

Article

Attention-Augmented PointPillars for Enhanced Mining Personnel Detection

Pingan Peng ^{1,2} , Chaowei Zhang ^{1,*}  and Bing Cui ³

¹ School of Resources and Safety Engineering, Central South University, Changsha 410083, China; ping_an@csu.edu.cn

² State Key Laboratory for Fine Exploration and Intelligent Development of Coal Resources, Xuzhou 221116, China

³ Norinmining Co., Ltd., Beijing 100055, China; cuibing@norinmining.com

* Correspondence: 235511057@csu.edu.cn

Featured Application

This methodology has direct application in enhancing real-time personnel safety monitoring in underground mining operations through LiDAR-based 3D perception. By enabling robust, tag-free detection of workers in complex tunnel environments—characterized by poor lighting, dust, occlusions, and blind zones—the proposed Attention-Augmented PointPillars system provides critical situational awareness for load-haul-dump (LHD) machinery and other mobile equipment. The approach supports the deployment of intelligent collision-avoidance systems that operate reliably even when visual or radio-frequency sensing fails, thereby significantly reducing human-machine accident rates and advancing autonomous safety protocols in hazardous subterranean workplaces.

Abstract

The intricate layout of underground mine tunnels and the operation of large-scale mining equipment create extensive blind zones, leading to an average of 17 personnel collision accidents per 100,000 working hours in China's metal mines. To tackle this issue, we constructed a specialized Jinchuan Underground Mining Personnel Dataset covering intersecting tunnels and long straight tunnels, with precise bounding box annotations for personnel locations under varying illumination and dust conditions. We propose the Attention-Augmented PointPillars for Enhanced Mining Personnel Detection. Incorporating Recursive Gated Convolutions into the feature extraction network enables long-range modeling and higher-order spatial interactions. Moreover, the pyramidal design in g^n Conv with channel width gradually increasing during spatial interactions enhances the model's efficiency in processing complex spatial information. Additionally, a Channel and Spatial Attention module integrating spatial and channel attention feature fusion strengthens feature expression via multiple weighting mechanisms. Field tests in Jinchuan underground mine show optimal performance with a batch size of 8, a learning rate of 0.003, and a spatial interaction order of 5, achieving 3% higher accuracy than the original network. Furthermore, comparisons with mainstream methods on the Underground Personnel Dataset confirm our method's state-of-the-art performance.

Keywords: mine; point cloud; personnel detection; PointPillars



Academic Editor: Andrea Prati

Received: 6 January 2026

Revised: 30 January 2026

Accepted: 9 February 2026

Published: 11 February 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](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Mineral resources provide essential materials and energy for the advancement of human society, playing a pivotal role in national economic vitality [1,2]. As a crucial piece of mining equipment for underground ore extraction, the scraper primarily handles ore loading, transportation, and unloading within the confined spaces of subterranean environments [3]. However, as shown in Figure 1a, underground tunnels typically form intricate networks with numerous forks and bends, obstructing the scraper operator's line of sight [4]. Moreover, as illustrated in Figure 1b, the underground scraper's substantial size and wide bucket further restrict visibility, increasing the risk of collisions with nearby personnel. Due to the aforementioned reasons, personnel collisions frequently occur in mine tunnels, resulting in numerous casualties [5–7]. Therefore, solving the issue of determining the position and distance perception of surrounding personnel during the shoveling machine's transport process is crucial. Addressing this problem can significantly reduce the danger of the operation and effectively lower the incidence of man-machine collision accidents.

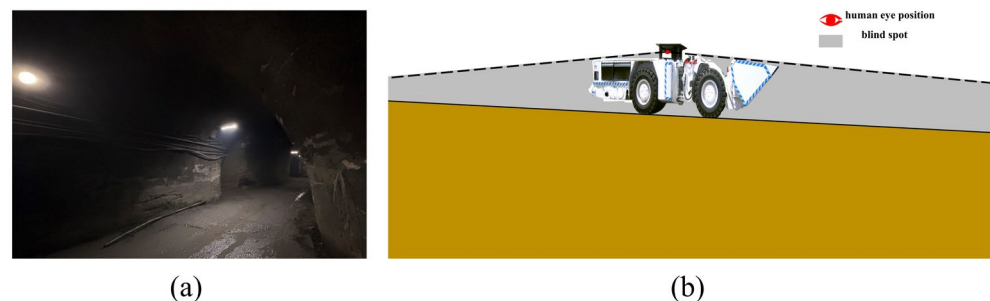


Figure 1. (a) The intricate and complex network of underground mine tunnels. (b) The visibility issues of the scraper, with the gray areas representing the blind spots caused by the scraper's large size.

To achieve the above objectives, LHD machines must detect nearby pedestrians and respond promptly during tunnel operations, as illustrated in Figure 2. Personnel detection around mining transport equipment is broadly categorized into tag-based approaches such as Global Positioning System (GPS), Radio Frequency Identification (RFID), and Ultra Wide Band (UWB), and non-tag-based approaches including cameras and LiDAR, with selection guided by operational conditions. GPS works well in open-pit mines but fails underground due to signal attenuation. RFID is widely used underground but suffers from concurrent tag collisions: dense tags overwhelm readers, causing misreads, reduced throughput, and bandwidth waste [8,9]; detection range also varies with antenna–tag orientation [10,11]. UWB enables precise localization but is highly vulnerable to multipath and occlusion in metal-rich tunnels [12,13]. Although tag-based systems offer high accuracy, they require personnel to wear tags. In practice, tags are often forgotten or missing; excessive tags induce interference, severely compromising reliability and efficiency.

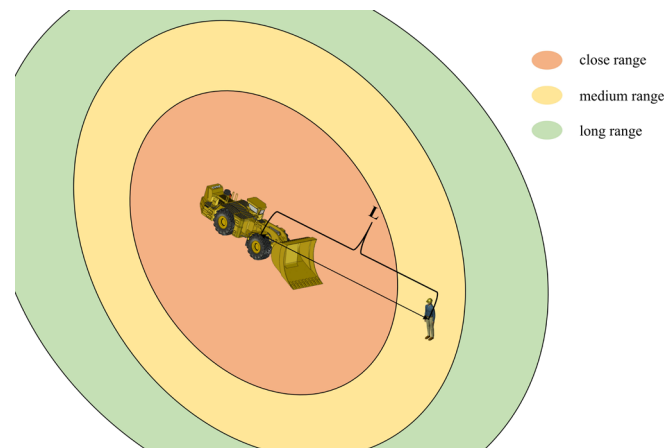


Figure 2. The scraper has different danger distances. L represents the distance from the center of the scraper to the worker, which is categorized into three ranges: short-medium range, far range, and very far range.

In recent years, deep learning has advanced vision-based personnel detection in mining. D. Xiao et al. [14] proposed a compact residual YOLOv3 to detect obstacles during haul truck transit, reducing blind-spot accidents. L. Wang et al. [15] enhanced pedestrian detection in dense underground crowds via YWSSv1/v2, integrating YOLOv2's skip connections with fully convolutional networks. However, uneven lighting, dust, and noise cause high false positives. To improve robustness, Wei X et al. [16] introduced PftNet for adverse conditions, and S. Guan et al. [17] adapted RefineDNet for dehazing in dusty mines. Thermal infrared imaging offers a dust-immune alternative: human radiation in the 7–14 μm band penetrates particulates. X. Li et al. [18] and J. S. Dickens et al. [19] applied it to personnel tracking in tunnels, yet nonuniform temperatures—due to depth, ventilation, and operations—blur target-background distinctions, lowering reliability. Other works addressed degradation: P. Xu et al. [20] found wavelet transform most effective for tunnel imagery; X. Zhou et al. [21] improved robustness via highlight/shadow suppression from scene data mining. However, these enhancements incur high computational costs, compromising real-time performance and collision avoidance. Overall, visible-light systems remain limited by dust, uneven illumination, vibration, and sensitivity to lighting, dust, and camera angles [22,23], degrading image quality and constraining real-time accuracy and deployment.

Recent advances in LiDAR technology have enabled scanners to measure distances to the surrounding environment and directly generate sparse three-dimensional point clouds. Target detection using such point cloud data effectively mitigates issues commonly associated with conventional cameras. LiDAR-based point cloud sensing algorithms acquire 3D point clouds of objects via LiDAR sensors and process them to determine target position and category. Compared with camera-captured images, LiDAR-generated point clouds are less affected by lighting conditions and dust. Point cloud detection methods are broadly divided into traditional approaches and deep learning-based approaches.

Traditional point cloud object detection involves four stages: preprocessing, feature extraction, clustering, and classification. Preprocessing removes ground points to reduce noise; feature extraction relies on handcrafted strategies to group similar points. For example, Yokoyama H. et al. [24] used k-nearest neighbor clustering; Zermas D. et al. [25] applied multi-planar fitting for ground removal and raster-wise line-by-line clustering; Nan Z. et al. [26] introduced a density-based function to adaptively adjust Euclidean thresholds; Wang K. et al. [27] developed a Bayesian-optimized KNN model for improved multi-category classification. With deep learning advances, 3D detection has gained attention [28].

Due to the unstructured nature of point clouds, 2D convolutions are ineffective. Qi C. R. et al. [29] proposed PointNet, using fully connected layers for per-point features and max pooling for global representation; PointNet++ [30] extended this with hierarchical set abstraction for multi-scale local features. Zhou Y. et al. [31] presented VoxelNet, which voxelizes point clouds, encodes intra-voxel points, applies 3D convolutions, and uses a region proposal network—yet suffers from high computational costs. Yan Y. et al. [32] addressed this with SECOND, employing sparse convolutions that operate only on non-empty voxels to accelerate inference. Shi S. et al. [33] introduced Point RCNN, generating initial proposals via foreground segmentation and refining them in a canonical coordinate system with semantic features. Z. Yang et al. [34] preserved foreground points in 3DSSD using farthest feature distance sampling. Shi S. et al. [35] combined 3D CNN proposals and PointNet’s receptive field in PV-RCNN, mapping multi-scale key points to ROI grids for refinement. Yin T. et al. [36] predicted center points instead of regressing azimuth directly, refining with auxiliary features; Zhe H. et al. [37] enhanced set abstraction by re-scoring foreground and boundary points and using a geometry-aware template for better semantic cues.

Compared with other 3D object detection algorithms, the original PointPillars demonstrates significant advantages in inference speed [38]. Its core innovation lies in encoding sparse point clouds into a pillar structure: by converting 3D spatial data into pseudo-2D images, it drastically reduces computational complexity, enabling high-speed inference while maintaining acceptable accuracy—critical for real-time obstacle avoidance in mining equipment requiring immediate response. Specifically, the algorithm extracts local features within each pillar using the Pillar Feature Network (PFN), then applies a 2D convolutional neural network to learn global features from the pseudo-images, completing detection through a streamlined pipeline. This design greatly alleviates the computational burden of 3D convolutions, achieving real-time performance with reasonable accuracy, and outperforms both voxelization-based and direct point cloud processing methods, well satisfying mining real-time requirements. To improve detection accuracy in tunnel environments, TANet [39] introduces an attention mechanism during feature extraction and splits FPN RPN into coarse and fine stages, effectively addressing occlusion. Inspired by TANet and CBAM [40], we designed a Channel and Spatial Attention (CASA) module that generates weight masks after feature extraction to assign varying importance to different spatial locations, yielding finer features and enhancing classification and regression accuracy. Furthermore, integrating Recursive Gated Convolutions (g^n Conv) [41] into the feature extraction network captures higher-order spatial dependencies and leverages spatial context to improve feature representation and prediction accuracy.

In summary, this work focuses on the task of 3D point cloud personnel detection in a tunnel environment, enhances the network to adapt to the uniqueness of the underground setting, and improves detection performance in both cross-tunnel and long-tunnel scenarios. Compared with the original algorithm, the proposed method achieves greater progress in recognition accuracy, lays a solid foundation for subsequent underground personnel perception tasks, and can significantly reduce the occurrence of underground collision accidents.

