*electronics*

MDPI

*Article*

# Robust 3D Target Detection Based on LiDAR and Camera Fusion

**Miao Jin [1,*], Bing Lu [1], Gang Liu [2], Yinglong Diao [1], Xiwen Chen [1] and Gaoning Nie [1]**

[1] China Electric Power Research Institute, Wuhan 430074, China; lubing@epri.sgcc.com.cn (B.L.); diaoyinglong@epri.sgcc.com.cn (Y.D.); niegaonign@epri.sgcc.com.cn (G.N.)
[2] Measurement Center, State Grid Sichuan Electric Power Co., Ltd., Chengdu 610045, China; gangliusc@163.com
* Correspondence: jinmiaohb@126.com

**Abstract**

Autonomous driving relies on multimodal sensors to acquire environmental information for supporting decision making and control. While significant progress has been made in 3D object detection regarding point cloud processing and multi-sensor fusion, existing methods still suffer from shortcomings—such as sparse point clouds of foreground targets, fusion instability caused by fluctuating sensor data quality, and inadequate modeling of cross-frame temporal consistency in video streams—which severely restrict the practical performance of perception systems. To address these issues, this paper proposes a multimodal video stream 3D object detection framework based on reliability evaluation. Specifically, it dynamically perceives the reliability of each modal feature by evaluating the Region of Interest (RoI) features of cameras and LiDARs, and adaptively adjusts their contribution ratios in the fusion process accordingly. Additionally, a target-level semantic soft matching graph is constructed within the RoI region. Combined with spatial self-attention and temporal cross-attention mechanisms, the spatio-temporal correlations between consecutive frames are fully explored to achieve feature completion and enhancement. Verification on the nuScenes dataset shows that the proposed algorithm achieves an optimal performance of 67.3% and 70.6% in terms of the two core metrics, mAP and NDS, respectively—outperforming existing mainstream 3D object detection algorithms. Ablation experiments confirm that each module plays a crucial role in improving overall performance, and the algorithm exhibits better robustness and generalization in dynamically complex scenarios.

**Keywords:** 3D target detection; reliability evaluation; multi-sensor fusion; deep learning

check for updates

Academic Editor: Kefeng Ji

Received: 16 July 2025
Revised: 23 September 2025
Accepted: 24 September 2025
Published: 27 October 2025

**Citation:** Jin, M.; Lu, B.; Liu, G.; Diao, Y.; Chen, X.; Nie, G. Robust 3D Target Detection Based on LiDAR and Camera Fusion. *Electronics* **2025**, *14*, 4186. https://doi.org/10.3390/electronics14214186

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

In the field of autonomous driving, accurate and reliable 3D object detection is one of the key technologies to achieve safe driving and intelligent decision making of vehicles. With the continuous development of autonomous driving technology, the requirements for 3D target detection algorithms are getting higher and higher. Traditional 3D object detection methods mainly rely on single sensor data, such as lidar or camera, but these methods have certain limitations in the face of complex environments. For example, the lidar point cloud data are sparse, and it is difficult to accurately describe the details of the object; although the camera image can provide rich semantic information, it lacks depth information and is easily affected by factors such as illumination. Therefore, in recent years, the 3D target detection method based on multimodal sensor fusion has attracted more

and more attention. By integrating the advantages of different sensors, the accuracy and robustness of detection is expected to be improved, as shown in Figure 1.

However, the existing multimodal 3D object detection methods still face some challenges. On the one hand, the data quality of different sensors may fluctuate due to environmental factors, resulting in unstable fusion results. On the other hand, the temporal information between consecutive frames in the video stream data has not been fully utilized, and this temporal information is of great significance for understanding the motion state of the target and improving the coherence of detection. In order to overcome these challenges, this paper proposes a multimodal video stream 3D object detection algorithm framework based on reliability evaluation. It aims to dynamically adjust the proportion of multimodal feature fusion by means of reliability evaluation and confidence weighted feature fusion, and use time series information to enhance the performance of object detection.

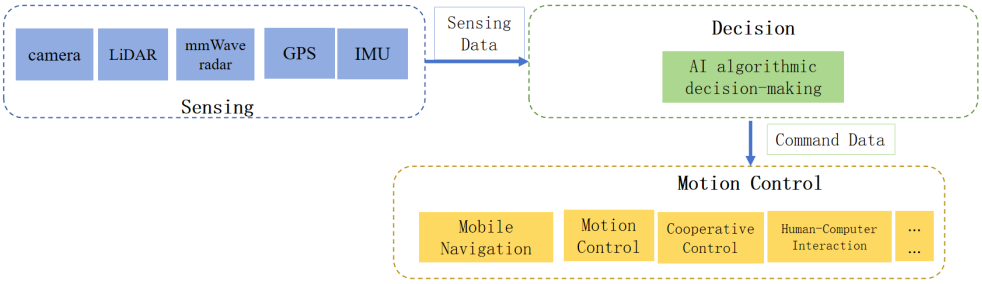



**Figure 1.** Intelligent perception and decision-making control loop.

In autonomous driving and intelligent perception systems, a single sensor modality (such as pure vision or pure LiDAR) is highly prone to failure when encountering occlusion, extreme weather, or hardware malfunctions. This leads to the loss of critical environmental information, which in turn causes misjudgments in the control system and breaks the decision-making chain. Multi-modal sensor fusion significantly improves the redundancy and robustness of the system by integrating the complementary advantages of heterogeneous data from cameras, LiDAR, and other sensors, as shown in Figure 1. However, achieving efficient and stable deep fusion still faces core challenges: it is necessary to solve the alignment problem of multi-source data in terms of time sequence, space, and semantics, and design an adaptive architecture to cope with dynamic environments. This paper aims to study a deep fusion method based on reliability assessment. By dynamically quantifying the confidence and quality indicators of each modal data, the fusion weights are adaptively adjusted, thereby achieving more efficient and stable acquisition of multimodal perception information in complex scenarios and providing technical support for safety-critical applications.

The core contribution of this paper lies in the synergy of three key modules: First, the Region of Interest (RoI)-level reliability assessment module dynamically quantifies the quality and confidence of each regional feature from cameras and LiDAR, providing refined weight guidance for the fusion process. Second, the confidence-weighted feature fusion module uses the evaluated reliability as weights and leverages the cross-attention mechanism to prioritize the integration of high-confidence feature data, effectively suppressing the impact of noise and abnormal inputs. Finally, the cross-frame optimization module that does not require explicit geometric reconstruction enhances the current frame's perception results using historical information by enforcing the consistency of temporal and spatial features, which significantly improves the model's robustness in complex scenarios such as occlusion and sensor absence.

## 2. Related Work

### 2.1. 3D Target Detection Technology Based on Pure Laser Radar

Point cloud is three-dimensional data obtained by sensors such as LiDAR, which can accurately describe the spatial position of objects in the environment. Since the laser radar measures the spatial position of the object by transmitting pulses and receiving reflected signals, the point cloud data obtained by the laser radar are sparse, disorderly, and uneven. Sparsity leads to a large number of blank areas in the point cloud, increasing computational complexity. The disorder makes it difficult for the point cloud to be directly applied to the traditional regularized data structure, and the inhomogeneity means that the density distribution of the point cloud in different regions is inconsistent, which affects the detection accuracy. According to different processing methods, point cloud processing methods are mainly divided into three types: point-based methods, the voxel-based method, and the cylinder-based method.

The point-based method directly learns the geometric features of the object from the unordered point cloud, retaining the original structural information of the point cloud. Its core idea is to enable the model to capture the spatial distribution and local morphology of the target object in the point-cloud data by layer-by-layer sampling and feature extraction. PointNet [1] pioneered the direct processing of disordered point cloud data, breaking through the dependence of traditional deep learning methods on regularized data structures. PointRCNN [2] adopts the Farthest Point Sampling (FPS) method to ensure global coverage and reduce computational overhead in a stepwise downsampling manner. The DBQ-SSD proposed by Yang et al. [3] adopts a dynamic feature aggregation mechanism so that the model can adaptively extract key information in local areas, reduce computational redundancy, and maintain detection accuracy.

The voxel-based method divides the point-cloud data into uniform voxel grids, and assigns each point in the three-dimensional space to the corresponding voxel according to its position, thereby transforming the sparse and disordered point-cloud data into a more regular data structure. On this basis, the convolution operation is used to extract voxel features to accelerate subsequent target detection. Vote3D proposed by Wang et al. [4] uses a sliding window method based on a voting mechanism to aggregate adjacent information in the voxel space, improving the expression ability of the characteristic of the target area. SECOND [5] proposed a spatial sparse convolutional network, which significantly improves the operation speed. The Associate-3Ddet model proposed by Du et al. [6] uses domain adaptation methods to provide new ideas for cross-domain feature learning and multimodal data fusion. Although voxelization can reduce the complexity of point-cloud data and improve computational efficiency, it still faces the problem of information loss caused by voxelization. Therefore, how to balance the accuracy and computational efficiency of voxelization is still the main challenge for this method.

