $\pm$ 0.04	0.46 $\pm$ 0.09
✓	✓	✓	63.17 $\pm$ 0.45	61.52 $\pm$ 0.30	27	0.23 $\pm$ 0.09	0.39 $\pm$ 0.06

Note: “✓” indicates that the corresponding component is enabled in the model configuration. Bold values indicate the best performance in each column.

As shown in Table 8, the baseline Voxel R-CNN achieves a BEV mAP@0.40 of 54.15 $\pm$ 0.24% and a 3D mAP@0.40 of 54.12 $\pm$ 0.22%. Introducing MFS increases the BEV mAP to 59.71 $\pm$ 0.33%, corresponding to an improvement of 5.56 percentage points, while the

3D mAP increases to $55.39 \pm 0.32\%$. This result demonstrates that MFS effectively enhances the feature representation of sparse objects through visually assisted voxel importance estimation and selective sparse propagation. In comparison, IAQH provides a more pronounced improvement in 3D detection, increasing the 3D mAP to $60.32 \pm 0.18\%$, an improvement of 6.20 percentage points over the baseline. Meanwhile, the ATE is reduced from 0.31 ± 0.08 m to 0.26 ± 0.04 m within 0–60 m and from 0.67 ± 0.15 m to 0.46 ± 0.09 m within 60–80 m, indicating that IAQH improves proposal localization quality, particularly for long-range objects.

When MFS and IAQH are jointly incorporated, the complete model achieves the best overall performance, with a BEV mAP@0.40 of $63.17 \pm 0.45\%$ and a 3D mAP@0.40 of $61.52 \pm 0.30\%$, representing improvements of 9.02 and 7.40 percentage points over the baseline, respectively. The ATE is further reduced to 0.23 ± 0.09 m within 0–60 m and 0.39 ± 0.06 m within 60–80 m. Across the three independent runs, the standard deviations of BEV and 3D mAP remain relatively small compared with the corresponding improvements over the baseline, indicating that the observed performance gains are consistently maintained across different random seeds. These results also demonstrate the complementary roles of MFS in sparse feature enhancement and IAQH in proposal localization and quality refinement.

To further investigate the contributions of the individual components within MFS, fine-grained ablation experiments are conducted on LFS, VSI, and VFE, as shown in Table 9.

Table 9. Fine-grained ablation experiments on MFS.

Baseline	LFS	VSI	VFE	BEV mAP@0.40	3D mAP@0.40	FPS	ATE (0–60 m)	ATE (60–80 m)
√				54.34	54.25	34	0.32	0.63
√	√			55.86	54.46	32	0.33	0.64
√		√		57.23	54.88	32	0.32	0.61
√	√	√		58.72	55.20	31	0.31	0.58
√	√	√	√	59.66	55.62	30	0.31	0.57

Note: “√” indicates that the corresponding component is enabled in the model configuration. Bold values indicate the best performance in each column.

As shown in Table 9, the improvement in detection accuracy is limited when LFS is used alone, and ATE increases slightly, indicating that unconstrained neighborhood propagation may introduce background interference. After VSI is introduced, visual edge, contour, and texture information supplements the semantic information of sparse point clouds. With the further introduction of VFE, the model selects expansion locations based on voxel importance prediction, allowing LFS propagation to focus on potential foreground regions. The complete MFS achieves a BEV mAP of 59.66% and a 3D mAP of 55.62%, demonstrating the complementary effects of LFS, VSI, and VFE.

To further investigate the sensitivity of VFE to the voxel selection ratio and justify the default setting, additional experiments were conducted by varying the proportion of voxels selected for neighborhood propagation. The complete MFS configuration was retained, while only the VFE selection ratio was changed. IAQH was excluded to isolate the effect of the VFE selection mechanism. BEV and 3D mAP are reported as mean $\pm$ standard deviation over three independent runs with different random seeds, whereas Recall, ATE, and FPS are reported under the fixed-seed controlled setting. The results are presented in Table 10.

Table 10. Ablation study on the VFE voxel selection ratio.