Building on PointPillars and considering the characteristics of the unstructured environment in underground mines, this work proposes the APEM (Attention-Augmented PointPillars for Enhanced Mining Personnel Detection) algorithm. The main contributions are as follows. Firstly, the integration of g^n Conv into feature extraction networks aims to enhance the spatial interaction capabilities of the PointPillars algorithm through multiscale higher-order information interactions. This enhancement extends self-attentive second-order interactions to arbitrary orders by integrating recursive gated convolutions,

providing high flexibility and customizability without introducing a significant computational burden. Additionally, we designed the CASA module and incorporated it into the PointPillars network to obtain richer global features by integrating spatial and channel attention, thus enhancing the accuracy of classification and regression. Finally, we introduced the point cloud-based personnel detection task into the mining environment, overcoming the limitations of traditional vision-based personnel detection methods in such scenarios and effectively reducing the occurrence of underground personnel collision accidents. Under extreme mining conditions, such as post-blasting dust storms, low-illumination tunnels, or high-humidity zones, camera-based detectors frequently lose visibility, while radio frequency signals suffer from attenuation or metal-induced multipath distortion. In such cases, LiDAR remains the only consistently operational sensing modality. Our system leverages this advantage to provide minimal yet critical perception capability when other sensing channels are compromised. Even with reduced detection accuracy, its ability to identify potential human presence in equipment blind spots enables timely intervention, directly contributing to accident prevention during high-risk operational periods.

2. Approach

APEM (Attention-Augmented PointPillars for Enhanced Mining Personnel Detection) presents an enhanced 3D point cloud personnel detection model tailored for unstructured conditions of underground mines. PointPillars serves as the baseline due to its utilization of a 3D point cloud for input, enabling end-to-end learning. Moreover, its swift and accurate inference makes it applicable across various real-world scenarios. Given the need for high real-time detection in the unstructured underground mine environment, this model selection effectively meets the specific demands of such settings.

The algorithmic framework of APEM is depicted in Figure 3, segmented primarily into the following components: Pillar Feature Net, 2D Backbone, CASA Module, and Detection Head. (1) The Pillar Feature Net converts the original sparse point cloud information into pillars, generating pseudo-2D images. (2) The 2D Backbone utilizes the Feature Pyramid Networks (FPN) structure [42], enabling continuous down-sampling and up-sampling of pseudo-2D images, ultimately stitching them together to produce 2D feature maps. (3) The CASA Module integrates the produced 2D feature maps through spatial attention and channel attention mechanisms, resulting in the output of an attention feature map. (4) The Detection Head (SSD) is responsible for generating size and position information for the three-dimensional detection frame from the detection feature map and mapping it back to the original point cloud.

The primary distinction between this algorithm and the original lies in the adaptations made to the 2D Backbone and CASA Module to suit the unique conditions of underground mines. The detailed structure of each component of the algorithm will be delineated below.

2.1. Pillar Feature Net

The APEM algorithm operates on pure point cloud data as input. Due to the inherent disorderliness and sparsity of the original point cloud, most pillars within it are empty, with only a sparse distribution of points in the non-empty pillars. Through PFN processing, the initial sparse point cloud is transformed into a pseudo-two-dimensional image rich in information density. This process is visually depicted in Figure 4.

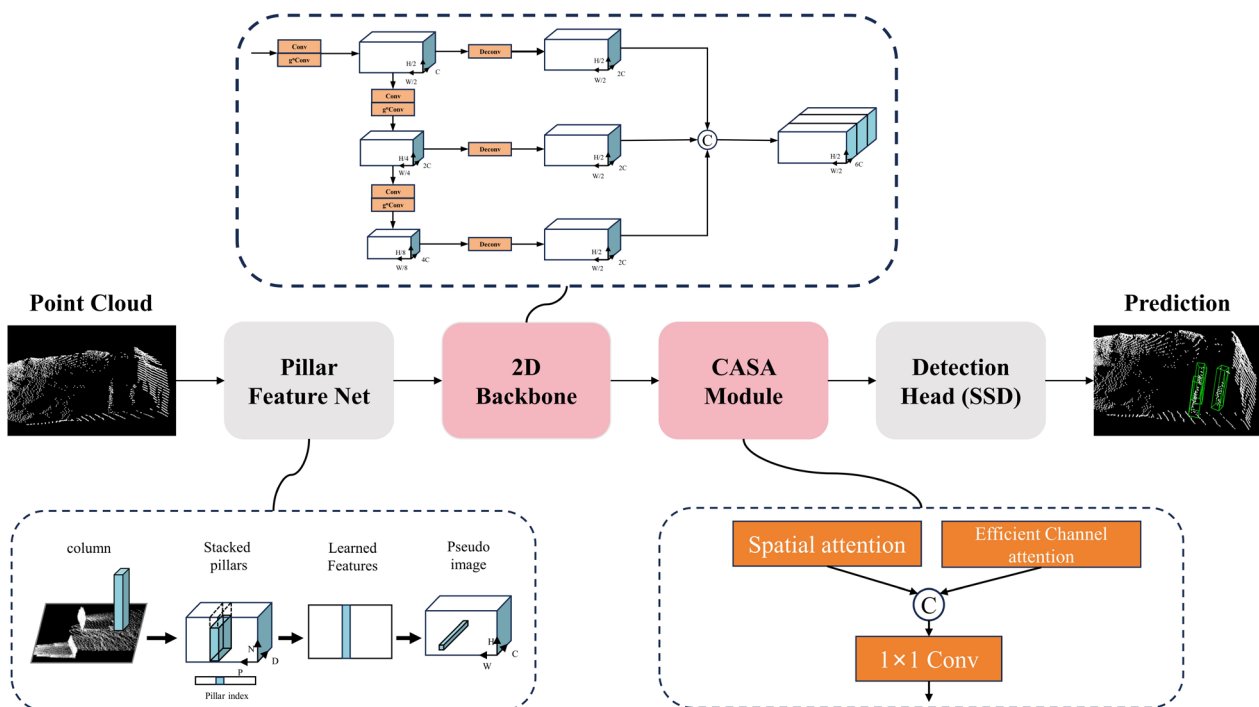


Figure 3. Overview of the framework. The framework provides an overview structured around four main modules: Pillar Feature Net, 2D Backbone, CASA Module, and Detection Head, visually represented by arrows in the workflow diagram. Initially, the sparse point cloud undergoes processing in PFN, transforming into a dense pseudo-2D map. Subsequently, the 2D Backbone and CASA Module are utilized to extract both spatial and channel feature information from the point cloud adaptively. Finally, the Detection Head predicts the 3D bounding box of the human using the features derived from the backbone.

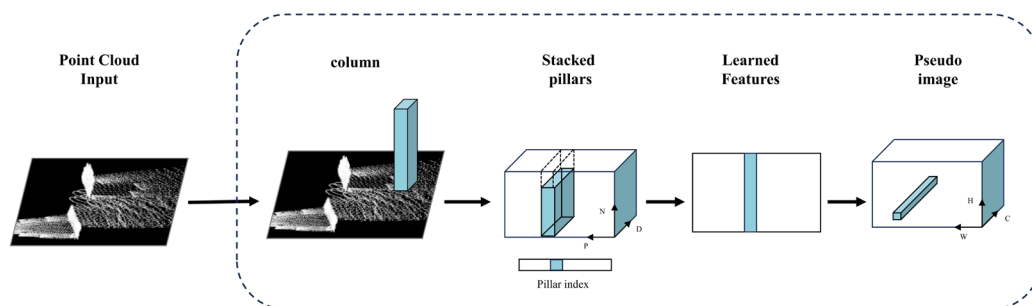


Figure 4. Pillar Feature Net (PFN) structure diagram. Initially, the original point cloud is divided into pillars, with multiple points allocated to each pillar. Subsequently, Multi-Layer Perceptron (MLP), Batch Normalization, Rectified Linear Unit (ReLU), and maximum pooling operations are employed to extract features from all points within each pillar. Consequently, a pseudo-image of size $(C \times H \times W)$ is generated.

The overarching process of PFN involves partitioning the original point cloud into several uniform square grids within the x-y plane. Each pillar extends along the axis aligned with its respective grid. As a result, the entire point cloud sample can be represented as a tensor with the shape (D, P, N) , where D represents the feature dimension of the point cloud, P denotes the number of valid pillars in the original point cloud, and N represents the number of points in each pillar. In the context of APEM, each point in the point cloud comprises ten feature dimensions, including the coordinate information (x, y, z) , intensity (r) , coordinate offsets (x_c, y_c, z_c) , and offsets relative to the geometric center of the pillar (x_p, y_p, z_p) . N denotes the maximum number of points in each pillar. If the number of

points is fewer than N , zero-padding is applied to complete the pillar; if the number exceeds N , N points are randomly sampled from the pillar.

After obtaining the tensor of shape (D, P, N) for feature extraction, the first step involves applying a Multi-Layer Perceptron to each point in every pillar. This MLP encodes the position and extends the feature dimensions from D to C , resulting in a tensor of shape (C, P, N) . Subsequently, a maximum pooling operation is conducted on each pillar to extract the most salient feature points, yielding a feature map with a shape of (C, P) that expresses the overall features of each pillar.

Finally, based on the index, the point cloud is remapped to its corresponding position on the original mesh, expanding P to (H, W) . This process yields a (C, H, W) feature representation of the pseudo-image, where H and W denote the height and width of the pseudo-image.

2.2. 2D Backbone

In the original point cloud, after the feature coding network generates the pseudo-two-dimensional image, a convolutional neural network can effectively extract features from this image. The structure of the backbone extraction network is depicted in Figure 5. Drawing inspiration from HorNet, we introduced Recursive Gated Convolutions (g^n Conv) into the feature extraction network. This addition addresses the challenge posed by the numerous long tunnels present in most mines, which often leads to ineffective feature extraction for personnel located at longer distances.

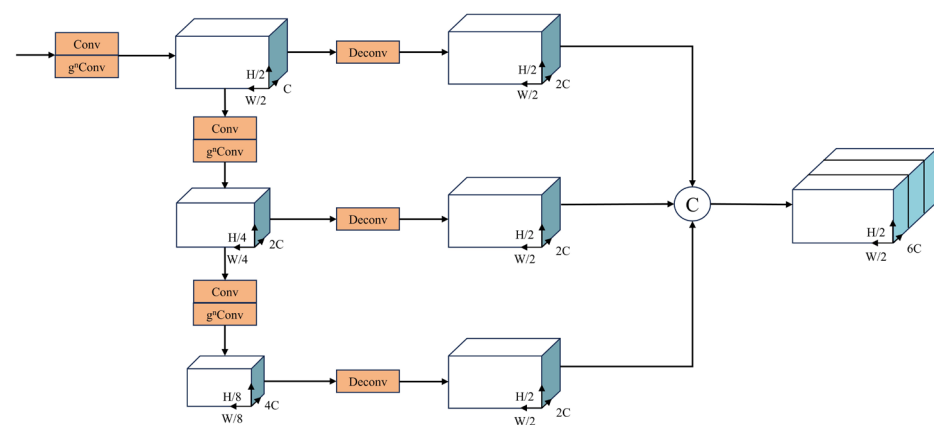


Figure 5. Backbone extraction network. Conv and gated g^n Conv layers are stacked to create a subnetwork for top-down operations, facilitating the extraction of features from progressively smaller spatial resolution feature maps. Another subnetwork is dedicated to upsampling features extracted from feature maps with varying resolutions to the same dimension through transposition convolution operations. Finally, multi-resolution feature maps are interconnected in series.