The cylinder-based method is an optimization of traditional voxelization, which reduces the computational burden by simplifying point cloud data into vertical cylinders. PointPillars [7] is a single-stage 3D object detection method, which significantly improves the computational efficiency by learning features on the cylinder to predict the 3D orientation frame of the object. Wang et al. [8] achieved target localization by directly predicting the bounding box parameters for each cylinder, avoiding the limitation of relying on anchor frames in traditional methods. Zhang et al. [9] proposed a semi-supervised single-stage 3D object detection model, adopted the teacher–student framework, and improved the efficiency of the teacher–student network by introducing Exponential Moving Average (EMA) and asymmetric data enhancement methods. Effective progress has been made in reducing the time and cost of dataset labeling and annotation.

### 2.2. 3D Target Detection Technology Based on Multi-Sensor Fusion

Due to the sparse, disordered, and uneven characteristics of point cloud data, they usually face the challenges of information sparsity and spatial geometric modeling in processing. Different from point clouds, camera images can provide rich semantic information, such as color, texture, and shape features of objects, which can effectively supplement the shortcomings of point cloud data in detail expression. Based on this, the current mainstream multi-sensor fusion methods mainly include pre-fusion, deep fusion, and post-fusion.

Pre-fusion refers to the early fusion of data from different sensors at the input layer, usually by directly combining point cloud and image data into the network for joint learning. F-PointNet [10] is a cascaded fusion network that uses CNN to generate 2D region proposals from RGB images and extend them to 'Frustum point clouds', followed by object instance segmentation and 3D bounding box regression to accurately detect the position and size of the target object. PointPainting [11] proposed a sequential fusion method, which projects the point cloud onto the output of the image semantic segmentation network, and appends the classification score to each point to make up for the lack of semantic information of the point cloud data. By using 2D detection results to generate dense 3D virtual points, MVP [12] enhances the sparse LiDAR point cloud and improves the performance and accuracy of 3D recognition.

Deep fusion extracts high-level feature representations from different data types by designing a special network architecture, and performs fine fusion in the feature space to achieve deep integration of image and point cloud features. MV3D [13] efficiently generates candidate boxes by using 3D point clouds represented by a bird's-eye view. By designing a deep fusion scheme combined with regional features from multiple perspectives, the interaction between different path intermediate layers is realized, thereby enhancing the fusion effect of multimodal data. ContFuse [14] designed an end-to-end learnable architecture that uses continuous convolution to fuse images and point cloud feature maps at different resolutions to achieve more accurate and efficient multi-sensor deep fusion. TransFusion [15] generates a preliminary bounding box through the transformer decoder, and uses the sparse object query module to extract information from the LiDAR point cloud, and adaptively fuses the object query features with the image features, which effectively improves the accuracy and generalization of the model.

The post-fusion realizes the fusion of multimodal information at the decision-making level of the model. After the independent completion of the respective detection branches, the output results from different sensors are directly combined. It avoids the complex interaction of intermediate features or input point clouds and is efficient. CLOCs [16] proposed a low-complexity multimodal fusion framework, which significantly improves the performance of single-modal detectors by fusing the candidate boxes of 2D and 3D detectors before non-maximum suppression. Based on the CLOCs network, Fast-CLOCs [17] introduced a lightweight 3D detector-guided 2D image detector (3D-Q-2D), which effectively improved the 3D detection accuracy and reduced the complexity of the network. Sec-CLOCs [18] combines the results of the improved YOLOv8s 2D detector and the SECOND 3D detector, which significantly improves the target detection performance under severe weather conditions.

### 2.3. Three-Dimensional Target Detection Technology Based on Multimodal Continuous Frames

In autonomous driving, the camera captures video streams rather than individual frames, which means that rich temporal information can be directly used for 3D object detection. Unlike traditional single-frame image methods, continuous frame sequences can provide more context and dynamic information, and have significant advantages in tracking object motion, predicting future state, and improving detection robustness. However, it

also has the problems of inter-frame information inconsistency, motion blur, and occlusion in fast-changing scenes, which will affect the detection accuracy and reliability. In order to avoid its shortcomings while using continuous frame data to enrich temporal information, researchers continue to explore methods that combine temporal information and spatial features to improve the accuracy and robustness of 3D object detection. According to the different ways of temporal information fusion, the current research can be divided into the method based on BEV perspective, the method based on Proposal, and the method based on Query.

The method based on BEV perspective aligns the BEV features of the previous frame by affine transformation to enhance the BEV features of the current frame by temporally fusing dense BEV features. Mgtanet [19] introduced short-term motion-aware voxel coding and long-term motion-guided BEV feature enhancement, and realized the fusion of multi-frame point cloud data from the perspective of BEV. Yin et al. [20] enhanced the attention to small objects and optimized the alignment of moving objects by transmitting network coding short-term time information through grid information. INT [21] proposed a framework based on streaming training and prediction, which can use unlimited frames for multi-frame fusion without increasing computational and memory costs. However, not all pixel-level BEV features have a positive effect on improving the detection performance of foreground objects. Some background regions or unrelated pixels may introduce noise and interfere with the accuracy of object detection. Therefore, how to effectively filter the favorable features for foreground object detection while excluding irrelevant information has become the key to improve the detection performance.

The proposal-based method usually relies on the region proposal network to generate a large number of 3D proposal boxes, and achieves temporal fusion by focusing on the features of foreground objects, which are extracted by computationally intensive 3D RoI operations. MPPNet [22] proposed a three-level framework to achieve efficient processing of long-term sequence point clouds and accurate fusion of multi-frame trajectory features. MSF [23] first generates a 3D proposal in the current frame and propagates it to the previous frame according to the estimation speed, and then realizes multi-frame fusion by pooling the points of interest and encoding them as proposal features. However, the effect of timing fusion is often limited by the quality of 3D proposal boxes. Low-quality proposal boxes may introduce unnecessary redundancy or noise, thus affecting the precise positioning and recognition of objects. Therefore, improving the accuracy and denoising ability of the 3D proposal box is crucial for achieving more effective timing fusion.

The query-based method continuously iteratively optimizes the query features by introducing a query mechanism to achieve feature extraction and target localization. By proposing a motion-guided temporal modeling (MTM) module, QTNet [24] reliably constructs the correlation between adjacent frames and achieves efficient multi-frame fusion; by introducing the time alignment mechanism, StreamLTS [25] can effectively deal with the problem of LiDAR sensor asynchrony and reduce the positioning error in the process of data fusion. EfficientQ3M [26] combines the 'modal balance' Transformer decoder so that Query can access all sensor modes in the decoder, which improves the performance of multimodal data in multi-frame fusion based on Query. Although the Query mechanism implements the feature similarity between the current frame and the previous frame query through cross-attention operations, the accuracy of this feature measure may be limited because it is difficult to distinguish 3D objects with similar geometric structures in adjacent frames.

This algorithm framework achieves structural enhancement and information completion under the problem of point cloud data sparsity. It effectively integrates generative networks into discriminative 3D object detection tasks, thereby improving the model's

ability to represent the geometric structure of sparse point clouds. On this basis, it further accomplishes robust multimodal fusion under the uncertainty caused by fluctuations in modal data quality in complex dynamic environments, significantly enhancing the algorithm's robustness in real-world scenarios. Additionally, it realizes the modeling of video stream temporal information in embodied perception scenarios, effectively exploring the continuous states and motion features of targets in the temporal dimension. Ultimately, this ensures the temporal stability and consistency of detection results, thus constructing a multimodal 3D object detection solution suitable for general autonomous driving scenarios, as shown in Figure 2.