VFE Selection Ratio	BEV mAP@0.40	3D mAP@0.40	Recall (60–80 m)	ATE (60–80 m)	FPS
25%	59.18 ± 0.31	55.34 ± 0.27	81.23	0.59	31
50%	59.71 ± 0.33	55.39 ± 0.32	83.40	0.57	30
75%	59.45 ± 0.37	55.51 ± 0.31	83.25	0.52	29
100%	59.08 ± 0.41	55.29 ± 0.33	82.51	0.61	29

As shown in Table 10, the detection performance first improves as the VFE voxel selection ratio increases from 25% to 50%, and then tends to saturate with only minor fluctuations at higher ratios. A relatively low selection ratio may restrict the propagation of potentially informative foreground features, resulting in a lower long-range Recall. In contrast, when the selection ratio is further increased, more low-importance voxels participate in neighborhood propagation, which weakens the foreground selectivity of VFE and may introduce additional background responses and redundant feature propagation. Consequently, increasing the selection ratio beyond 50% does not lead to consistent improvements in BEV/3D mAP or long-range Recall, while the computational cost slightly increases. Although the 75% setting achieves the lowest ATE in the 60–80 m range, this improvement is not accompanied by consistent gains in the other detection metrics. Therefore, considering detection accuracy, long-range Recall, and computational efficiency together, a VFE voxel selection ratio of 50% is adopted in the proposed model.

To further investigate the contributions of the individual components within IAQH, component-wise ablation experiments are conducted on class-specific regression (CSR), class-conditional quality (CQ), and class-agnostic IoU quality (IoUQ). Since CQ relies on the class-specific refinement structure, its contribution is evaluated on top of CSR rather than independently. IoUQ is separately introduced on top of CSR to evaluate the effect of class-agnostic localization quality estimation. MFS is enabled for all configurations, with the VFE selection ratio fixed at 50%. The results are presented in Table 11.

Table 11. Component-wise ablation study of IAQH.

MFS	CSR	CQ	IoUQ	BEV mAP@0.40	3D mAP@0.40	Recall (60–80 m)	ATE (60–80 m)	FPS
√				59.71 ± 0.33	55.39 ± 0.32	83.40	0.57	30
√	√			60.65 ± 0.27	57.25 ± 0.27	82.92	0.58	29
√	√	√		61.82 ± 0.36	58.38 ± 0.29	84.65	0.55	29
√	√		√	61.48 ± 0.14	58.92 ± 0.28	85.23	0.51	28
√	√	√	√	63.17 ± 0.45	61.52 ± 0.30	86.47	0.41	27

Note: “√” indicates that the corresponding component is enabled in the model configuration. Bold values indicate the best performance in each column.

As shown in Table 11, introducing CSR increases the BEV mAP@0.40 from 59.71% to 60.65% and the 3D mAP@0.40 from 55.39% to 57.25%, demonstrating that class-specific geometric modeling improves the overall 3D detection performance. Adding CQ on top of CSR further increases the BEV and 3D mAP to 61.82% and 58.38%, respectively, while the long-range Recall improves to 84.65%, indicating the benefit of class-conditioned proposal quality estimation. When IoUQ is introduced on top of CSR, the 3D mAP reaches 58.92%, accompanied by a higher long-range Recall of 85.23% and a lower ATE of 0.51 m, demonstrating the effectiveness of class-agnostic localization quality estimation for proposal ranking. When CSR, CQ, and IoUQ are jointly incorporated, the complete IAQH achieves the best overall performance, with BEV mAP@0.40 and 3D mAP@0.40 reaching 63.17% and 61.52%, respectively. Meanwhile, the long-range Recall increases to 86.47%

and the ATE decreases to 0.41 m. These results demonstrate that CSR, CQ, and IoUQ provide complementary benefits in class-specific geometric refinement, class-conditioned quality estimation, and localization-aware proposal ranking.