The g^n Conv module is a convolutional operator that enables higher-order spatial interactions through iterative neighborhood aggregation. This design is particularly well suited to underground mining environments for three reasons. First, mining operations require real-time perception under strict computational constraints. Unlike self-attention mechanisms [43] that scale quadratically with input size, g^n Conv uses a convolution-based implementation with linear complexity, ensuring efficient inference on embedded systems deployed on load-haul-dump machines. Second, the geometric layout of mine tunnels involves narrow passages, irregular surfaces, and frequent occlusions. Detecting personnel in such settings demands robust modeling of long-range spatial dependencies among sparse point cloud targets. The pyramidal structure of g^n Conv progressively expands the receptive field while increasing channel width, allowing high-order interactions to be captured without exceeding practical memory or latency budgets. Third, LiDAR scans in

underground mines suffer from non-uniform point density due to dust, water spray, and sensor mounting variations. The translation invariance inherited from standard convolutions ensures that g^n Conv produces stable feature responses across diverse scanning conditions, which is essential for reliable detection in highly dynamic and degraded sensing environments. When $n = 1$, the first-order spatial interaction forms the base layer of the architecture, as illustrated in Figure 6.

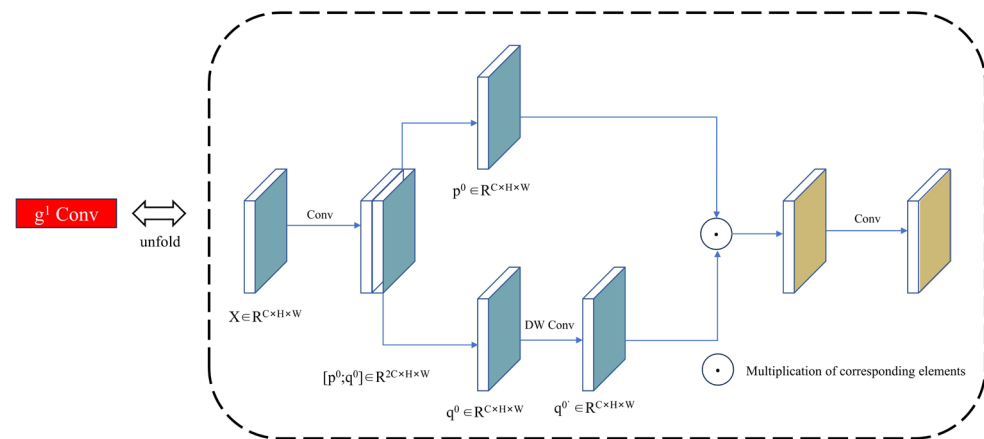


Figure 6. g^1 Conv structure diagram. Utilizing standard convolution, deep separable convolution, and element-wise multiplication enables the effective operation of long-term and high-order spatial interactions. This approach features an input-adaptive spatial mixing function akin to self-attention.

Let X represent the input feature with dimensions (C, W, H) . The first-order spatial interaction, denoted as $Y = g^1 \text{Conv}(X)$, proceeds as follows. Initially, the input feature X undergoes a convolution operation to increase the number of channels to $2C$. The channel layer is then evenly split into two parts, q^0 and p^0 . The p^0 part serves as the backbone, preserving global semantic information, which directly contributes to task processing. On the other hand, q^0 functions as the recursive component, utilizing dynamic convolution and a recursive mechanism to extract more complex contextual information. Through the recursive gating mechanism, a dynamically adjusted recursive path is constructed, which enhances the ability to extract and express features while retaining global information, thus improving the model's capacity for sparse data handling.

$$\left[p^0 \in R^{C \times H \times W}; q^0 \in R^{C \times H \times W} \right] = \phi_{in}(X) \in R^{2C \times H \times W} \quad (1)$$

where ϕ_{in} represents the input convolutional layer, H is the height of the feature, W is the width of the feature, and C is the number of channels of the feature. Then perform deep separable convolution (DW Conv) on q^0 :

$$q^{0'} = f(q^0) \in R^{C \times H \times W} \quad (2)$$

where f represents the depth separable convolution layer. Then take the dot product of p^0 and $q^{0'}$ to get p^1 :

$$p^1 = q^{0'} \odot p^0 \in R^{C \times W \times H} \quad (3)$$

Finally, the result of spatial interaction is mapped to the specified dimension through convolution:

$$Y = \phi_{out}(p^1) \in R^{C \times H \times W} \quad (4)$$

In this work, the spatial interaction is implemented with an order of five, in accordance with the empirical analysis presented in Section 4.3. Our evaluation—spanning interaction

orders from one through ten—reveals that model performance improves markedly up to five, after which additional orders deliver only marginal gains while substantially increasing computational burden. Given the stringent latency and memory constraints inherent to real-time personnel detection in mining environments, an order of five represents a well-justified and operationally viable configuration. The architecture of the five-order spatial interaction module and its integration within the overall network are illustrated in Figure 7.

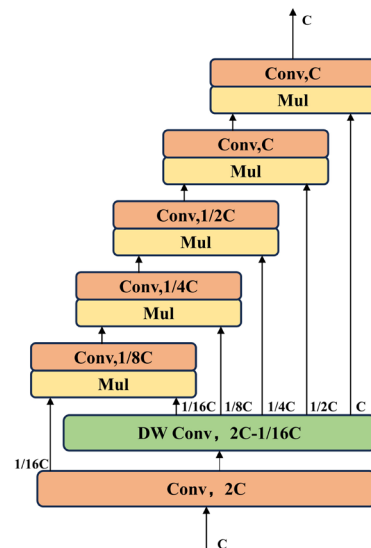


Figure 7. g^5 Conv structure diagram. The efficient implementation of gated convolution and recursive design is used to realize the fifth-order spatial interaction (the reason for selecting the fifth-order spatial interaction is shown in Section 5). The original feature map with channel C is convolved into channel $2C$ by ordinary convolution, and the feature map of $(2C-1/16C)$ is deeply separable. Convolution and element multiplication, along with convolution with the remaining channel, produce a $1/16C$ feature map, finally achieving long-term and high-order spatial interaction.

The Backbone extraction network produces feature maps with varying receptive fields at each trunk stage by stacking Conv layers with g^n Conv layers, as illustrated in Figure 5. Subsequently, the large, medium, and small feature maps generated by the trunk are upsampled to the same size using deconvolution to create three output layers with uniform resolution. The output layers are concatenated along the depth direction to form a comprehensive feature fusion tensor, which is then transferred to subsequent steps.

2.3. CASA Module

Due to the sparsity of point cloud data detected by LiDAR, long tunnel distances in mines often encounter issues of false detection and omission during the detection process. Addressing these challenges and considering the specific conditions underground, this work introduces Channel and Spatial Attention (CASA) by adapting the Convolutional Block Attention Module (CBAM). CASA comprises Efficient Channel Attention and Spatial Attention components. The overall structure of CASA is depicted in Figure 8.

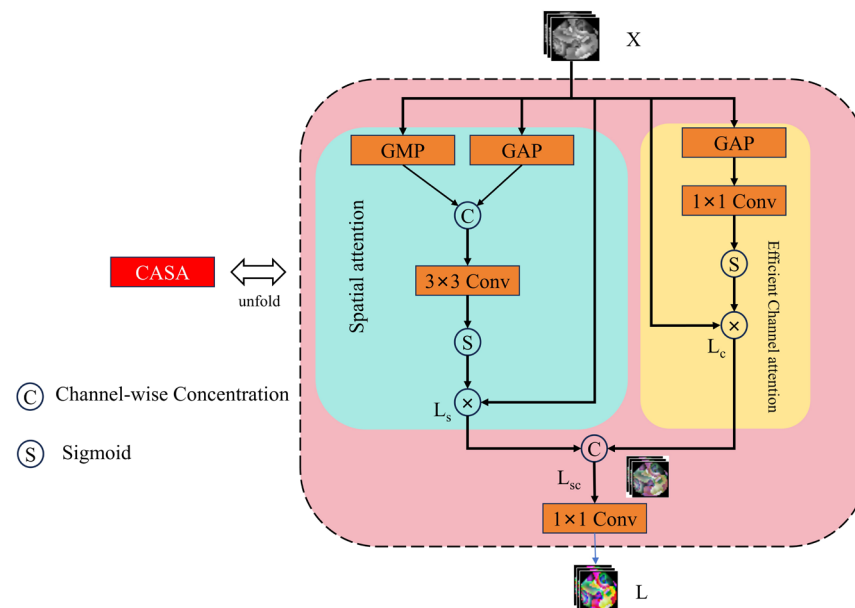


Figure 8. Overall structure diagram of the CASA module. Utilizing Efficient Channel Attention (ECA) and Spatial Attention (SA) mechanisms, the original feature map receives channel attention weights and spatial attention weights, which are subsequently stacked along the channel dimension. Finally, a convolutional kernel with a size of 1 is employed to integrate spatial and channel attention, yielding the feature map with adaptive space-channel attention.

The Efficient Channel Attention (ECA) module dynamically allocates attention to the weight between each channel by learning the intrinsic relationship within the feature map's channels. It focuses on channels with rich feature information to enhance the overall performance of the network. The overall structure of the ECA is depicted in Figure 9. Initially, the model takes the feature map output from the feature extraction network and computes the average value of each channel using a global average pooling layer, thus learning network features. Subsequently, a 1×1 convolution is applied to adjacent T dimensions within each channel to derive the attentional weights of the channels, which are then applied to each channel of the input feature map, resulting in a weighted combination of different channels in the feature map. Finally, the adjusted features are normalized using the Sigmoid function and multiplied with the initial feature map to obtain the channel attention-weighted feature map. The overall process is represented in Equation (5):

$$L_c = \sigma(f_1(\text{AvgPool}(X))) \cdot X \quad (5)$$

The Spatial Attention (SA) module is designed to enhance the feature representation of key regions by transforming the spatial information in the original feature map into another space while preserving essential details. This process generates a weight mask for each location, enabling the amplification of specific target regions of interest while attenuating irrelevant background regions. The overall structure of SA is illustrated in Figure 10. This module takes the feature map output from the feature extraction network as input and computes the maximum and average values for each feature point along the channels to integrate the depth dimension, yielding two descriptive feature maps. These maps are concatenated along the depth dimension, and attention values for different regions at the same depth are computed using convolution and a sigmoid activation function. These values are then multiplied with the initial feature map to obtain the spatial attention weight feature map, as detailed in Equation (6).

$$L_s(X) = \sigma(f_2([\text{AvgPool}(X); \text{MaxPool}(X)])) \cdot X \quad (6)$$

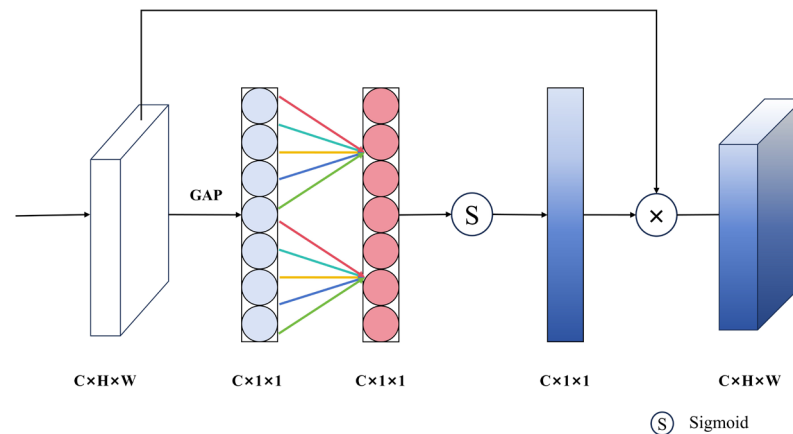


Figure 9. ECA module structure diagram. The parameters C , W , and H denote the dimensions of the feature map, while GAP signifies the global average pooling layer. This layer extracts channel features and facilitates the acquisition of new feature maps via adaptive one-dimensional convolution. Utilizing the sigmoid activation function, attention weights are mapped between 0 and 1, and subsequently multiplied by the original feature map. Finally, the feature map with adaptive channel weights is derived.