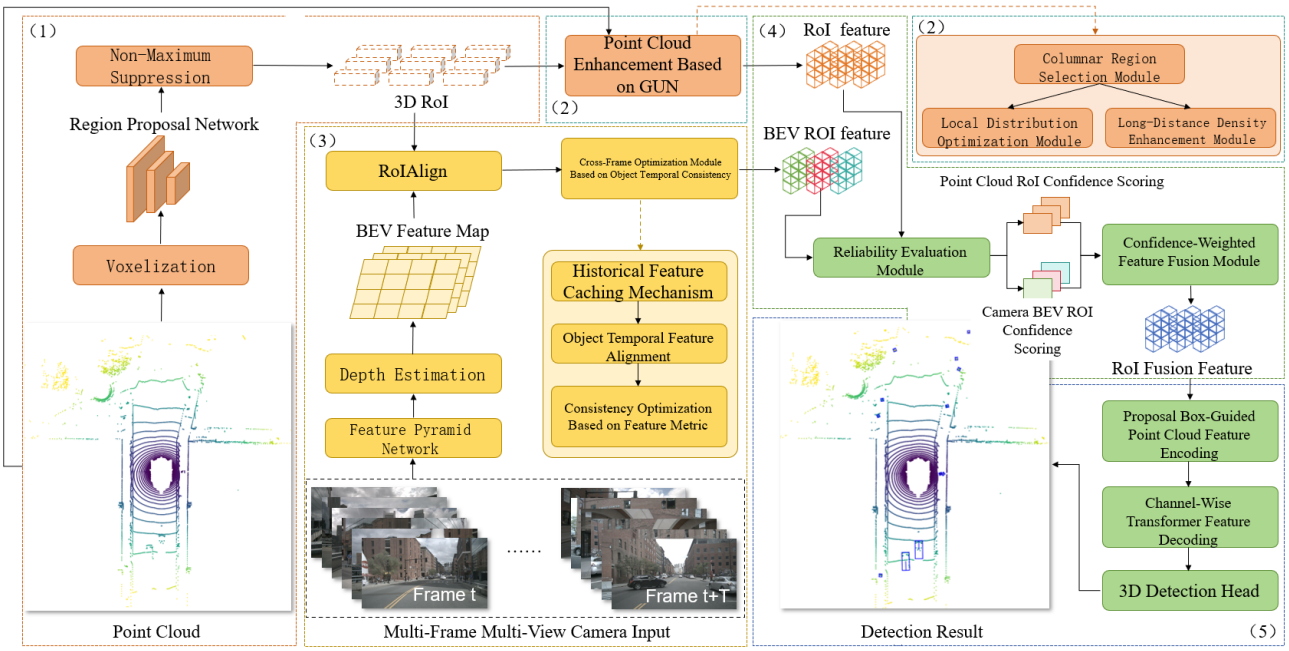



**Figure 2.** Detection algorithm framework diagram.

## 3. Multimodal Video Stream 3D Object Detection Based on Reliability Evaluation

### *3.1. Reliability Evaluation Module*

In multimodal target detection, the quality of sensor data has a crucial impact on the detection performance. In the real complex environment, the quality of sensor data inevitably fluctuates or degrades. The direct and simple fusion of point cloud and image BEV (bird's-eye view) features may lead to low-quality modal information interference detection results and reduce the performance and stability of the overall system. Therefore, we need to evaluate the reliability of sensor data. In this study, a reliability evaluation module is proposed to quantify the feature quality of the camera and LiDAR in each candidate Region of Interest (RoI). The module provides a confidence score for the subsequent fusion process by analyzing the integrity and consistency of the sensor data. Specifically, for each RoI, the module calculates the confidence scores of point cloud and image features to dynamically adjust the fusion strategy to ensure that high detection accuracy can still be maintained when data quality declines.

Assuming that the 3D target detection network based on LiDAR predicts the candidate region (RoI) in the first stage, the point cloud feature is expressed as $F_{RoI-LiDAR}$, and the image BEV feature is expressed as $F_{RoI-Camera}$. We first stitch the RoI features of these two modes to construct a joint feature representation, as follows:

$$F_{RoI} = concat(F_{RoI-LiDAR}, F_{RoI-Camera}) \quad (1)$$

The purpose of the joint feature representation is to jointly evaluate the accuracy of the RoI predicted by the point cloud network from the two dimensions of point cloud spatial features and image appearance semantic features, and then evaluate whether the current point cloud data are reliable. Then, based on this joint representation, we calculate the overall confidence score of RoI prediction, as follows:

$$C_{RoI} = \sigma(W_{RoI}F_{RoI} + b_{RoI}) \tag{2}$$

At the same time, in order to cope with the extreme situation of serious degradation of point cloud quality, we further introduce the full image BEV feature representation $F_{BEV-Camera}$ and calculate its global confidence score, as follows:

$$C_{BEV} = \sigma(W_{BEV}F_{BEV-Camera} + b_{BEV}) \tag{3}$$

To evaluate whether the camera BEV feature has sufficient reliability when the overall quality of the point cloud RoI is insufficient, as an alternative detection scheme.

When the overall confidence of point cloud RoI is high, we further evaluate the quality of RoI internal features in a fine-grained manner, and define the point cloud RoI feature confidence score and the camera BEV RoI feature confidence score as

$$C_{RoI-LiDAR} = \sigma(W_{RoI-LiDAR}F_{RoI-LiDAR} + b_{RoI-LiDAR}) \tag{4}$$

$$C_{RoI-Camera} = \sigma(W_{RoI-Camera}F_{RoI-Camera} + b_{RoI-Camera}) \tag{5}$$

Both $W_*$ and $b_*$ in the above formulas are learnable parameters, and $\sigma(\cdot)$ represents the Sigmoid activation function.

The specific implementation details of the above process are given in Algorithm 1.

### *3.2. Confidence Weighted Feature Fusion Module*

In the reliability evaluation module, we propose an adaptive modal feature evaluation strategy, which dynamically determines the detection scheme according to the confidence score of the overall RoI. However, even if it has been determined to enter the RoI-based local feature fusion path, the feature quality of different modalities in a specific region may still be significantly different. If simple strategies such as splicing or average fusion are directly adopted, it is impossible to dynamically adjust according to the actual feature quality, resulting in the interference information of low-confidence modes masking the effective features of high-quality modes, thereby affecting the overall detection performance.

In order to achieve a more discriminative and adaptive fusion strategy, this study designed a confidence weighted feature fusion module. The module performs weighted fusion of point cloud and image features according to the confidence score provided by the reliability evaluation module. The modal features with high confidence will occupy a greater weight in the fusion process, while the modal features with low confidence will be appropriately suppressed. In this way, the fused features can more accurately reflect the geometric structure and semantic information of the target, thereby improving the robustness and accuracy of detection. In addition, the module also introduces a cross-attention mechanism to further enhance the complementarity between point cloud and image features and improve the fusion effect. The structure of the designed confidence weighted feature fusion module is shown in Figure 3.

**Algorithm 1** Reliability Assessment Process (Reliability Evaluation Module Pseudo-code)

**Input:**
The predicted RoI set from the first-stage LiDAR network $R_{LiDAR} \in \mathbb{R}^N$, where $N$ is the number of RoIs;
Point cloud RoI features $F_{RoI-LiDAR} \in \mathbb{R}^{N \times M \times C_1}$, where $M$ is the number of points in each RoI, and $C_1$ is the feature dimension;
Image BEV RoI features $F_{RoI-Camera} \in \mathbb{R}^{N \times H_R \times W_R \times C_2}$, where $H_R, W_R$ are the dimensions of the RoI extracted from the image BEV, and $C_2$ is the feature dimension;
Image BEV features $F_{BEV-Camera} \in \mathbb{R}^{N \times H \times W \times C_3}$, where $H, W$ are the dimensions of the image BEV, and $C_3$ is the feature dimension;
The predicted confidence threshold from the first-stage LiDAR network $T_{RoI}$;
The confidence threshold from the camera data $T_{Camera}$;
**Reliability Assessment Process:**
1. Pool the point cloud features for each RoI: $F_{RoI-LiDAR-Pool} = \text{Pool}(F_{RoI-LiDAR}) \rightarrow \mathbb{R}^{N \times C_1}$
2. Pool the image BEV RoI features for each RoI: $F_{RoI-Camera-Pool} = \text{Pool2D}(F_{RoI-Camera}) \rightarrow \mathbb{R}^{N \times C_2}$
3. Concatenate multi-modal features to obtain a joint feature representation: $F_{RoI} = \text{concat}(F_{RoI-LiDAR-Pool}, F_{RoI-Camera-Pool}) \rightarrow \mathbb{R}^{N \times (C_1 + C_2)}$
4. Calculate the confidence score for each predicted RoI: $C_{RoI} \in \mathbb{R}^N$
5. Calculate the overall confidence score for RoI: $C_{RoI_mean} = \text{mean}(C_{RoI}) \in \mathbb{R}^1$
**Calculate the global confidence score of image BEV features** $C_{BEV} \in \mathbb{R}^1$
**if** $C_{RoI_mean} \geq T_{RoI}$ **then**
  **for** each $r \in R_{LiDAR}$ **do**
    **if** $C_{RoI}[r] \geq T_{RoI}$ **then**
      1. Calculate the point cloud RoI confidence score $C_{RoI-LiDAR} \in \mathbb{R}^1$
      2. Calculate the camera BEV RoI confidence score $C_{RoI-Camera} \in \mathbb{R}^1$
      3. Input $F_{RoI-LiDAR}[r]$ and $F_{RoI-Camera}[r]$ into the subsequent confidence-weighted feature fusion module
    **else**
      Discard the current RoI features $F_{RoI-LiDAR}[r]$ and $F_{RoI-Camera}[r]$
    **end if**
  **end for**
**else if** $C_{RoI_mean} < T_{RoI}$ and $C_{BEV} \geq T_{Camera}$ **then**
  Input $F_{BEV-Camera}$ into the pure image BEV detection module
**else**
  The current sensor data are unreliable, do not perform detection
**end if**