4. Discussion

In terms of class-specific results, the proposed method achieves the highest detection accuracy for seven of the eight object classes, with the exception of wide-body dump trucks. Particularly notable improvements are observed for pushdozers, tracked transport vehicles, excavators, and pedestrians, which typically exhibit sparse point clouds, weak local features, or substantial geometric differences. MFS uses shallow visual information to assist voxel importance estimation and strengthens potential foreground responses through FSConv, whereas IAQH improves bounding box geometry and ranking through class-specific regression and dual-quality probability calibration. For Wide-Truck, the proposed method achieves BEV AP@0.40 and 3D AP@0.40 values of 97.45% and 93.71%, respectively. Frustum PointNets achieves the highest AP for this category, while MambaFusion also slightly exceeds the proposed method in 3D AP (94.07% vs. 93.71%). Nevertheless, the proposed method achieves the best performance for the remaining seven categories, with particularly clear advantages for Pushdozer, Track, Excavator, and Pedestrian.

To further investigate the limitations of the proposed model, Figure 7 presents several failure cases, focusing on challenging scenarios involving small and weak objects.



Figure 7. Failed Samples in Weak and Small Object Detection. (a) Original Frame. (b) This Work. In (b), green boxes denote the ground-truth bounding boxes, while red boxes denote the predicted bounding boxes.

Among these failure cases, pedestrians, as typical small and weak objects, pose a substantial challenge to the detection task. Pedestrian objects contain only a very small number of LiDAR points, making it difficult to extract sufficient features. Neither the proposed model nor comparison methods such as MVXNet and Frustum PointNets can accurately identify the pedestrian class in these cases. The proposed model and some comparison

methods exhibit different degrees of missed detections and bounding box offsets, indicating that existing models still face substantial performance bottlenecks when extreme point cloud sparsity, interference from nearby objects, and environmental noise coexist. Future improvements may therefore focus on extremely sparse point cloud enhancement, suppression of multi-object interference, and robust multimodal fusion under sensor degradation.

5. Conclusions

This study proposes a Voxel R-CNN-based camera–LiDAR multimodal 3D object detection method to address long-range point cloud sparsity, complex background interference, and inconsistency between 3D proposal classification confidence and localization quality in intelligent driving scenarios in open-pit mines. The proposed Multimodal Focal Sparse Voxel Enhancement (MFS) algorithm improves sparse-object representation and suppresses irrelevant background expansion through voxel importance prediction, selective neighborhood propagation, and shallow visually conditioned enhancement. Furthermore, the proposed IoU-Aware Quality and Geometry Refinement Head (IAQH) performs proposal quality estimation, class-specific geometric regression, and ranking calibration. MFS and IAQH operate on first-stage feature enhancement and second-stage proposal refinement, respectively, providing an effective approach to multimodal fusion and 3D object detection in unstructured mining environments.

Under conditions involving extremely sparse point clouds, severe occlusion, interference from nearby objects, and sensor degradation, the model still exhibits missed detections and bounding box offsets. Future work will focus on feature enhancement for extremely sparse point clouds, suppression of multi-object interference, and robust multimodal fusion under sensor degradation to further improve model stability and adaptability in complex open-pit mine operating environments.

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

Funding: This research was funded by the National Natural Science Foundation of China under the project “UWB Radar Life Information Feature Extraction and Quantitative Identification for Mine Drill Hole Rescue” (grant number 52174197).

Data Availability Statement: The Automine dataset (detection_v2.1 subset) used in this study is available upon reasonable request from the dataset owner via the application page: <https://automine.cc/download/> (accessed on 9 September 2026). Restrictions apply to the availability of this dataset, which was used under license for this research.

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

References

1. He, X.; Wang, Z.; Lin, J.; Nai, K.; Yuan, J.; Li, Z. DO-SA&R: Distant Object Augmented Set Abstraction and Regression for Point-Based 3D Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 5852–5864.
2. Lai, X.; Chen, Y.; Lu, F.; Liu, J.; Jia, J. Spherical Transformer for LiDAR-Based 3D Recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 17545–17555.
3. Lu, Y.; Hao, X.; Li, Y.; Chai, W.; Sun, S. Range-Aware Attention Network for LiDAR-Based 3D Object Detection With Auxiliary Point Density Level Estimation. *IEEE Trans. Veh. Technol.* **2025**, *74*, 292–305.