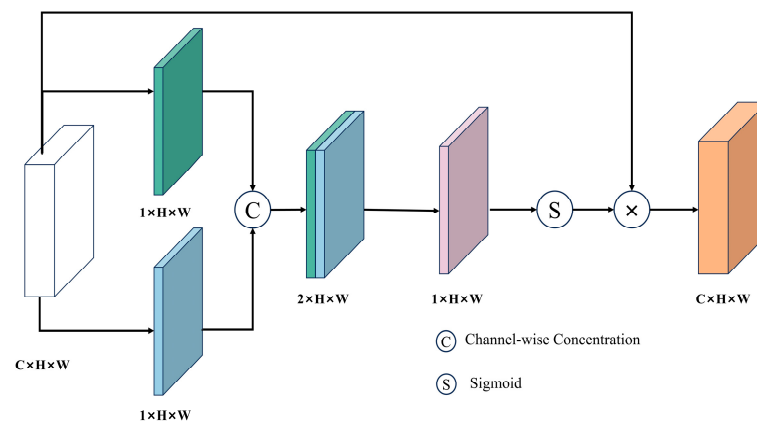


Figure 10. SA module structure diagram. The parameters C , W , and H represent the dimensions of the feature map. Max and Mean denote operations where the maximum and mean values, respectively, are computed at each position along each channel, followed by stacking and quadratic convolution. The sigmoid activation function is applied to map attention weights between 0 and 1, which are then multiplied by the dot product of the original feature maps to ultimately obtain feature maps with adaptive spatial weights.

In CASA, the output feature layers of ECA and SA are concatenated along the depth channel and subsequently convolved. The resulting feature map is then outputted for detection, as illustrated in the following formula:

$$\begin{aligned}
 L &= f(L_{sc}) \\
 &= f([L_s; L_c]) \\
 &= f([\sigma(f_2([AvgPool(X); MaxPool(X)])) \cdot X; \sigma(f_1(AvgPool(X))) \cdot X])
 \end{aligned} \tag{7}$$

Motivated by the unique characteristics of underground tunnels—including repetitive geometric structures, low-texture surfaces, and severe signal degradation from dust and moisture—we adapt the CASA attention mechanism by integrating SA and ECA. This hybrid design dynamically recalibrates the importance of both spatial locations and feature channels in response to scene content. In narrow, cluttered mine passageways, personnel often appear at varying distances and under partial occlusion; SA helps the model focus on relevant regions despite background clutter, while ECA selectively amplifies discriminative

channels that encode human shape and posture cues, which remain stable even when intensity or point density fluctuates. By fusing these two complementary attention pathways, CASA enhances feature expressiveness without introducing significant computational overhead—making it especially suitable for real-time safety-critical perception on mining vehicles.

2.4. Detection Head

The SSD is employed as the detection head to predict the position, orientation, and category of 3D humans. The Intersection over Union (IOU) on the Bird's Eye View (BEV) plane is utilized for matching between prediction frames and real labels. Additionally, height information is computed separately as a regression target. In real tunnel scenes, all humans can be approximated to lie within the same plane in three-dimensional space within LiDAR's scanning range, with minimal height differences between scanned humans. Therefore, the SmoothL1 function is directly applied as the regression loss for bounding box coordinate refinement. This loss function combines the best properties of both L1 and L2 losses: it behaves like the L2 loss for small residuals, ensuring smooth gradients near zero, and switches to the L1 loss for larger residuals, making it less sensitive to outliers. This hybrid behavior leads to more stable and efficient training compared to using pure L1 or L2 losses. Consequently, its use yields improved regression accuracy for human bounding box localization. Furthermore, the Feature Pyramid Network (FPN) is integrated into the detection head to handle humans of various sizes, facilitating more accurate human localization by extracting multi-scale features.

3. Experiments

3.1. Dataset Details

To enhance the robustness of the experiments, we meticulously planned and executed a series of tests in the Jinchuan No. 2 Mining Area. This site was specifically chosen due to its complex geological conditions and extensive tunnel networks, which together formed an ideal testbed for our research.

The data collection phase was a critical component of the study. We employed state-of-the-art technology, specifically the Ouster OS1 LiDAR, to collect a large volume of personnel-related data in underground environments. The Ouster OS1 is a high-performance 64-line LiDAR system that has significantly advanced 3D data acquisition. It demonstrates exceptional capabilities in distance detection and reflectance sensitivity: it can detect targets with reflectivity as low as 10% at distances up to 90 m, with a maximum range extending to 200 m. This performance enables comprehensive scanning of underground tunnels, ensuring the capture of detailed spatial information about personnel and their surrounding environments. The detailed specifications of the LiDAR sensor are provided in Table 1.

Table 1. LiDAR specifications. This table provides a comprehensive overview of the LiDAR system's technical specifications.

Parameter Name	Parameter Specification
Model	Ouster OS1
Number of Lines	32/64
Laser Wavelength	865 nm
Range (at 10% reflectivity)	90 m
Accuracy	±2.5 cm
Horizontal Field of View	360°
Vertical Field of View	+21.2° to −21.2°
Frame Rate	10/20 Hz

Scanning operations were carried out in the intersecting and long tunnels of the Jinchuan No. 2 Mining Area. These tunnels present unique challenges for data collection, such as variable lighting conditions, airborne dust particles, and complex geometric structures. Nevertheless, the Ouster OS1 LiDAR largely overcame these challenges, generating high-quality point cloud data. Point cloud data comprises discrete 3D spatial points that, when integrated, form a detailed representation of the scanned objects and their environments.

After data collection, the entire dataset was named the “Jinchuan Underground Mining Personnel Dataset” (as illustrated in Figure 11). The raw dataset contained a vast amount of information, but not all of it was relevant to our research on underground personnel. Therefore, a rigorous screening and processing procedure was implemented, involving noise filtering, removal of points unrelated to personnel or relevant tunnel objects, and data normalization to facilitate subsequent analysis. Through this meticulous screening and processing, we successfully isolated 3700 frames of point cloud data that specifically contain information about underground personnel. Despite its limited scale, this in-house dataset was intentionally curated to capture extreme scene complexity and a broad spectrum of real-world interference factors commonly encountered in underground mining environments. The selected site features persistent high-concentration dust, near-complete darkness or severely dim lighting, intense laser reflections from metallic support structures, highly non-standard human postures—including bending and squatting—and dynamic occlusions caused by large haul trucks and roadheaders.



Figure 11. Specifics of the underground mine personnel dataset: (a) the Ouster OS1 mounted on top of a shovel loader, (b) a 2D image frame of the underground mine dataset captured by cameras, and (c) a 3D point cloud image frame of the underground mine dataset captured by LiDAR.

The dataset annotation follows the format of the KITTI dataset [44]. In the KITTI dataset, each detection target in a frame of the point cloud is represented by 14 values: category, truncation, occlusion, observation angle, 2D bounding box coordinates, human length, width, height, and 3D bounding box coordinates. As this dataset consists solely of point cloud data without two-dimensional image information, the 2D bounding box coordinates of the human are set to 0 and are not utilized in model training.

3.2. Data Augmentation

To achieve satisfactory results on this dataset, data augmentation proves to be a crucial technique. In this experiment, the following procedures were implemented for data augmentation.

Initially, a database is constructed from the training dataset, comprising comprehensive ground-truth annotations paired with their corresponding point cloud frames. Human instances are often sparse in LiDAR scans, and many frames contain few or even zero annotated pedestrians. This sparsity can hinder the model’s ability to learn robust detection features. To address this data imbalance and enrich the diversity of training samples, an

augmentation strategy inspired by practices in PointPillars-like frameworks is adopted. Specifically, 15 authentic labeled human instances are randomly sampled from the entire training set and synthetically inserted into each training frame. The number 15 is chosen to strike a balance between increasing instance density to strengthen the learning signal and avoiding excessive clutter that could lead to occlusion, ambiguity, or unrealistic scene configurations. The inserted instances are carefully placed in free space to ensure they do not overlap with or occlude any existing objects in the original point cloud.

Additionally, to minimize the impact of noise, it is essential to independently and randomly transform each labeled truth value and point cloud data. Each background truth value undergoes random rotation, with the degree of rotation following a uniform distribution between -1.57 and 1.57 radians.

Lastly, global scaling and global random flip along the x-axis were applied to the entire point cloud. The global scaling adheres to a uniform distribution between 0.95 and 1.05, while the global random flip along the x-axis has a probability of 50%.

3.3. Parameter Configuration

The direction of the point cloud in the radar coordinate system is illustrated in Figure 12. The x-axis points forward, the y-axis points to the left, and the z-axis points upward. The range of the point cloud scene on the x-axis and y-axis is $(0, 40.96)$ and $(-6.4, 6.4)$ respectively, while the range on the z-axis is $(-3, 3)$. The dimensions of each pillar are set to differ: 0.16 m in length, 0.16 m in width, and 6 m in height.

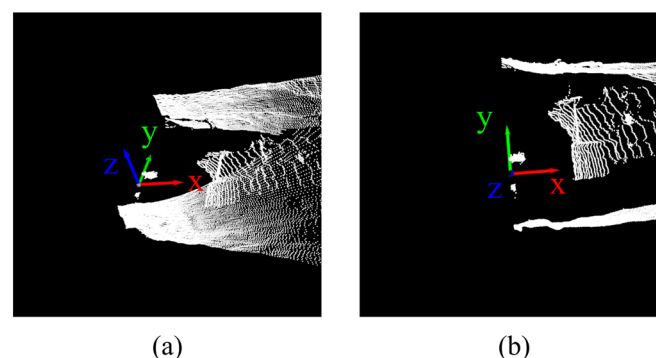


Figure 12. Schematic diagram of the radar coordinate system. (a) Side view; (b) Bottom view. Red, green, and blue arrows represent the forward x-axis, left y-axis, and upward z-axis, respectively.

The maximum number of pillars generated in each frame of the point cloud is 16,000. In each pillar, 32 points are randomly sampled. If the number of points in some pillars is less than 32, they are completed with zero-padding. In generating the initial anchor frames, each anchor frame is determined by seven parameters, $(c_x, c_y, c_z, w, l, h, \theta)$, where c_x , c_y , and c_z represent the coordinates of the center point of the anchor box; the width w , length l , and height h facilities of the initial anchor frame are $(0.6, 0.8, 2.0)$; and the direction of the initial anchor frame θ is set to two vertical directions $(0, 1.57)$. When setting positive and negative samples through IOU, the samples with IOU greater than 0.5 are set as positive samples, the samples with IOU less than 0.35 are set as negative samples, and the samples between $(0.35, 0.5)$ are ignored in the calculation of loss. Finally, the threshold value is set to 0.5 when performing non-maximum suppression.

3.4. Model Training

The computer environment for training and testing this algorithm comprises an Ubuntu 20.04 operating system, an Intel® Core™ i9-9900K CPU running at 3.6 GHz, and an NVIDIA GeForce GTX 1080 Ti (NVIDIA Corporation, Santa Clara, CA, USA).

graphics card with 12 GB of video memory. APEM is an improved model based on the PointPillars algorithm within the OpenPCDet framework, implemented in Python 3.8. As an open-source point cloud object detection library built on PyTorch 1.13.1, OpenPCDet integrates advanced techniques including data augmentation, optimized optimizers, and sophisticated learning strategies to enhance model performance, thereby improving the detection accuracy of the trained network.

For optimization, the Stochastic Gradient Descent (SGD) algorithm was adopted, with an initial learning rate of 0.001 and a momentum factor of 0.9 to accelerate convergence and mitigate the risk of local optima. The model was trained using the Jinchuan Underground Mining Personnel Dataset constructed in this study, where deep-layer parameters were iteratively updated via the optimizer until the detection accuracy on the validation set stabilized. In each training epoch, the model updated its parameters on the training set, followed by performance evaluation on the validation set.

Considering that the dataset, with a size of 3700 frames, is a small-scale dataset, we specifically adopted the 5-fold cross-validation method in the experiment to maximize the reliability and stability of the evaluation results. In the specific implementation process, we first conducted strict preprocessing on the original dataset, removing the existing abnormal frames and duplicate frames to ensure the validity of the data. Subsequently, the dataset was randomly and evenly divided into 5 subsets by means of stratified sampling, with each subset containing 740 frames of data, and each subset being highly consistent with the original dataset in terms of data distribution characteristics. In each round of validation, 4 of these subsets were selected as the training set, and the remaining 1 subset was used as an independent test set to evaluate the generalization ability of the model after training in that round. This process was repeated for 5 experiments, with a different subset used as the test set each time. Moreover, the model parameters were reinitialized before each round of training to avoid the interference of previous training results on subsequent experiments. Finally, we took the average of each performance index obtained from the 5 tests as the final evaluation result of the model. This approach can make full use of the limited dataset, effectively reduce the evaluation bias caused by the randomness of data division, and more comprehensively reflect the performance of the model on different data subsets, thereby making the experimental results more convincing and credible.