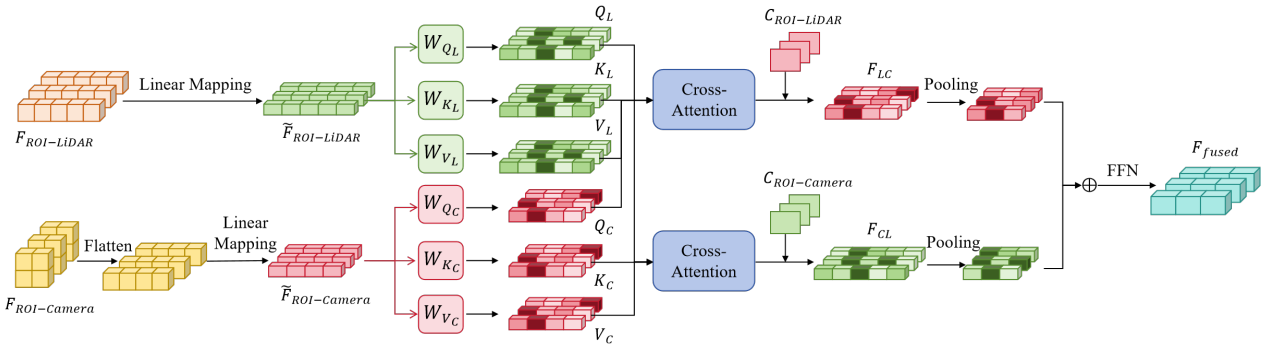



**Figure 3.** Confidence weighted feature fusion module schematic diagram.

Specifically, let the point cloud feature in the $i$ RoI region be $F^{(i)}_{RoI-LiDAR} \in \mathbb{R}^{M \times C_1}$ and the image BEV feature be $F^{(i)}_{RoI-Camera} \in \mathbb{R}^{H_R \times W_R \times C_2}$. Since the original features of the two modalities have significantly different data structures and dimensions, we first serialize them in a unified way to achieve subsequent cross-modal information interaction. For the

point cloud mode, the feature itself is in the form of point sequence without additional processing; for the image modality, we expand the features in the spatial dimension into a one-dimensional sequence, which is transformed into a two-dimensional feature matrix of $[(H_R \times W_R), C_2]$. Then, we perform linear mappings on the serialized features of point clouds and images, and uniformly map them into the feature space of the same dimension $d$ to obtain

$$\widetilde{F}^{(i)}_{RoI-LiDAR} = Linear\left(F^{(i)}_{RoI-LiDAR}\right) \in \mathbb{R}^{M\times d} \tag{6}$$

$$\widetilde{F}^{(i)}_{RoI-Camera} = Linear\left(Flatten\left(F^{(i)}_{RoI-Camera}\right)\right) \in \mathbb{R}^{(H_R\cdot W_R)\times d} \tag{7}$$

On this basis, we generate the query, key, and value vectors required for the attention mechanism for each modality, which are represented as

$$Q^{(i)}_L = \widetilde{F}^{(i)}_{RoI-LiDAR}W_{Q_L}, K^{(i)}_L = \widetilde{F}^{(i)}_{RoI-LiDAR}W_{K_L}, V^{(i)}_L = \widetilde{F}^{(i)}_{RoI-LiDAR}W_{V_L} \tag{8}$$

$$Q^{(i)}_C = \widetilde{F}^{(i)}_{RoI-Camera}W_{Q_C}, K^{(i)}_C = \widetilde{F}^{(i)}_{RoI-Camera}W_{K_C}, V^{(i)}_C = \widetilde{F}^{(i)}_{RoI-Camera}W_{V_C} \tag{9}$$

Among them, $W_{Q_\sim}, W_{K_\sim}, W_{V_\sim}$ are the independent learning parameter matrices of the corresponding modes, and the dimensions are all $d \times d$, where $d = 256$.

We introduce the modal score $C^{(i)}_{RoI-LiDAR}, C^{(i)}_{RoI-Camera}$ obtained from the reliability evaluation module. In the form of cross-attention, the query vector of one modality interacts with the key and value vectors of the other modality to calculate the following:

1. Point cloud features guide the cross-attention of image feature enhancement, as follows:

$$F^{(i)}_{LC} = C^{(i)}_{RoI-LiDAR}\cdot softmax\left(\frac{Q^{(i)}_C\left(K^{(i)}_L\right)^T}{\sqrt{d}}\right)V^{(i)}_L \tag{10}$$

2. Image features guide the cross-attention of point cloud feature enhancement, as follows:

$$F^{(i)}_{CL} = C^{(i)}_{RoI-Camera}\cdot softmax\left(\frac{Q^{(i)}_L\left(K^{(i)}_C\right)^T}{\sqrt{d}}\right)V^{(i)}_C \tag{11}$$

This mechanism enables the fusion process to dynamically adjust the weight distribution of each modal feature in the interaction process based on confidence. Finally, we perform global pooling on the two cross-modal enhancement features, compress them into a vector representation of a unified dimension and add them, and input them into the Feed-Forward Network (FFN) to generate the fusion feature representation of the RoI, as follows:

$$F^{(i)}_{fused} = FFN\left(Pool\left(F^{(i)}_{LC}\right) + Pool\left(F^{(i)}_{CL}\right)\right) \tag{12}$$

This fusion feature combines geometric structure and semantic details, and strengthens the guiding role of reliable modes through the confidence weighting mechanism. Finally, it is transmitted as input to the proposal correction network based on channel-by-channel Transformer to complete target classification and bounding box correction.

### *3.3. Cross-Frame Optimization Module Based on Target Timing Consistency*

In the video stream data, the motion trajectory and appearance features of the target have certain continuity and consistency in the time dimension. This study proposes a

cross-frame optimization module based on target timing consistency, which is used to improve the stability and accuracy of detection results. The module constructs a target-level semantic soft matching graph, and combines spatial self-attention and temporal cross-attention mechanisms to mine the spatio-temporal correlation between consecutive frames to complete and enhance the target features. At the same time, the structural consistency loss is introduced to reversely optimize the cross-frame features to ensure that the detection results of the target in different frames are consistent, thereby alleviating the detection performance degradation caused by occlusion, blurring, or scale changes. The module is mainly composed of three parts: historical feature caching mechanism, target timing feature alignment, and consistency optimization based on feature measurement. Its structure is shown in Figure 4.

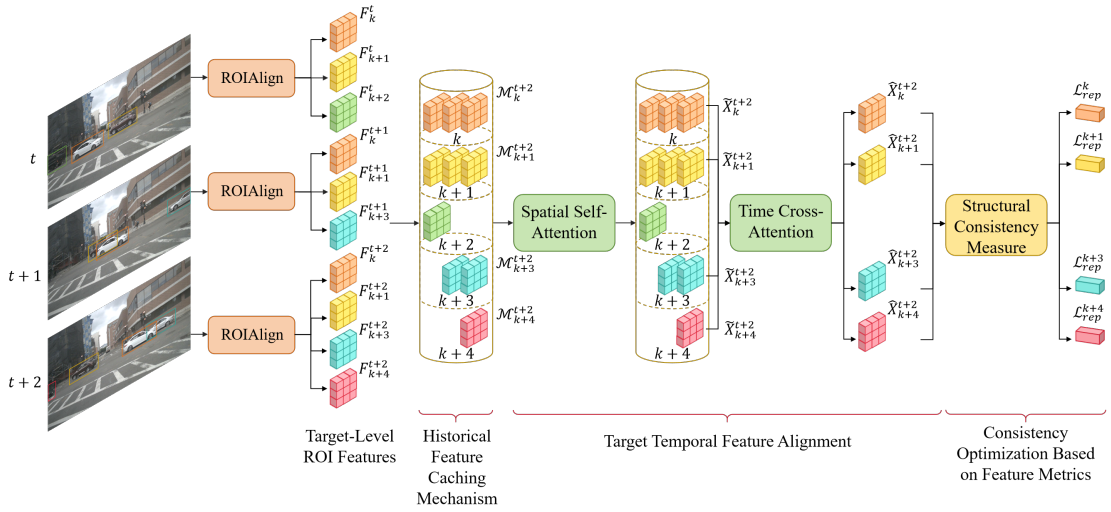



**Figure 4.** Cross-frame optimization module based on target temporal consistency.

1. Historical feature caching mechanism
   In order to achieve cross-frame target-level semantic modeling, the model needs to have the ability to perceive the target state in the historical frame. We use the RoI feature $F_k \in \mathbb{R}^{H \times W \times C}$ of the target $k$ in the BEV space as its state representation. Assuming that the current time frame is $t$ and a sliding time window with a length of $T$ is defined, the historical memory bank of the target at the current time can be expressed as

$$\mathcal{M}_k^t = \left\{ F_k^{t-T+1}, F_k^{t-T+2}, \ldots, F_k^t \right\} \tag{13}$$

   where $F_k^t$ denotes the characteristic of target $k$ at time $t$. The entire feature sequence $\mathcal{M}_k^t$ is arranged in chronological order to form a continuous historical state trajectory of the target $k$ before time $t$.
2. Target timing feature alignment
   In order to enhance the feature expression ability of the current frame target, this section introduces the spatial and temporal attention mechanism to fuse the historical feature sequence of the target. The process includes two stages: intra-frame spatial self-attention and cross-frame temporal cross-attention.
   First, $F_k^t$ is transformed into a two-dimensional feature sequence $X_k^t$, and then the spatial self-attention calculation is performed on the feature sequence as follows:

$$\widetilde{X}_k^t = AttS\left(X_k^t\right) = softmax\left(\frac{Q_S^t \left(K_S^t\right)^T}{\sqrt{d}}\right) V_S^t \tag{14}$$

After obtaining the spatial enhancement feature $\widetilde{X}_k^t$, the model needs to further model the consistent expression between the current frame and the historical frame. For the current frame $t$, we use its corresponding spatial enhancement feature as a query, and use the enhancement feature of the historical frame as a key and value for time cross-attention fusion. The calculation is as follows:

$$\widehat{X}_k^t = AttT\left(\widehat{\mathcal{M}}_k^t\right) = \sum_{l=t-T+1}^{t-1} softmax\left(\frac{Q_t^t\left(K_t^l\right)^T}{\sqrt{d}}\right)V_t^l \tag{15}$$

3. Consistency optimization based on feature measure

Although the target temporal feature alignment mechanism can enhance the semantic representation ability of the current frame target, it is difficult to ensure the structural consistency of the prediction results of the same target in different frames only relying on feature level enhancement. Because the detection process of each frame is independent of each other, the model is prone to problems such as position fluctuation and attitude jump when facing moving targets, partial occlusion, or illumination changes, which affects the rationality of the detection results and the stability of subsequent tasks. To this end, we introduce the idea of target-level structure optimization, and reversely adjust the timing prediction results based on the target cross-frame feature consistency. This idea is similar to the traditional Bundle Adjustment [27]; that is, if the target semantic key areas can correspond in different frames, their spatial projection positions should be consistent, and if inconsistent, there is an error in pose or position estimation.