4. Xie, Z.; Zhao, Y.; Zhang, F.; Luo, B.; Hu, W. Mine-SSD: Dual-threshold set abstraction and radius-adaptive grouping for 3D object detection in open-pit mines. *Neurocomputing* **2025**, *657*, 131507.
5. Zhang, H.; Xie, Z.; Zhang, F.; Zhao, Y. Multi-scale hierarchical voxel-aware transformer network for 3D object detection in open-pit mines. *Expert Syst. With Appl.* **2026**, *300*, 130467.
6. Sindagi, V.A.; Zhou, Y.; Tuzel, O. MVX-Net: Multimodal VoxelNet for 3D Object Detection. In *2019 International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2019; pp. 7276–7282.
7. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 918–927.
8. Yang, J.; Jie, Z.; Chen, S.; Chen, J.; Ma, L. MSMD Fusion: Fusing LiDAR and Camera at Multiple Scales with Multi-Depth Seeds for 3D Object Detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 21643–21652.
9. Xie, Y.; Xu, C.; Rakotosaona, M.-J.; Rim, P.; Tombari, F. SparseFusion: Fusing Multi-Modal Sparse Representations for Multi-Sensor 3D Object Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 17545–17556.
10. Qin, Y.; Wang, C.; Kang, Z.; Ma, N.; Li, Z. SupFusion: Supervised LiDAR-Camera Fusion for 3D Object Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 21957–21967.
11. Yan, J.; Liu, Y.; Sun, J.; Jia, F.; Li, S. Cross Modal Transformer: Towards Fast and Robust 3D Object Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 18222–18232.
12. Li, X.; Fan, B.; Tian, J.; Fan, H. GAFusion: Adaptive Fusing LiDAR and Camera with Multiple Guidance for 3D Object Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 21209–21218.
13. Yin, J.; Shen, J.; Chen, R.; Li, W.; Yang, R. IS-Fusion: Instance-Scene Collaborative Fusion for Multimodal 3D Object Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 14905–14915.
14. Gunn, J.; Lenyk, Z.; Sharma, A.; Donati, A.; Buburuzan, A. Lift-Attend-Splat: Bird’s-Eye-View Camera-LiDAR Fusion Using Transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2024; pp. 4526–4536.
15. Cai, H.; Yin, D.; Yu, F.R.; Xiong, S. DSTR: Dual Scenes Transformer for Cross-Modal Fusion in 3D Object Detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2025; pp. 3064–3073.
16. Lee, Y.; Jeong, H.M.; Jeon, Y.; Kim, S. EVT: Efficient View Transformation for Multi-Modal 3D Object Detection. In *2025 IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 26632–26642.
17. Wang, H.; Gao, J.; Hu, W.; Zhang, Z. Height-Fidelity Dense Global Fusion for Multi-Modal 3D Object Detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2025; pp. 26664–26674.
18. Park, K.; Kim, Y.; Kim, D.; Choi, J.W. Resilient Sensor Fusion under Adverse Sensor Failures via Multi-Modal Expert Fusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 6720–6729.
19. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of Localization Confidence for Accurate Object Detection. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 816–832.
20. Li, J.; Luo, S.; Zhu, Z.; Dai, H.; Krylov, A.S. 3D IoU-Net: IoU Guided 3D Object Detector for Point Clouds. *arXiv* **2020**, arXiv:2004.04962.
21. Zheng, W.; Tang, W.; Chen, S.; Jiang, L.; Fu, C.-W. CIA-SSD: Confident IoU-Aware Single-Stage Object Detector From Point Cloud. In *AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2021; Volume 35, pp. 3555–3562.
22. Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; Jia, J. Focal Sparse Convolutional Networks for 3D Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 5418–5427.
23. Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y. Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection. In *AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2021; Volume 35, pp. 1201–1209.
24. Yin, T.; Zhou, X.; Krähenbühl, P. Center-based 3D Object Detection and Tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 11784–11793.

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.