The optimizer chosen for training was SGD. Throughout the training of APEM, the loss change curve, depicted in Figure 13, illustrates a continual decrease in loss value, ultimately reaching a convergence state.

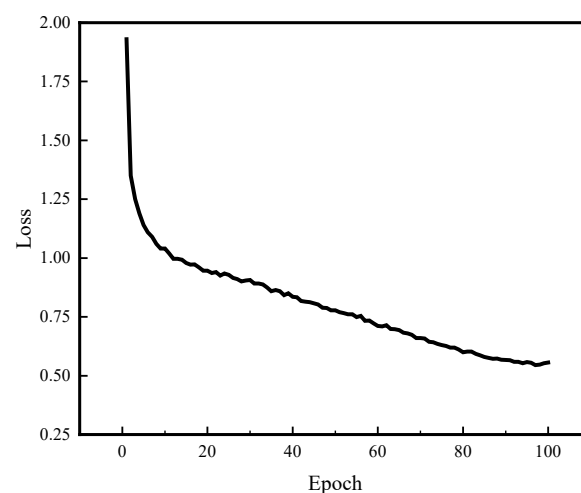


Figure 13. Loss change curve of APEM.

3.5. Evaluation Metrics

In the realm of object detection, Intersection over Union (IOU), Precision Rate (Precision), Recall Rate (Recall), and Average Precision (AP) serve as primary metrics to gauge algorithm performance. IOU quantifies the intersection and union ratio, denoted by Equation (8), signifying the extent of overlap between actual and predicted frames.

$$IOU = \frac{Detection\ Result \cap Ground\ Truth}{Detection\ Result \cup Ground\ Truth} \quad (8)$$

IOU is utilized to ascertain the accuracy of the generated detection box. Typically, a threshold value (T) is established. If the IOU between the detection box and the actual box exceeds T, the detection is deemed correct; otherwise, it is considered incorrect. Based on this criterion, the following scenarios can be delineated:

TP (True Positives): Refers to the number of positive samples that are correctly categorized, predicted to be positive, and actually positive.

TN (True Negatives): Refers to the number of negative samples that are correctly categorized, predicted to be negative, and actually negative.

FP (False Positives): Refers to the number of negative samples that are incorrectly labeled as positive samples, which are actually negative samples but predicted as positive samples.

FN (False Negatives): Refers to the number of positive samples that are incorrectly labeled as negative samples, which are actually positive but predicted to be negative.

The concepts of Precision and Recall can be derived based on the Intersection over Union (IOU). Precision represents the proportion of correctly identified positive samples out of all samples predicted as positive. This metric primarily reflects prediction accuracy but may result in missed detections (FN). The formula is as follows:

$$precision = \frac{TP}{TP + FP} \quad (9)$$

Recall represents the ratio of predicted positive samples to the total number of actual positive samples. This metric aims to encompass as many positive samples as possible but may lead to false positives (FP). The formula is as follows:

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

Average Precision (AP) is a combination of points using different Precision and Recall values. AP is the area enclosed by the Precision–Recall (P-R) curve and the coordinate axes. The formula is as follows:

$$AP_i = \int_0^1 P_i(R_i) dR \quad (11)$$

In general, the better the object detection model, the higher its AP value. In order to compare with existing algorithms, AP is also used as the main detection metric.

3.6. Ablation Experiment

To systematically evaluate the efficacy of the proposed enhancement, we designed ablation studies to quantify the contribution of each module to overall detection performance. The model was first evaluated on the validation set and benchmarked against the original PointPillars (PP), with results summarized in Table 2. Here, 3D denotes the average precision (AP) in three-dimensional space, while BEV refers to AP from the bird's-eye-view perspective. Specifically, we investigated the individual and combined effects of g^n Conv and the CASA module: PP& g^n represents PointPillars augmented solely with g^n convolu-

tion; PP&CASA indicates performance when only the channel-spatial attention mechanism is integrated; and PP&gⁿ&CASA reflects the joint integration of both components. To ensure statistical robustness, all variants were trained independently five times under identical hyperparameter settings, each initialized with a distinct random seed (42, 156, 784, 2026, 864). Final metrics are reported as mean $\pm$ standard deviation across the five runs—rather than relying on a single best trial—to more faithfully capture both central tendency and performance variability. All experiments were conducted with a fixed epoch count of 100 and evaluated using the AP(R40) metric for consistency. As shown in Table 2, the concurrent incorporation of gⁿ Conv and CASA consistently yields measurable gains in detection accuracy.

Table 2. Results of ablation experiments to study the effects of different setup scenarios of gⁿ Conv and CASA.

Sequences	Method	g ⁿ Conv	CASA	AP _{BEV} (%)	AP _{3D} (%)	FPS
①	PP	×	×	66.5 ($\pm$ 0.1)	47.3 ($\pm$ 0.2)	181
②	PP&g ⁿ	✓	×	66.6 ($\pm$ 0.1)	49.7 ($\pm$ 0.1)	169
③	PP&CASA	×	✓	68.6 ($\pm$ 0.2)	50.0 ($\pm$ 0.1)	179
④	PP&g ⁿ &CASA	✓	✓	69.3 ($\pm$ 0.1)	50.4 ($\pm$ 0.2)	166

During training, the SGD optimizer was selected, and ablation experiments were conducted under identical conditions. The loss curves before and after model improvements during training are depicted in Figure 14. From the graph, it is evident that the loss values of all four methods in the ablation experiments steadily decrease and eventually converge.

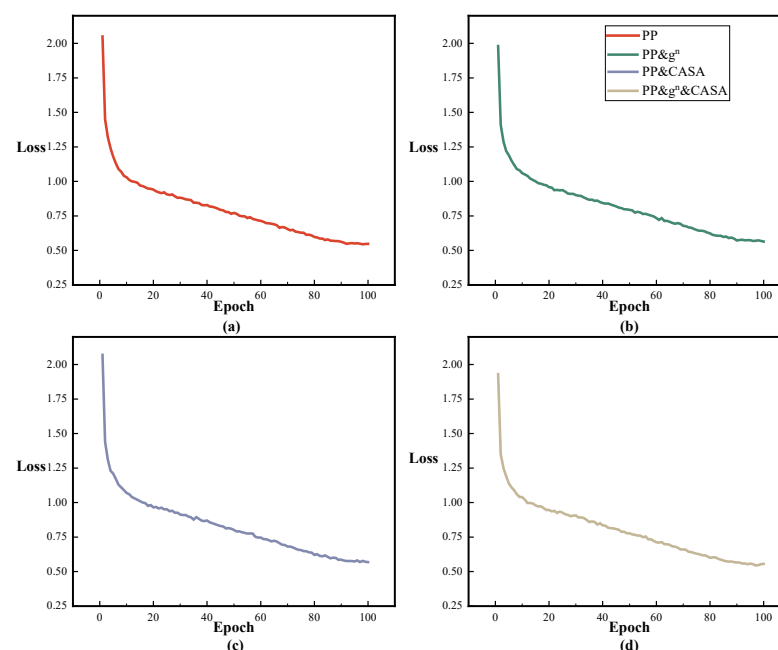


Figure 14. The graph below shows the loss variation curves for the ablation experiments. Four different colors represent the changes in the loss function for four models: (a) PP, (b) PP&gⁿ, (c) PP&CASA, and (d) PP&gⁿ&CASA.

In terms of 3D detection accuracy, integrating the gⁿ Conv module improves AP_{3D} by 2.4%. When deployed independently, this component significantly enhances the model's ability to capture high-order spatial interactions and global semantic structures—directly evidenced by the AP_{3D} gain. However, its impact on bird's-eye-view (BEV) detection

remains marginal, with AP_{BEV} increasing by only 0.1%, indicating that while g^n Conv excels in 3D feature extraction, its contribution to bird's-eye-view projection is limited.

By contrast, integrating the CASA module in isolation yields gains of 2.7% in AP_{3D} and 2.1% in AP_{BEV} . CASA dynamically reweights input feature maps via a unified channel-spatial attention mechanism, directing the model's focus toward more discriminative regions. This effect is especially evident in the marked improvement in AP_{3D} , corroborating CASA's ability to enrich 3D feature representations; simultaneously, the consistent rise in AP_{BEV} underscores its efficacy in enhancing bird's-eye-view projection quality, thereby demonstrating its comprehensive contribution to overall detection performance.

When g^n Conv and CASA are jointly integrated, AP_{3D} and AP_{BEV} improve by 3.1% and 2.8%, respectively. The two modules act in synergy: g^n Conv leverages its recursive architecture to increase feature complexity and more effectively capture long-range spatial dependencies, whereas CASA enhances feature saliency through attention-guided emphasis. This complementary interplay preserves global semantic integrity while uncovering richer contextual relationships, resulting in consistent performance gains across both the 3D and bird's-eye-view domains and ultimately yielding a more robust and accurate detection system.

To objectively evaluate the performance gap between PointPillars and our proposed APEM, we performed five independent trials on the core metrics AP_{BEV} and AP_{3D} (random seeds: 42, 156, 784, 2026, 864). As shown in Table 2, APEM consistently outperforms the baseline, achieving $69.3\% \pm 0.1\%$ in AP_{BEV} (compared to $66.5\% \pm 0.1\%$ for PointPillars, representing a +2.8% improvement) and $50.4\% \pm 0.2\%$ in AP_{3D} (compared to $47.3\% \pm 0.2\%$ for PointPillars, a +3.1% gain). Given the limited number of repetitions, we adopted non-parametric statistical inference to ensure reliable conclusions. Specifically, a one-sided Wilcoxon signed-rank test was conducted to assess whether APEM systematically surpasses PointPillars—a well-established approach for paired, small-sample comparisons in machine learning that makes no assumption about data normality. Building upon this framework, we further computed the Hodges–Lehmann estimator and its 95% confidence interval (CI) to deliver a robust and unbiased quantification of the performance gain. The results indicate median improvements of 0.029 for AP_{BEV} (95% CI: [0.027, 0.032]) and 0.030 for AP_{3D} (95% CI: [0.028, 0.033]), with both CIs strictly bounded away from zero—providing strong evidence of a genuine positive effect. Furthermore, the one-sided Wilcoxon test yields $p = 0.03125$ for both metrics, allowing us to reject the null hypothesis that APEM is not superior to PointPillars. Notably, APEM achieved higher scores in all five independent runs, highlighting the remarkable consistency and robustness of its performance advantage.

The addition of the g^n Conv and CASA modules results in varying reductions in the algorithm's Frames Per Second (FPS). As shown in Figure 15, the original model operates at 181 FPS, whereas the inclusion of g^n Conv reduces the FPS to 169. For the models with PP&CASA and PP& g^n &CASA, the FPS are 179 and 166, respectively. This suggests that although both g^n Conv and CASA introduce additional computational complexity, the overall speed does not degrade significantly. The slight increase in FPS for the PP&CASA model suggests that CASA is relatively lightweight and has minimal impact on inference speed. In contrast, the introduction of g^n Conv leads to a more noticeable drop in FPS, which can be attributed to its increased computational demands arising from the recursive mechanism and higher-order spatial interactions. Despite this reduction, the improved model still operates at a high FPS of 166, which remains sufficient for real-time detection requirements, particularly for applications such as personnel detection in underground tunnels. Furthermore, the detection accuracy of the state-of-the-art (SOTA) model for personnel detection is preserved, confirming that the trade-off between accuracy and speed remains acceptable for real-world deployment in these environments.