Based on the above ideas, we construct a feature metric consistency mechanism based on cross-frame semantic correspondence to constrain the geometric consistency of target multi-frame prediction results. Different from the traditional bundle adjustment method, this mechanism does not need to rely on explicit geometric projection and 3D point reconstruction, but constructs cross-frame semantic consistency constraints in the feature metric space. As shown in Figure 5, by establishing the feature-level correspondence of the key regions of the target in the multi-frame view, a semantic point soft matching graph is formed, and the target pose estimation parameters are optimized to ensure that the same semantic region is aligned at each frame projection position. This mechanism directly imposes structural consistency constraints on the feature space, has differentiability, supports end-to-end training, and effectively compensates for the shortcomings of a single-frame detector in time series modeling at the structural level.

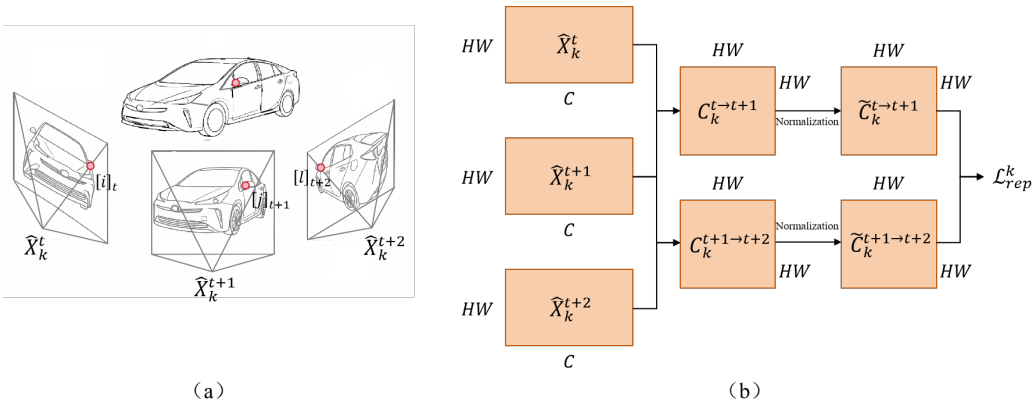



**Figure 5.** Schematic diagram of feature metric consistency mechanism based on cross-frame semantic correspondence. (**a**) Schematic diagram of the correspondence between target semantic features in multi-frame views. (**b**) Calculation of reprojection error process for semantic point feature soft matching graph.

Specifically, for any two frames $t$ and $t^{'}$, we first calculate the feature similarity between the feature points $\widehat{X}_k^t, \widehat{X}_k^{t^{'}}$ of the target $k$ in these two frames:

$$C_k^{t \to t^{'}}[i,j] = \left\langle \widehat{X}_k^t[i], \widehat{X}_k^{t^{'}}[j] \right\rangle \tag{16}$$

where $\widehat{X}_k^t[i]$ denotes the $i$-th local characteristic of $\widehat{X}_k^t$, and $i \in \{1, 2, \ldots, H \times W\}$, $\langle \cdot \rangle$ denote the inner product operation. Then, all the positions in the space are normalized to obtain the soft matching probability across frames, as follows:

$$\widetilde{C}_k^{t \to t^{'}}[i,j] = softmax\left(C_k^{t \to t^{'}}[i,j]\right) \tag{17}$$

Among them, the meaning of $\widetilde{C}_k^{t \to t^{'}}[i,j] \in \mathbb{R}^{(H \cdot W) \times (H \cdot W)}$ is the matching degree of the $i$-th feature point in frame $t$ and the $j$-th point in frame $t^{'}$ in the feature space.
We hope that the feature representation learned by the model can meet the requirements of structural consistency; that is, the semantic matching points can be aligned in three-dimensional space after the pose transformation of their respective frames. To this end, we introduce the Featuremetric Object Bundle Adjustment [28], which extends the structural consistency modeling from the traditional geometric space to the feature space. The alignment process of the target structure is indirectly constrained by the semantic similarity between features. The reprojection error can be expressed as

$$\mathcal{L}_{rep}^k = \sum_{i=1}^{H \cdot W} \sum_{t=1}^{T} \sum_{t^{'}=1}^{T} \left\| \widehat{X}_k^t[i] - \widehat{X}_k^{t^{'}}[i] \right\|_2^2 \tag{18}$$

Here, we use the feature soft matching graph $\widetilde{C}_k^{t \to t^{'}}[i,j]$ of semantic points to replace the L2 loss of Equation (18) as a measure of structural consistency, and take the logarithm of the score and add a negative sign to construct the maximum likelihood loss function as follows:

$$\mathcal{L}_{F-rep}^k = -\sum_{i=1}^{H \cdot W} \sum_{t=1}^{T} \sum_{t^{'}=1}^{T} \log\left(\widetilde{C}_k^{t \to t^{'}}[i,i]\right) \tag{19}$$

Compared with directly calculating the Euclidean distance of cross-frame semantic points, the feature soft matching graph directly comes from the feature distribution itself, so $\mathcal{L}_{F-rep}^k$ avoids the dependence on clear geometric correspondence and has better training stability. For each traceable target $k$ in the time window, a structural consistency metric loss $\mathcal{L}_{F-rep}^k$ is constructed. After introducing the loss, the overall training target of the model is updated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{RPN} + \mathcal{L}_{IoU} + \mathcal{L}_{conf} + \mathcal{L}_{reg} + \lambda_{F-rep} \sum_k \mathcal{L}_{F-rep}^k \tag{20}$$

Among them, $\lambda_{F-rep}$ is the loss weight coefficient.

## 4. Experimental Verification

### *4.1. Experiment Setting*

In order to verify the effectiveness and feasibility of the proposed multimodal video stream 3D object detection algorithm based on reliability evaluation, we conducted experiments on the nuScenes dataset [29]. nuScenes is a large-scale autonomous driving dataset designed to provide rich annotation data and diverse scenarios for multimodal sensing

tasks. In the experiment, we use the standard training and validation data partitioning method and perform performance evaluation on the validation set. The experimental environment configuration is shown in Table 1.

In the reliability evaluation module, the RoI confidence threshold $T_{RoI}$ generated by the LiDAR branch is set to 0.25, and the camera data confidence threshold $T_{Camera}$ is set to 0.38. After introducing the video stream, the weight coefficient $\lambda_{F-rep}$ of the structural consistency metric loss is set to 0.15, which is used to balance its impact on the total loss. At the same time, in order to reduce redundant interference, for the target that the number of frames in the sequence is less than 10 or the average number of key points is less than 5, it does not participate in the timing consistency optimization. The whole model is trained in an end-to-end manner. Since the multimodal algorithm takes up a large amount of memory, the batch size is set to 1, the optimizer is Adam [30], and the number of training rounds is 80 epochs. The learning rate is scheduled by a cosine annealing strategy, and the maximum learning rate is set to $5 \times 10^{-4}$.

**Table 1.** Experimental environment configuration.

| Name | Configuration |
|---|---|
| CPU | 11th Gen Intel(R) Core(TM) i7-11700K |
| GPU | NVIDIA GeForce RTX 3090 |
| Memory | 24 GB |
| Storage | 30 GB |
| Operating System | Ubuntu 20.04 |
| CUDA | 11.2 |
| Deep Learning Framework | Pytorch 1.10.0 |

### *4.2. Detection Index Results and Visual Analysis*

The systematic evaluation results on the nuScenes verification set show that the proposed method is significantly better than the existing mainstream 3D object detection algorithms in terms of overall performance and key category accuracy. The experimental results are shown in Table 2. From the perspective of overall indicators, the method in this paper achieves the best results of 67.3% and 70.6% on the two core indicators of average accuracy (mAP) and comprehensive detection score (NDS), respectively, reflecting strong global detection ability and stable multimodal fusion effect. From the performance of each category, this paper achieves the highest detection accuracy in small volume and easy occlusion categories such as pedestrians and traffic cones. In addition, the temporal consistency modeling mechanism introduced in this paper further explores the continuous motion characteristics and appearance evol. Froution trend of the target in the video stream, and enhances the stability and coherence of the model in cross-frame detection, so as to achieve better detection performance in dynamic perception scenarios.

In order to further verify the effectiveness of the image semantic feature deep fusion and cross-frame temporal consistency module in the complete multi-modal video stream perception framework proposed in this paper, we select the pure point cloud detection algorithm as the baseline for comparison. Table 3 shows the performance of the two models on the nuScenes verification set. Considering the typical target characteristics in autonomous driving scenarios, we mainly focus on the detection performance of three representative targets: vehicles, pedestrians, and cyclists.