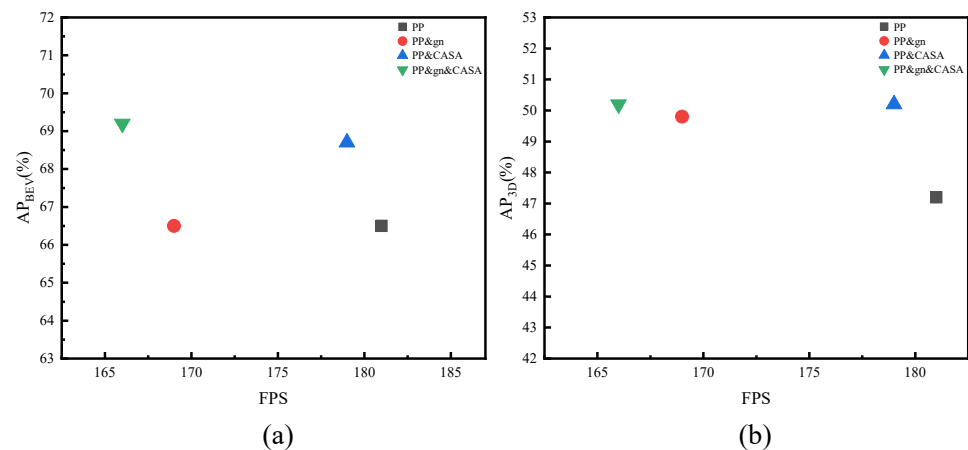


Figure 15. FPS and average precision (AP) results for each configuration in the ablation study. (a) APBEV (%) versus FPS; (b) AP3D (%) versus FPS. Although the improved algorithm exhibits a slight reduction in FPS, real-time detection capability is preserved, while both AP3D and APBEV improve to varying degrees.

The visualization results presented in Figure 16 demonstrate that APEM effectively detects and visualizes people in both crossways and long tunnels using 3D box selection. The results are sufficiently robust to fulfill the task of detecting people in the unstructured environment of underground tunnels.

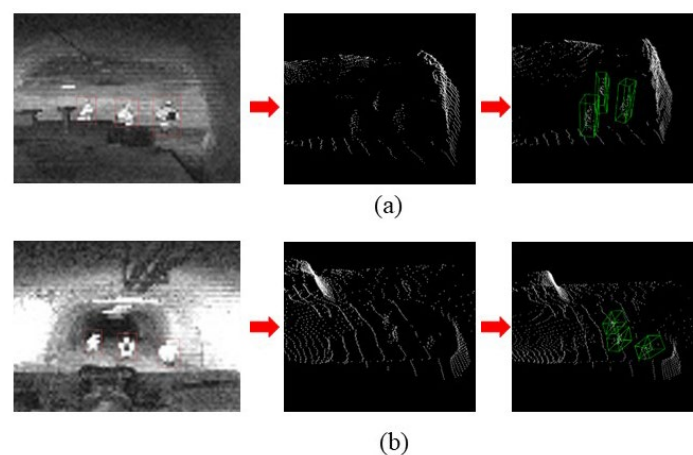


Figure 16. Personnel detection results in a real underground tunnel environment. (a) Intersection tunnel scenario; (b) Long straight tunnel scenario. In each case, the left panel shows the RGB image from the LiDAR-integrated camera, the middle panel displays the corresponding LiDAR point cloud, and the right panel presents the detection output of the APEM algorithm, where green bounding boxes indicate detected personnel.

4. Discussion

4.1. Selection of the Model Hyperparameters

In this section, we delve into the impact of two crucial model hyperparameters, namely learning rate and batch size, on the model's accuracy. Our objective is to explore the effects of different learning rates and batch sizes on the detection performance of the APEM model and determine the optimal configuration for achieving the highest accuracy.

The learning rate directly influences the model's convergence speed, while the batch size plays a pivotal role in determining the model's generalization ability. To assess the effects of varying learning rates and batch sizes on the APEM model, we conducted experiments across a range of learning rates from 0.001 to 0.008, and tested batch sizes of 2, 4, and 8, respectively. The results of these experiments are summarized in Table 3.

Table 3. Comparison of APEM accuracy under different hyperparameters.

Learning Rate	Batch Size					
	2		4		8	
	AP _{3D} (%)	AP _{BEV} (%)	AP _{3D} (%)	AP _{BEV} (%)	AP _{3D} (%)	AP _{BEV} (%)
0.001	49.3	69.0	48.4	68.9	47.1	69.2
0.003	47.3	68.9	46.6	68.8	50.2	69.2
0.005	44.2	66.4	46.8	68.8	47.2	69.2
0.008	41.4	66.4	43.9	66.1	46.6	68.9

The best detection results are highlighted in bold.

The data from Table 3 and Figure 17 shows that when the learning rate is set to 0.001 and the batch size is 2, the AP value for the 3D direction of the APEM model reaches 49.3%, while the AP value for the BEV direction reaches 69.0%. However, as the learning rate increases, both AP values begin to decrease. Similarly, when the batch size is set to 4 and the learning rate is 0.001, the algorithm achieves an AP value of 48.4% in the 3D direction and 68.9% in the BEV direction, with both values gradually declining as the learning rate increases.

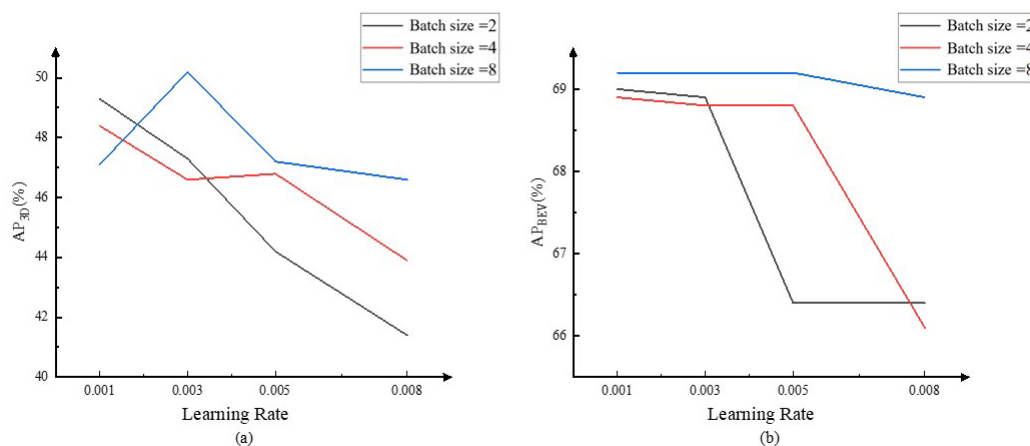


Figure 17. Model accuracy under varying hyperparameters. (a) AP_{3D} versus training epoch; (b) AP_{BEV} versus training epoch. Each curve corresponds to a different batch size, indicated by color.

Notably, when the batch size is 8 and the learning rate is set to 0.003, the performance of the APEM model peaks, with the AP value for the 3D direction reaching 50.2% and the AP value for the BEV direction reaching 69.2%. This hyperparameter configuration yields the best precision performance for the model. While the AP_{3D} increases with the higher learning rate, the AP_{BEV} remains relatively stable, suggesting that the 3D detection task is more sensitive to the learning rate and batch size.

In conclusion, a learning rate of 0.003 and a batch size of 8 result in the most significant improvement in model performance, enhancing 3D detection accuracy while maintaining a high AP_{BEV} value. In contrast, both lower and higher learning rates lead to training instability, negatively affecting the model's performance. Therefore, an optimal learning rate is critical to achieving stable and efficient training.

4.2. Spatial Interaction Order

Gated convolution (g^n Conv) enables spatial interactions of arbitrary orders, enhancing the model's feature extraction capability and high-density prediction performance. However, the order of spatial interactions affects the final feature extraction results. Therefore, we explore the impact of varying spatial interaction orders on detection performance.

In $g^{(i)}$ Conv, i denotes the spatial interaction order at each position. The data in Table 4 demonstrate that the spatial interaction order in gated convolution (g^n Conv) significantly influences the model's feature extraction and detection performance. Experimental results indicate that, within a certain range, increasing the spatial interaction order helps improve the model's detection accuracy.

Table 4. Recognition accuracy of personnel with different spatial interaction orders.

Spatial Interaction Order (g^n Conv)	AP _{BEV} (%)	AP _{3D} (%)
g^1 Conv	68.1	49.2
g^2 Conv	68.5	49.7
g^3 Conv	68.9	49.8
g^4 Conv	68.5	48.7
g^5 Conv	69.2	50.2
g^6 Conv	68.8	49.1
g^7 Conv	69.0	48.9
g^8 Conv	68.7	48.7
g^9 Conv	68.6	48.5
g^{10} Conv	68.6	48.4

The best detection results are highlighted in bold.

Specifically, the model achieves its best performance at fifth-order interaction (g^5 Conv), with an AP_{BEV} of 69.2% and an AP_{3D} of 50.2%—the highest values observed across all configurations. This suggests that fifth-order neighborhood aggregation effectively captures fine-grained geometric details while preserving contextual semantic coherence, leading to optimal detection accuracy.

However, when the spatial interaction order is further increased to 6 or 10, the detection performance degrades. This degradation can be attributed to the additional computational complexity introduced by higher-order interactions, which makes it difficult for the model to effectively process high-dimensional features and may further lead to information redundancy or training instability. Therefore, while increasing the spatial interaction order can enhance feature extraction capability, excessively high orders introduce unnecessary complexity that adversely affects the overall performance of the model. For instance, the 3D Average Precision (AP_{3D}) for g^6 Conv drops to 49.1%, and for g^{10} Conv, the Bird's Eye View Average Precision (AP_{BEV}) and AP_{3D} are recorded at 68.6% and 48.4%, respectively.

These results demonstrate that there exists an optimal range for spatial interaction order. Excessively high orders provide no additional benefit and may even impair model performance due to over-smoothing or optimization instability. Consequently, fifth-order spatial interaction strikes the best balance between representational capacity and model complexity in the current architecture.

4.3. Comparison with Existing Methods

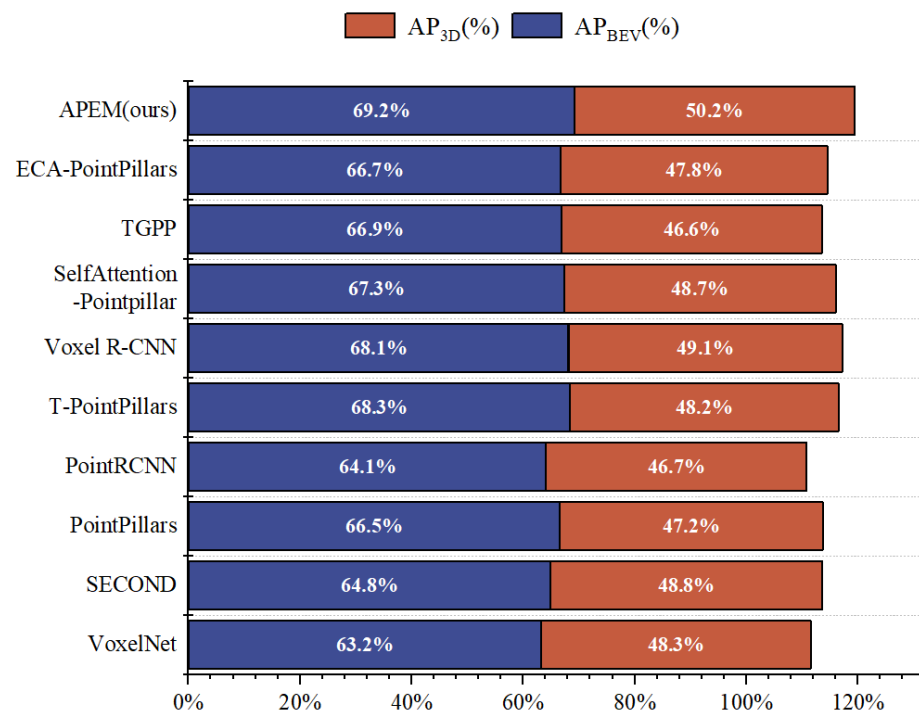
To validate the effectiveness of our proposed method, we conducted a series of systematic experiments using our self-constructed Underground Personnel Dataset. Several point cloud object detection methods were retrained in tunnel environments, including VoxelNet [31], SECOND [32], PointPillars [38], PointRCNN [33], T-PointPillars [45], Voxel R-CNN [46], SelfAttention-PointPillars [47], TGPP [48], and ECA-PointPillars [49]. These methods, through efficient feature extraction and object prediction, are capable of significantly improving detection speed while maintaining high accuracy, making them well-suited for dynamic and complex underground tunnel environments. The quantitative test results are presented in Table 5.