From the perspective of overall indicators, after the introduction of image depth fusion and timing consistency modules, the mAP and NDS of the model increased by 5.2% and 2.0%, respectively, which was significantly better than the point cloud baseline. This reflects the positive effect of multi-modal fusion and video stream timing modeling strategy on the overall detection accuracy and stability. Specific to each category, although the structure of

the vehicle target is relatively stable, it often faces occlusion and background interference in the actual scene. The additional semantic features of the image modality are introduced to make the model better describe the boundary and appearance of the target. The mAP of the vehicle category is increased from 85.4% of the baseline to 88.1%. The pedestrian category is small and easy to be occluded. The continuous dynamic information is captured by the cross-frame temporal consistency modeling module, which effectively improves the detection accuracy of this kind of target from 86.4% to 89.2%. The cyclist category is particularly prominent. Due to its fine structure and obvious dynamic characteristics, the point cloud baseline model is limited in this category (27.1%). After introducing image semantic deep fusion and temporal information, the detection accuracy is improved by 20.1% compared with the baseline model, which fully verifies the modeling advantages of the complete multi-modal framework proposed in this paper on dynamic and weak structural targets.

**Table 2.** The performance comparison between the proposed algorithm and the baseline algorithm on the nuScenes verification set. We report NDS, mAP, and various types of mAP.

| Model | mAP (%) | NDS (%) | Vehicle (%) | Truck (%) | Construction Vehicle (%) | Bus (%) | Trailer (%) | Obstruction (%) | Motorcycle (%) | Cyclist (%) | Pedestrian (%) | Traffic Cone (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PointPillar [7] | 40.1 | 55.0 | 76.0 | 31.0 | 11.3 | 32.1 | 36.6 | 56.4 | 34.2 | 14.0 | 64.0 | 45.6 |
| Pointpainting [11] | 46.4 | 58.1 | 77.9 | 35.8 | 15.8 | 36.2 | 37.3 | 60.2 | 41.5 | 24.1 | 73.3 | 62.4 |
| BEVFormer [31] | 48.1 | 56.9 | 67.7 | 39.2 | 22.9 | 35.7 | 39.6 | 62.5 | 47.9 | 40.7 | 54.4 | 70.3 |
| BEVDet [32] | 42.4 | 48.8 | 64.3 | 35.0 | 16.2 | 35.8 | 35.4 | 61.4 | 44.8 | 29.6 | 41.1 | 60.1 |
| DETR3D [33] | 41.2 | 47.9 | 60.3 | 33.3 | 17.0 | 29.0 | 35.8 | 56.5 | 41.3 | 30.8 | 45.5 | 62.7 |
| 3D-CVF [34] | 52.7 | 62.3 | 83.0 | 45.0 | 15.9 | 48.8 | 49.6 | 65.9 | 51.2 | 30.4 | 74.2 | 62.9 |
| CBGS [35] | 52.8 | 63.3 | 81.1 | 48.5 | 10.5 | 54.9 | 42.9 | 65.7 | 51.5 | 22.3 | 80.1 | 70.9 |
| CT3D [36] | 59.5 | 65.3 | 84.8 | 53.7 | 19.4 | 64.2 | 55.4 | 72.1 | 59.5 | 24.8 | 83.4 | 77.9 |
| CenterPoint [37] | 60.3 | 67.3 | 85.2 | 53.5 | 20.0 | 63.6 | 56.0 | 71.1 | 59.5 | 30.7 | 84.6 | 78.4 |
| SECOND [5] | 62.0 | 50.9 | 81.6 | 51.9 | 18.0 | 68.5 | 38.2 | 57.8 | 40.1 | 18.1 | 77.4 | 56.9 |
| FUTR3D [38] | 64.2 | 68.0 | 86.3 | 61.5 | 26.0 | 71.9 | 42.1 | 64.4 | **73.6** | **63.3** | 82.6 | 70.1 |
| MVP [12] | 66.4 | 70.5 | 86.8 | 58.5 | 26.1 | 67.4 | 57.3 | 74.8 | 70.0 | 49.3 | 87.9 | 83.6 |
| **Ours** | **67.3** | **70.6** | **88.1** | **61.8** | **29.4** | **72.7** | **59.7** | **80.1** | 68.3 | 47.2 | **89.2** | **86.3** |

**Table 3.** Performance comparison between the complete multimodal perception algorithm (Ours) and the baseline point cloud model (Ours-L) on the nuScenes validation set. The reported metrics include NDS, mAP, and category-specific mAP.

| Model | mAP (%) | NDS (%) | Vehicle (%) | Pedestrian (%) | Cyclist (%) |
|---|---|---|---|---|---|
| **Ours-L** | 62.1 | 68.6 | 85.4 | 86.4 | 27.1 |
| **Ours** | 67.3 ↑ **5.2** | 70.6 ↑ **2.0** | 88.1 ↑ **2.7** | 89.2 ↑ **2.8** | 47.2 ↑ **20.1** |

As shown in Figure 5, we further visualize the detection effects of the multimodal video stream algorithm (Ours) and the pure point cloud baseline method (Ours-L) proposed in this paper on the continuous video frames of the nuScenes verification set. The experiment extracts the detection results of three consecutive frames from the same scene to analyze the performance differences of the model at different time steps. On the whole, the proposed method obviously shows stronger spatial perception and cross-frame correlation ability, especially in long-distance target perception, occlusion target detection, and target tracking stability.

First, in the area at the far end of the scene, the algorithm proposed in this paper can accurately detect the foreground target that has just entered the field of view from a distance in the first frame, while the point cloud baseline algorithm completely misses it, demonstrating the supplementary ability of the deep fusion mechanism of image semantic features in distant target perception. Subsequently, with continuous observation across frames, the detection box position of the algorithm proposed in this paper becomes more stable and accurate, reflecting the role of the introduced cross-frame temporal consistency modeling strategy in improving localization accuracy.

Second, in the area near the vehicle, the algorithm proposed in this paper can effectively perceive nearby foreground targets and accurately output bounding boxes, while the point cloud baseline algorithm exhibits significant missed detections, indicating that the depth fusion strategy possesses stronger geometric modeling capabilities in near-field multi-target dense scenes.

For complex areas, both methods face the issues of occlusion and sparse point cloud interference. However, the algorithm proposed in this paper can obtain more complete and stable detection results through the accumulation of multi-frame information. It gradually improves the bounding box prediction from the first frame to the third frame, demonstrating the advantage of information integration in the temporal dimension.

In addition, in the inter-frame continuity analysis, the fourth target is lost in the point cloud baseline algorithm, and the algorithm in this paper successfully maintains the target consistency by means of the temporal feature caching mechanism, which avoids the unstable tracking phenomenon.

The fifth point shows that the proposed algorithm outputs more accurate orientation estimation for the target, which further verifies the advantage of its feature decoding ability in direction modeling.

Finally, the sixth target cannot be recovered after being lost by the point cloud baseline algorithm in the middle frame, while the algorithm in this paper can effectively backtrack and continuously detect the target through cross-frame semantic memory, which reflects the robustness and timing perception ability of the proposed method in dealing with problems such as instantaneous occlusion or temporary missing.

### *4.3. Ablation Experiment*

The multi-modal video stream sensing algorithm proposed in this paper is based on the pure point cloud baseline method (Ours-L), and further gradually introduces the confidence weighted feature fusion module and the cross-frame timing consistency optimization module to improve the detection performance and stability of the model under complex sensing conditions. In order to verify the effectiveness of the above modules, we analyzed the specific contributions of each module through ablation experiments. The experimental results are shown in Table 4.

**Table 4.** Horizontal comparison of 3D target detection indicators in ablation experiments.

| Model | Confidence Weighted Feature Fusion Module | Cross-Frame Optimization Module Based on Target Timing Consistency | mAP (%) | NDS (%) | Vehicle (%) | Pedestrian (%) | Cyclist (%) |
|---|---|---|---|---|---|---|---|
| Ours-L | | | 62.1 | 68.6 | 85.4 | 86.4 | 27.1 |
| Ours | ✓ | | 64.8 | 69.2 | 86.7 | 87.7 | 36.4 |
| | ✓ | ✓ | **67.3** | **70.6** | **88.1** | **89.2** | **47.2** |

Compared with the point cloud baseline model, after introducing the confidence weighted feature fusion module, the mAP is increased by 2.7%, the NDS is increased by 0.6%, and the detection accuracy of the 'pedestrian' category is increased from 86.4% to 87.7%. It shows that the module introduces reliable image features, provides rich semantic context information for point cloud branches, and makes up for the lack of geometric details in sparse or structurally blurred areas of point clouds. After further adding the cross-frame optimization module based on the target timing consistency, the overall performance is greatly improved again, especially in the 'cyclist' category from

36.4% to 47.2%, reflecting the module's ability to model the continuity of the target state in a dynamic scene. Because cyclists often have small target size and unstable motion, cross-frame semantic enhancement can effectively use the features of historical frames to complete the incomplete information of the current frame and improve its detectability.