Table 5. The comparison of performances in bird's-eye-view (BEV) detection AP (%) and 3D object detection AP (%) on the Underground Personnel Dataset test set.

Method	Type	FPS	AP _{BEV} (%)	AP _{3D} (%)
VoxelNet	Voxel-based	103	63.2	48.3
SECOND	Voxel-based	143	64.8	48.8
PointPillars	Pillar-based	181	66.5	47.2
PointRCNN	Point-based	88	64.1	46.7
T-PointPillars	Pillar-based	157	68.3	48.2
Voxel R-CNN	Voxel-based	135	68.1	49.1
SelfAttention-PointPillars	Pillar-based	173	67.3	48.7
TGPP	Pillar-based	169	66.9	46.6
ECA-PointPillars	Pillar-based	171	66.7	47.8
APEM (ours)	Pillar-based	166	69.2	50.2

The best detection results are highlighted in bold.

The proposed method is compared with several state-of-the-art methods from recent years, as shown in Figure 18, using the Underground Personnel Dataset for training and testing. From Table 5, it can be observed that the proposed method is compared with voxel-based methods such as VoxelNet, SECOND, and Voxel R-CNN, demonstrating significant advantages in terms of FPS and detection accuracy. For instance, APEM outperforms VoxelNet, SECOND, and Voxel R-CNN by 6%, 4.4%, and 1.1% in AP_{BEV}, and by 1.9%, 1.4%, and 1.1% in AP_{3D}, respectively. Compared to pillar-based methods, such as PointPillars, T-PointPillar, SelfAttention-PointPillars, TGPP, and ECA-PointPillars, APEM achieves improvements of 2.7%, 0.9%, 1.9%, 2.3%, and 2.5% in AP_{BEV}, and 3%, 2%, 1.5%, 3.6%, and 2.4% in AP_{3D}, respectively. In comparison to point-based methods such as PointRCNN, APEM performs 5.1% better in AP_{BEV} and 3.5% better in AP_{3D}. Furthermore, APEM also demonstrates superior detection speed compared to most of the previously mentioned methods.

**Figure 18.** Comparison of 3D object detection overall accuracy with other point cloud object detection methods.

In summary, APEM maintains an advantage in detection speed while also surpassing the current mainstream single-stage point cloud object detection methods in accuracy. This demonstrates that the proposed improvement is both feasible and practical. In actual engineering experiments, APEM can reduce the probability of human–machine collisions, effectively enhancing operational safety in unstructured tunnel environments.

4.4. Distance Grading Accuracy Check

The performance of LiDAR-based personnel detection inevitably degrades at distance extremes due to fundamental sensing constraints. As documented in prior work [50], detection accuracy declines with increasing range owing to point cloud sparsity. In underground mining tunnels—where LiDAR units are typically mounted atop scrapers and operational zones span tens of meters—this challenge is exacerbated by both near-field occlusion and far-field signal attenuation.

To systematically assess APEM under these conditions, we trained the model on a custom long-tunnel dataset and partitioned the scene into three distance regimes relative to the sensor: short–medium ([5, 20.36] m), far ([20.36, 40.84] m), and very far ([40.84, 66.44] m). The interval [0, 5] m corresponds to the blind zone between the LiDAR and the scraper bucket, which is inherently unoccupied by personnel. Qualitative examples across these ranges are shown in Figure 19.

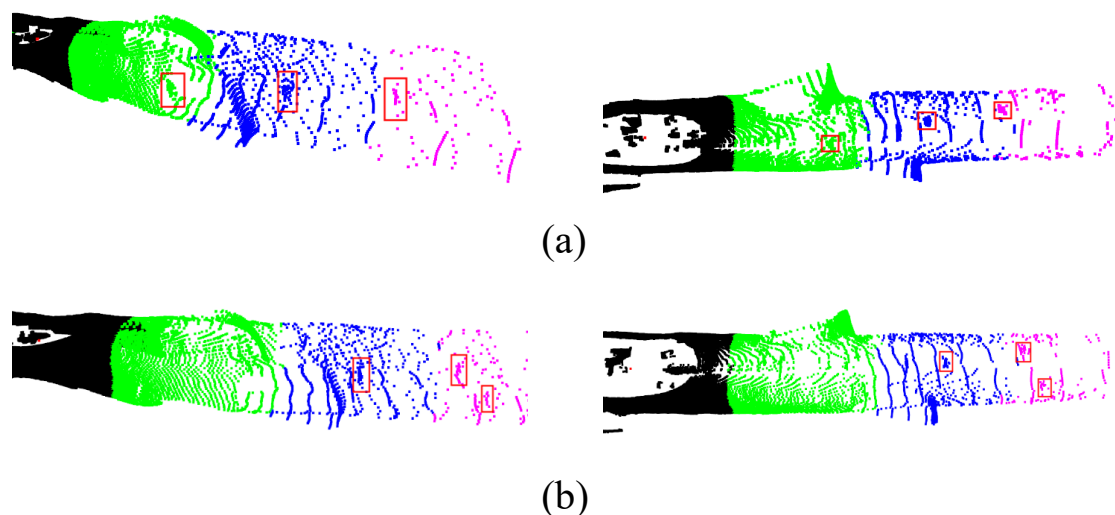


Figure 19. Point cloud maps of typical frames from the underground mine personnel dataset. (a) Uniform distribution of personnel; (b) Personnel at relatively long distances. Red represents very far-range regions ([40.84, 66.44]), blue denotes far-range regions ([20.36, 40.84]), and green indicates short-medium range regions ([5, 20.36]). Red frames mark the locations of personnel, and red dots represent LiDAR positions.

After personnel object detection across various ranges, the results are presented in Table 6. In Figure 20a, it is evident that when personnel are in close proximity to the LiDAR mounted on the scraper, numerous humans, including the scraper bucket and various communication equipment, obstruct the LiDAR’s view, resulting in partial loss of point cloud information from nearby personnel. This impedes the accuracy of short-medium range personnel detection, with an AP value of 44.98% in BEV and 44.77% in 3D. Figure 20b illustrates that when detecting personnel at a very far distance, the sparse nature of the point cloud makes it challenging for the LiDAR sensor to extract human characteristics beyond 40 m. Consequently, the point cloud information representing personnel characteristics becomes too scarce, making detection difficult for the algorithm. In this dataset, when the distance between the person and the LiDAR exceeds 40.84 m, the AP value is merely 20.61%

in BEV and 25.37% in 3D. Detection accuracy for personnel at far distances in the tunnel environment remains relatively stable, with an AP value of 44.5% in BEV and 44.45% in 3D.

Table 6. Personnel recognition accuracy at different distances.

Distance	AP _{BEV} (%)	AP _{3D} (%)
5–20.36 m	44.98	44.77
20.36–40.84 m	44.50	44.45
40.84–66.44 m	20.61	25.37

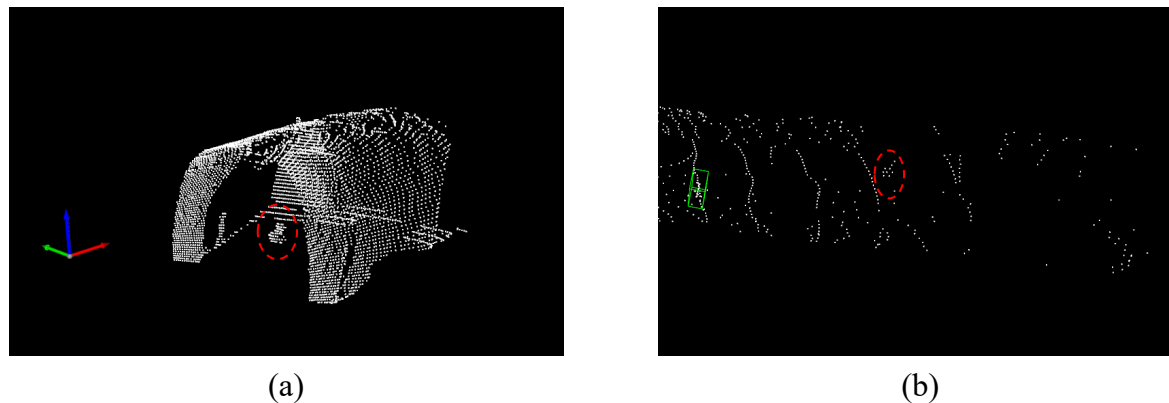


Figure 20. LiDAR point cloud quality at different ranges. (a) Short-to-medium range; (b) Very far range. Regions with undetected or unrecognizable structures are highlighted by red dashed circles. In (a), the lower half of the medium-range point cloud is not captured due to sensor field-of-view limitations; in (b), the point cloud becomes excessively sparse at long range, degrading detection reliability.

Critically, these limitations arise not from deficiencies in our attention-augmented architecture, but from inherent trade-offs in mobile LiDAR physics within confined, reflective environments, a constraint shared by all single-sensor 3D detectors in mining applications. Consequently, while APEM cannot overcome hardware-imposed sensing boundaries, it achieves reliable performance precisely in the intermediate corridor where human–machine interaction is most frequent. When integrated into a multi-layered safety system, APEM provides actionable situational awareness within its physically viable operating regime, supporting practical deployment.

5. Conclusions

In this work, to reduce human–machine collision accidents in mining environments, traditional personnel detection methods are analyzed, and a point cloud-based personnel detection method is proposed in combination with specific underground conditions. An improved PointPillars algorithm, named APEM, is introduced here, which is specifically designed for the unique environment of mine tunnels to enhance personnel detection performance and mitigate collision incidents. This enhanced algorithm is equipped with a refined feature extraction network, capable of accurately capturing long-range modeling information and high-order spatial interactions, thereby strengthening the feature extraction capability of the original algorithm during the feature encoding stage while minimizing feature loss. Additionally, we integrate the CASA module before the SSD detection head, enabling it to adaptively adjust the importance of different channels and spatial locations, thus achieving more accurate human localization in scenarios with varying target sizes. Through ablation experiments, we verify the effectiveness and necessity of these improvements; in mine tunnel environments, compared with the original algorithm, the personnel

detection accuracy is significantly improved by 3 percentage points (from 47.2% to 50.2%). Among single-stage point cloud object detection algorithms capable of real-time detection under lane conditions, the comprehensive detection accuracy of the proposed method reaches the state-of-the-art level. Furthermore, we explore the relatively high accuracy of APEM in detecting personnel targets within long-range distances. In summary, this method enables rapid and accurate detection of personnel positions around equipment, addressing the issue of personnel collision injuries caused by blind spots in underground mines and thereby enhancing operational safety.

6. Limitations and Future Work

Despite the promising performance of the proposed APEM algorithm in enhancing mining personnel detection, several limitations need to be acknowledged to guide future research.

Firstly, the dataset has constraints in diversity and scale. The Jinchuan Underground Mining Personnel Dataset, while specialized, is collected from a single mining area and focuses on intersecting tunnels and long straight tunnels. This geographical specificity may limit the model's generalizability to other mining regions with distinct geological structures, tunnel layouts, or operational conditions. Additionally, the dataset size is relatively small compared to mainstream public datasets, potentially restricting the model's ability to learn rare but critical scenarios. Although the dataset covers varying illumination and dust conditions, the limited variation in environmental factors fails to fully capture extreme cases, which pose potential risks in underground mining.

Secondly, environmental adaptability remains limited. Underground mining involves dynamic conditions such as vibrations from moving equipment, which can cause LiDAR misalignment and point cloud distortion. Moreover, the algorithm's robustness to sensor noise and temporary obstructions has not been thoroughly evaluated. These factors may compromise real-time detection reliability in complex operational scenarios.

To address the aforementioned limitations, our future work will focus on the following three directions. First, we will collect multi-source LiDAR data spanning diverse mineral types and geographic regions to enhance the model's generalization capability in complex industrial mining scenarios. Second, we will systematically define critical engineering constraints for LiDAR deployment, including mounting position, ranging accuracy, and frame rate, to ensure the reliability and adaptability of the perception system in real-world mining environments. Third, we plan to develop a multilevel self-diagnostic mechanism that includes a LiDAR heartbeat monitor to periodically verify the integrity and timeliness of the point cloud stream. Upon detecting anomalies such as sensor failure, communication loss, or abnormally sparse point clouds, the system will immediately trigger an alert and report the fault status, thereby improving the overall robustness and maintainability of the system.

By addressing these limitations and pursuing the proposed directions, the APEM framework can be refined into a more robust, adaptable, and practical solution for underground mining personnel safety.

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

Funding: This work was supported by the Research Fund of The State Key Laboratory for Fine Exploration and Intelligent Development of Coal Resources, CUMT (SKLCRSM20KF003), the Na-

tional Natural Science Foundation of China under grant 52374168, the National Key Research and Development Program of China under grants 2022YFC2904105 and 2023YFC2907403, the Science and Technology Innovation Program of Hunan Province under grant 2023RC3069, and in part by the Fundamental Research Funds for the Central Universities of Central South University under Grant 2025ZZTS0544.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request, due to restrictions imposed by the collaborative mining enterprise to protect confidential production information.

Conflicts of Interest: Author Bing Cui was employed by the company Norinmining Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GPS	Global Positioning System
RFID	Radio Frequency Identification
UWB	Ultra Wide Band
APEM	Attention-Augmented PointPillars for Enhanced Mining Personnel Detection
CASA	Channel and Spatial Attention
FPN	Feature Pyramid Networks
PFN	Pillar Feature Net
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
IOU	Intersection over Union

References

1. Wang, J.; Wang, L.; Peng, P.; Jiang, Y.; Wu, J.; Liu, Y. Efficient and Accurate Mapping Method of Underground Metal Mines Using Mobile Mining Equipment and Solid-State Lidar. *Measurement* **2023**, *221*, 113581. [[CrossRef](#)]
2. Zhang, G.; Chen, C.-H.; Cao, X.; Zhong, R.Y.; Duan, X.; Li, P. Industrial Internet of Things-Enabled Monitoring and Maintenance Mechanism for Fully Mechanized Mining Equipment. *Adv. Eng. Inform.* **2022**, *54*, 101782. [[CrossRef](#)]
3. Gu, Q.; Liu, L.; Bai, G.; Li, K.; Meng, Y.; Ma, F. Longitudinal and Lateral Trajectory Planning for the Typical Duty Cycle of Autonomous Load Haul Dump. *IEEE Access* **2019**, *7*, 126679–126695. [[CrossRef](#)]
4. Ranjith, P.G.; Zhao, J.; Ju, M.; Silva, R.V.S.D.; Rathnaweera, T.D.; Bandara, A.K.M.S. Opportunities and Challenges in Deep Mining: A Brief Review. *Engineering* **2017**, *3*, 546–551. [[CrossRef](#)]
5. Godwin, A.A.; Eger, T.R.; Salmoni, A.W.; Dunn, P.G. Virtual Design Modifications Yield Line-of-Sight Improvements for LHD Operators. *Int. J. Ind. Ergon.* **2008**, *38*, 202–210. [[CrossRef](#)]
6. Eger, T.; Salmoni, A.; Whissell, R. Factors Influencing Load–Haul–Dump Operator Line of Sight in Underground Mining. *Appl. Ergon.* **2004**, *35*, 93–103. [[CrossRef](#)]
7. Schwabe, N.; Godwin, A. Implications of Articulating Machinery on Operator Line of Sight and Efficacy of Camera Based Proximity Detection Systems. *Safety* **2017**, *3*, 17. [[CrossRef](#)]
8. Anee, R.-E.-A.; Karmakar, N.C. Anti-Collision Methods for Chipless RFID Systems. In Proceedings of the 2015 Asia-Pacific Microwave Conference (APMC), Nanjing, China, 6–9 December 2015; Volume 2, pp. 1–3.
9. Lasantha, L.; Karmakar, N.C.; Ray, B. Chipless RFID Sensors for IoT Sensing and Potential Applications in Underground Mining—A Review. *IEEE Sens. J.* **2023**, *23*, 9033–9048. [[CrossRef](#)]
10. Ruff, T.M.; Hession-Kunz, D. Application of Radio-Frequency Identification Systems to Collision Avoidance in Metal/Nonmetal Mines. *IEEE Trans. Ind. Appl.* **2001**, *37*, 112–116. [[CrossRef](#)]
11. Chae, S.; Yoshida, T. Application of RFID Technology to Prevention of Collision Accident with Heavy Equipment. *Autom. Constr.* **2010**, *19*, 368–374. [[CrossRef](#)]
12. Tong, H.; Xin, N.; Su, X.; Chen, T.; Wu, J. A Robust PDR/UWB Integrated Indoor Localization Approach for Pedestrians in Harsh Environments. *Sensors* **2020**, *20*, 193. [[CrossRef](#)] [[PubMed](#)]

13. Cheng, T.; Venugopal, M.; Teizer, J.; Vela, P.A. Performance Evaluation of Ultra Wideband Technology for Construction Resource Location Tracking in Harsh Environments. *Autom. Constr.* **2011**, *20*, 1173–1184. [\[CrossRef\]](#)
14. Xiao, D.; Shan, F.; Li, Z.; Le, B.T.; Liu, X.; Li, X. A Target Detection Model Based on Improved Tiny-Yolov3 Under the Environment of Mining Truck. *IEEE Access* **2019**, *7*, 123757–123764. [\[CrossRef\]](#)
15. Wang, L.; Li, W.; Zhang, Y.; Wei, C. Pedestrian Detection Based on YOLOv2 with Skip Structure in Underground Coal Mine. In Proceedings of the 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 3–5 October 2017; pp. 1216–1220.
16. Wei, X.; Zhang, H.; Liu, S.; Lu, Y. Pedestrian Detection in Underground Mines via Parallel Feature Transfer Network. *Pattern Recognit.* **2020**, *103*, 107195. [\[CrossRef\]](#)
17. Guan, S.; Wang, J.; Wang, X.; Zhang, B.; Liang, H. Improved RefineDNet Algorithm for Precise Environmental Perception of Autonomous Earthmoving Machinery under Haze and Fugitive Dust Conditions. *Adv. Eng. Inform.* **2024**, *59*, 102326. [\[CrossRef\]](#)
18. Li, X.; Wang, S.; Liu, B.; Chen, W.; Fan, W.; Tian, Z. Improved YOLOv4 Network Using Infrared Images for Personnel Detection in Coal Mines. *J. Electron. Imaging* **2022**, *31*, 013017. [\[CrossRef\]](#)
19. Dickens, J.S.; van Wyk, M.A.; Green, J.J. Pedestrian Detection for Underground Mine Vehicles Using Thermal Images. In Proceedings of the IEEE Africon' 11, Livingstone, Zambia, 13–15 September 2011; pp. 1–6.
20. Xu, P.; Zhou, Z.; Geng, Z. Safety Monitoring Method of Moving Target in Underground Coal Mine Based on Computer Vision Processing. *Sci. Rep.* **2022**, *12*, 17899. [\[CrossRef\]](#)
21. Zhou, X.; Chen, K.; Zhou, Q. Human Tracking by Employing the Scene Information in Underground Coal Mines. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
22. Sammarco, J.J.; Lutz, T. Visual Performance for Incandescent and Solid-State Cap Lamps in an Underground Mining Environment. *IEEE Trans. Ind. Appl.* **2011**, *47*, 2301–2306. [\[CrossRef\]](#)
23. Wang, J.; Li, D.; Long, Q.; Zhao, Z.; Gao, X.; Chen, J.; Yang, K. Real-Time Semantic Segmentation for Underground Mine Tunnel. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108269. [\[CrossRef\]](#)
24. Yokoyama, H.; Date, H.; Kanai, S.; Takeda, H. Pole-like Objects Recognition from Mobile Laser Scanning Data Using Smoothing and Principal Component Analysis. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, XXXVIII-5-W12, 115–120. [\[CrossRef\]](#)
25. Zermas, D.; Izzat, I.; Papanikolopoulos, N. Fast Segmentation of 3D Point Clouds: A Paradigm on LiDAR Data for Autonomous Vehicle Applications. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 5067–5073.
26. Nan, Z.; Zhu, G.; Zhang, X.; Lin, X.; Yang, Y. A Novel High-Precision Railway Obstacle Detection Algorithm Based on 3D LiDAR. *Sensors* **2024**, *24*, 3148. [\[CrossRef\]](#)
27. Wang, K.; Zhang, Z.; Wu, X.; Zhang, L. Multi-Class Object Detection in Tunnels from 3D Point Clouds: An Auto-Optimized Lazy Learning Approach. *Adv. Eng. Inform.* **2022**, *52*, 101543. [\[CrossRef\]](#)
28. Liu, H.; Du, J.; Zhang, Y.; Zhang, H. Extracting Geometric and Semantic Point Cloud Features with Gateway Attention for Accurate 3D Object Detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106227. [\[CrossRef\]](#)
29. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation 2017. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
30. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Dutchess County, NY, USA, 2017; Volume 30.
31. 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; IEEE: Salt Lake City, UT, USA, 2018; pp. 4490–4499.
32. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*, 3337. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud 2019. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
34. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3DSSD: Point-Based 3D Single Stage Object Detector. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11037–11045.
35. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10529–10538.
36. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3D Object Detection and Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Nashville, TN, USA, 2021; pp. 11779–11788.
37. Huang, Z.; Wang, Y.; Wen, J.; Wang, P.; Cai, X. An Object Detection Algorithm Combining Semantic and Geometric Information of the 3D Point Cloud. *Adv. Eng. Inform.* **2023**, *56*, 101971. [\[CrossRef\]](#)
38. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.

39. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. TANet: Robust 3D Object Detection from Point Clouds with Triple Attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11677–11684. [\[CrossRef\]](#)
40. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19.
41. Rao, Y.; Zhao, W.; Tang, Y.; Zhou, J.; Lim, S.-N.; Lu, J. HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10353–10366.
42. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Honolulu, HI, USA, 2017; pp. 936–944.
43. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. *arXiv* **2018**, arXiv:1803.02155. [\[CrossRef\]](#)
44. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
45. Kong, W.; Du, Y.; He, L.; Li, Z. Improved 3D Object Detection Based on PointPillars. *Electronics* **2024**, *13*, 2915. [\[CrossRef\]](#)
46. Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; Li, H. Voxel R-CNN: Towards High Performance Voxel-Based 3D Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1201–1209. [\[CrossRef\]](#)
47. Liu, Y.; Pan, X.; Zhu, J. 3D Pedestrian Detection Based on Pointpillar: SelfAttention-Pointpillar. In Proceedings of the 2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 21–23 November 2024; Volume 9, pp. 1–10.
48. Zhang, L.; Meng, H.; Yan, Y.; Xu, X. Transformer-Based Global PointPillars 3D Object Detection Method. *Electronics* **2023**, *12*, 3092. [\[CrossRef\]](#)
49. Wang, S.; Chen, M. A LiDAR Multi-Object Detection Algorithm for Autonomous Driving. *Appl. Sci.* **2023**, *13*, 12747. [\[CrossRef\]](#)
50. Zhang, Y.; Wang, L.; Dai, Y. PLOT: A 3D Point Cloud Object Detection Network for Autonomous Driving. *Robotica* **2023**, *41*, 1483–1499. [\[CrossRef\]](#)

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.