On the whole, the two modules play their respective roles in multimodal information integration and cross-time consistency modeling. The former enhances the quality of feature fusion, and the latter improves the temporal stability of detection. The synergistic effect of the two modules finally achieves a comprehensive performance gain in dynamic and complex scenes.

As shown in the Figure 6, first, in the far distal area (Area 1), the proposed algorithm detects distant new foreground targets at the first frame, while the point cloud baseline algorithm completely misses them—highlighting the deep image semantic fusion mechanism's value for long-distance perception. Its detection boxes also stabilize with inter-frame observation, thanks to the cross-frame temporal consistency strategy.

Second, in the vehicle-near area (Area 2), the proposed algorithm accurately detects nearby targets and outputs bounding boxes, whereas the baseline algorithm shows obvious misses. This confirms the stronger geometric modeling ability of the point cloud structure completion and deep fusion in near-field dense scenarios.

For the complex area (Area 3), both methods face occlusion and sparse point cloud issues. However, the proposed algorithm achieves more stable results via multi-frame information accumulation, improving bounding box prediction from Frame 1 to 3 and demonstrating temporal integration advantages. In inter-frame continuity analysis, the baseline algorithm loses the target at Area 4, but the proposed algorithm maintains consistency via temporal feature caching, avoiding unstable tracking. Area 5 shows that the proposed algorithm provides more accurate target orientation estimation, verifying its feature decoding advantage in direction modeling.

Finally, after the baseline algorithm loses the target at Area 6 (unrecoverably) in middle frames, the proposed algorithm traces back and sustains detection via cross-frame semantic memory—proving its robustness and temporal perception for transient occlusion or temporary target absence.

As shown in Table 5, experimental results show that, under the condition of severe camera input loss, the performance of all types of models decreases significantly. Pure vision methods (such as DETR3D) almost completely fail under extreme loss conditions, highlighting their strong dependence on multi-view input. Fusion methods generally exhibit better robustness, among which the method proposed in this paper performs best in terms of both performance degradation amplitude and absolute accuracy. This is attributed to its confidence-guided fusion mechanism, which can dynamically adjust modal weights and effectively suppress the negative interference from missing modalities. When dealing with lens occlusion, although traditional fusion methods can maintain basic detection capabilities, the local semantic loss caused by occlusion still leads to a significant drop in accuracy. Through cross-frame temporal consistency modeling and the historical feature enhancement mechanism, the proposed method in this paper effectively compensates for the local information loss caused by occlusion, achieving the smallest performance decline. This proves the key role of long-term temporal dependency modeling in improving the model's occlusion resistance. In conclusion, the fusion framework proposed in this study demonstrates stronger robustness when addressing the two challenges of sensor loss and local occlusion. Its core innovations lie in the design of an adaptive modal weighting strategy and a long-term temporal feature fusion mechanism, which provide an effective solution for reliable multi-modal 3D detection in complex environments.

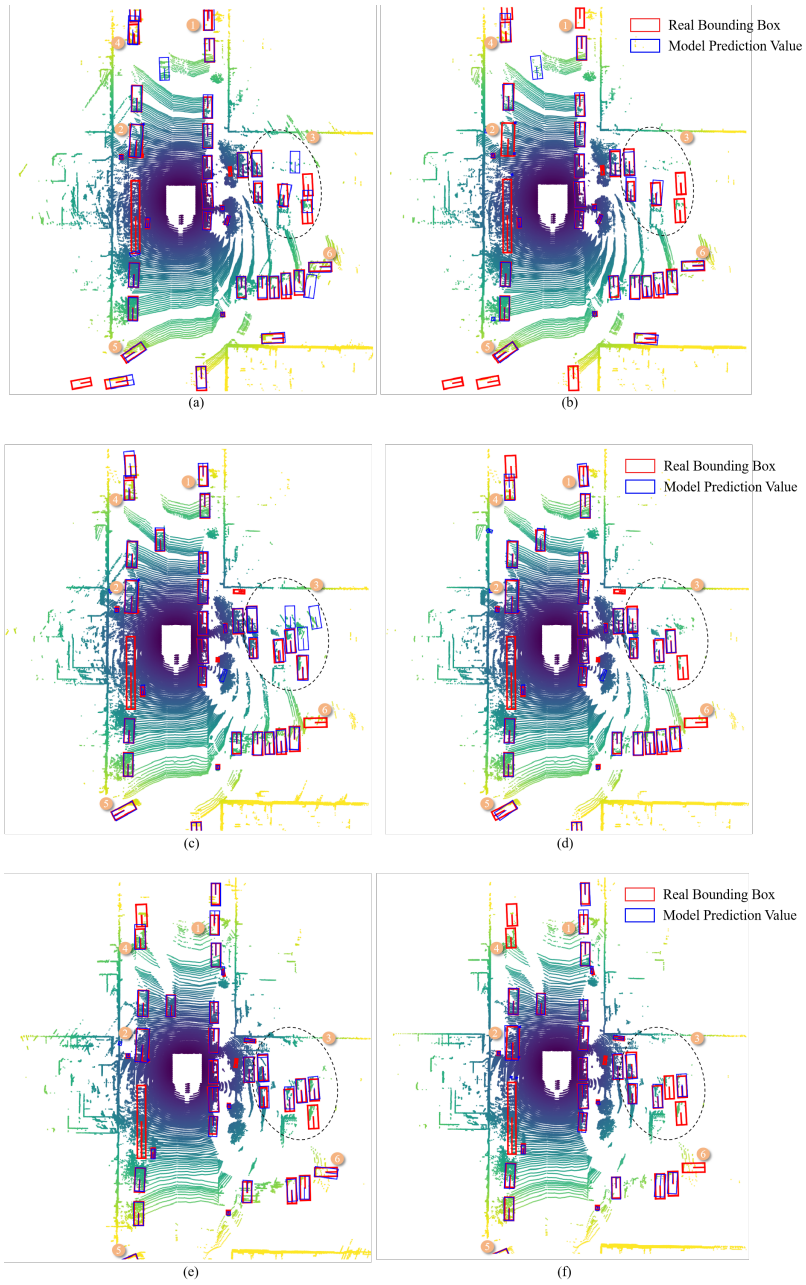



**Figure 6.** The visualization results of our algorithm and point cloud baseline model (Ours-L) on the nuScenes validation set in this article. (**a**) $t_{ours} = 1$, (**b**) $t_{ours-L} = 1$, (**c**) $t_{ours} = 2$, (**d**) $t_{ours-L} = 2$, (**e**) $t_{ours} = 3$, (**f**) $t_{ours-L} = 3$.

**Table 5.** Performance comparison of different methods on the nuScenes validation set under various camera anomaly scenarios (mAP/NDS). In input modalities, C represents pure camera input, and L + C represents the fusion of camera and LiDAR.

| Model | Input Modality | Ideal Condition | Front View Missing | Only Front View Remains | 50% Lens Occlusion |
|---|---|---|---|---|---|
| DETR3D | C | 41.2/47.9 | 30.6/41.8 | 8.8/26.3 | 18.0/32.5 |
| 3D-CVF | L + C | 52.7/62.3 | 47.3/58.1 | 36.4/49.9 | 45.2/57.2 |
| FUTR3D | L + C | 64.2/68.0 | 53.7/59.8 | 51.2/52.6 | 50.3/58.5 |
| MVP | L + C | 66.4/70.5 | 61.2/67.7 | 58.8/63.9 | 60.6/66.4 |
| Ours | L + C | **67.3/70.6** | **63.4/70.1** | **62.4/69.3** | **61.8/70.0** |

We also conducted a comparative analysis of the detection performance differences between the proposed algorithm and point cloud baseline methods under the condition of large-scale missing (50% discarded) in the target point cloud region, as shown in Figure 7. In Scene 1, especially in the region of close-range small targets, the proposed method can still accurately perceive targets and generate stable bounding boxes, demonstrating better robustness than the point cloud baseline method. In Scene 2, even in the face of significant point cloud loss of long-range targets, the proposed method can still maintain effective prediction by relying on image semantic supplementation, while the point cloud baseline method completely fails in most cases. Overall, since the point cloud baseline method relies entirely on point cloud geometric information, obvious missed detections and bounding box drifts occur after key point loss, reflecting its high sensitivity to point cloud integrity. In contrast, the proposed method provides effective semantic compensation in target regions with incomplete structures by introducing a deep fusion mechanism of image features. At the same time, it uses a confidence-guided strategy to dynamically adjust the information fusion weight between modalities, thereby improving the overall stability and anti-interference ability.

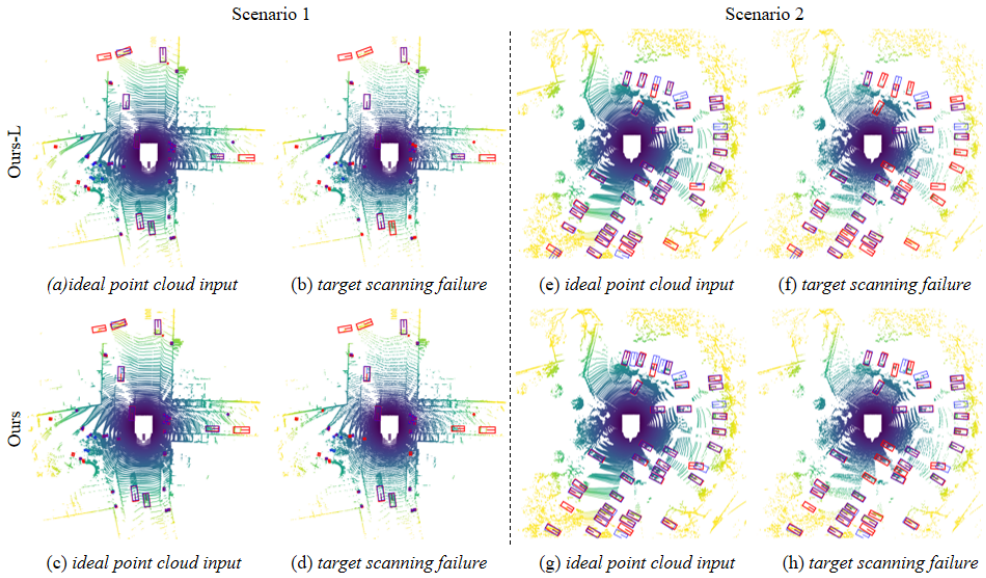



**Figure 7.** Visual comparison between ours and Ours-L under the target scanning failure scenario.

## 5. Conclusions

This paper proposes a 3D object detection algorithm for multimodal video streams based on reliability evaluation, with key achievements as follows:

1. Dynamically perceives the reliability of each modal feature and outputs confidence scores. It not only leverages the complementarity of point cloud and image BEV (Bird's-Eye View) features but also introduces pure visual BEV features to address extreme scenarios where point cloud quality degrades severely, ensuring the system's robustness under different environmental conditions.
2. Adaptively adjusts the contribution ratio of each feature in the fusion process based on the aforementioned confidence scores. This enables the fused features to more accurately reflect the target's geometric structure and semantic information, improving detection robustness and accuracy.
3. By explicitly modeling the correspondence between the internal semantic features of targets in consecutive frames, it achieves cross-time-scale target structure consistency constraints and collaborative optimization. This extends single-frame perception to

temporal memory and cross-frame structure inference, effectively enhancing the accuracy and stability of 3D target prediction to adapt to dynamic and complex scenarios.

In the experimental validation phase, the proposed algorithm was systematically evaluated on the nuScenes dataset and compared with existing mainstream 3D object detection algorithms. The results show that the proposed algorithm achieves an optimal performance of 67.3% and 70.6% in the two core metrics of mAP (mean Average Precision) and NDS (NuScenes Detection Score), respectively, which is significantly better than other existing algorithms. Ablation experiments further confirm that the confidence-weighted fusion module and cross-frame temporal consistency optimization module contribute significantly to the overall performance improvement of the model, and their synergistic effect ultimately achieves comprehensive performance gains in dynamic and complex scenarios.

In summary, this algorithm framework achieves innovative results in multiple aspects, effectively improving the stability and generalization ability of the overall detection system and providing a feasible solution for 3D object detection in autonomous driving. In the future, with the advancement of sensor technology and the further development of deep learning models, this algorithm framework is expected to play a role in a wider range of autonomous driving scenarios and promote the further development and popularization of intelligent driving technology.

**Author Contributions:** Conceptualization, M.J. and B.L.; methodology, G.L.; software, Y.D.; validation, X.C., M.J., and B.L.; formal analysis, G.N.; investigation, Y.D.; resources, G.L.; data curation, G.N.; writing—original draft preparation, X.C.; writing—review and editing, Y.D.; visualization, B.L.; supervision, M.J.; project administration, G.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by STATE GRID Corporation of China, the science and technology projects of research on Optimization Design and Intelligent Calibration Technology of Combined Transformer (5700-202326259A-1-1-ZN).

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

**Conflicts of Interest:** Author M.J., B.L., Y.D., X.C. and G.N. were employed by the China Electric Power Research Institute. Author G.L. was employed by State Grid Sichuan Electric Power Co., Ltd. The authors declare that this study received funding from STATE GRID Corporation of China. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## References

1. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
2. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
3. Yang, J.; Song, L.; Liu, S.; Mao, W.; Li, Z.; Li, X.; Sun, H.; Sun, J.; Zheng, N. Dbq-ssd: Dynamic ball query for efficient 3d object detection. *arXiv* **2022**, arXiv:2207.10909.
4. Wang, D.Z.; Posner, I. Voting for voting in online point cloud object detection. In *Robotics: Science and Systems*; Springer Proceedings in Advanced Robotics; Springer: Rome, Italy, 2015; Volume 1, pp. 10–15.
5. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. [CrossRef] [PubMed]
6. Du, L.; Ye, X.; Tan, X.; Feng, J.; Xu, Z.; Ding, E.; Wen, S. Associate-3ddet: Perceptual-to-conceptual association for 3d point cloud object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13329–13338.
7. 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; pp. 12697–12705.

8. Wang, Y.; Fathi, A.; Kundu, A.; Ross, D.A.; Pantofaru, C.; Funkhouser, T.; Solomon, J. Pillar-based object detection for autonomous driving. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part XXII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 18–34.
9. Zhang, J.; Liu, H.; Lu, J. A semi-supervised 3d object detection method for autonomous driving. *Displays* **2022**, *71*, 102117. [CrossRef]
10. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum pointnets for 3d object detection from rgb-d data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 918–927.
11. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4604–4612.
12. Yin, T.; Zhou, X.; Krähenbühl, P. Multimodal virtual point 3d detection. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16494–16507.
13. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
14. Liang, M.; Yang, B.; Wang, S.; Urtasun, R. Deep continuous fusion for multi-sensor 3d object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 641–656.
15. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
16. Pang, S.; Morris, D.; Radha, H. Clocs: Camera-lidar object candidates fusion for 3d object detection. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 10386–10393.
17. Pang, S.; Morris, D.; Radha, H. Fast-clocs: Fast camera-lidar object candidates fusion for 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 187–196.
18. Gong, R.; Fan, X.; Cai, D.; Lu, Y. Sec-clocs: Multimodal back-end fusion-based object detection algorithm in snowy scenes. *Sensors* **2024**, *24*, 7401. [CrossRef] [PubMed]
19. Koh, J.; Lee, J.; Lee, Y.; Kim, J.; Choi, J.W. Mgtanet: Encoding sequential lidar points using long short-term motion-guided temporal attention for 3d object detection. In Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1179–1187.
20. Yin, J.; Shen, J.; Gao, X.; Crandall, D.J.; Yang, R. Graph neural network and spatiotemporal transformer attention for 3d video object detection from point clouds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *45*, 9822–9835. [CrossRef] [PubMed]
21. Xu, J.; Miao, Z.; Zhang, D.; Pan, H.; Liu, K.; Hao, P.; Zhu, J.; Sun, Z.; Li, H.; Zhan, X. Int: Towards infinite-frames 3d detection with an efficient framework. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 193–209.
22. Chen, X.; Shi, S.; Zhu, B.; Cheung, K.C.; Xu, H.; Li, H. Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 680–697.
23. He, C.; Li, R.; Zhang, Y.; Li, S.; Zhang, L. Msf: Motion-guided sequential fusion for efficient 3d object detection from point cloud sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 5196–5205.
24. Hou, J.; Liu, Z.; Zou, Z.; Ye, X.; Bai, X. Query-based temporal fusion with explicit motion for 3d object detection. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 75782–75797.
25. Yuan, Y.; Sester, M. Streamlts: Query-based temporal-spatial lidar fusion for cooperative object detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2025; pp. 34–51
26. Van Geerenstein, M.R.; Ruppel, F.; Dietmayer, K.; Gavrila, D.M. Multimodal object query initialization for 3d object detection. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 12484–12491.
27. Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle adjustment—A modern synthesis. In *International Workshop on Vision Algorithms*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 298–372.
28. Lindenberger, P.; Sarlin, P.; Larsson, V.; Pollefeys, M. Pixel-perfect structure-from-motion with featuremetric refinement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 5987–5997.
29. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11621–11631.
30. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
31. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; Dai, J. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv* **2022**, arXiv:2203.17270.
32. Huang, J.; Huang, G. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv* **2022**, arXiv:2203.17054.

33. Wang, Y.; Guizilini, V.C.; Zhang, T.; Wang, Y.; Zhao, H.; Solomon, J. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In Proceedings of the Conference on Robot Learning, London, UK, 8–11 November 2021; PMLR: Auckland, New Zealand, 2022; pp. 180–191.
34. Yoo, J.H.; Kim, Y.; Kim, J.; Choi, J.W. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In Proceedings of the Computer Vision–ECCV 2020: 16th European conference, Glasgow, UK, 23–28 August 2020; part XXVII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 720–736.
35. Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; Yu, G. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv* **2019**, arXiv:1908.09492. [CrossRef]
36. Sheng, H.; Cai, S.; Liu, Y.; Deng, B.; Huang, J.; Hua, X.; Zhao, M. Improving 3d object detection with channel-wise transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 2743–2752.
37. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 11784–11793.
38. Chen, X.; Zhang, T.; Wang, Y.; Wang, Y.; Zhao, H. Futr3d: A unified sensor fusion framework for 3d detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 172–181.


**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.