Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081658

ARTICLE

CG-MAE: BEV Masked Autoencoders Based on Cross-Modal Guidance for 3D Object Detection in Autonomous Driving

Junchen Huo[1], **Song Wang**[1,*], **Enqing Chen**[1], **Yingqiang Ding**[1] **and Shouyi Yang**[2]

[1]School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, China
[2]Tianping College of Suzhou University of Science and Technology, Nanjing, China

*Corresponding Author: Song Wang. Email: ieswang@zzu.edu.cn

Received: 06 March 2026; Accepted: 18 May 2026; Published: 23 July 2026

ABSTRACT: Multi-modal 3D object detection, which leverages the complementary strengths of LiDAR point clouds and camera RGB images, has emerged as a critical component of 3D perception in autonomous driving. As a critical challenge in multi-modal learning, modality alignment aims to establish accurate semantic correspondences across distinct modalities. However, existing methods encounter significant difficulties in achieving robust alignment when data from one modality is obscured, such as in the presence of object occlusion or adverse environmental conditions, including illumination variations and inclement weather. To alleviate this issue, we present CG-MAE, a dual-branch Bird's-Eye-View (BEV) masked autoencoder framework based on cross-modal guidance for 3D object detection in autonomous driving. Specifically, a cross-modal guided reconstruction module is developed to predict the representations of the obscure objects in the BEV space, lowering the difficulty of the modality alignment during the multi-modal fusion process. To mimic the object obscuration caused by occlusion or adverse environments, this paper proposes a Ground truth-based foreground masking strategy to cover up the objects, such as vehicles and pedestrians, thereby encouraging the reconstruction module to focus on modality alignment in the foreground regions with high information density. Considering that all the modality data can be obscured, this paper builds a dual-branch BEV masked autoencoder to implement the reconstruction of the data from both the camera and LiDAR modalities. Extensive experiments on the nuScenes dataset with camera-LiDAR inputs demonstrate that the proposed framework achieves superior performance over existing state-of-the-art multi-modal learning methods.

KEYWORDS: 3D object detection; cross-modal guided reconstruction; ground truth-based foreground masking; multi-modality fusion; dual-branch BEV masked autoencoder

1 Introduction

3D object detection plays an indispensable role in 3D environmental perception, which is critical for autonomous driving [1–4]. Unlike 2D detection, it requires an accurate localization of objects in 3D space using onboard sensors such as cameras and LiDAR. In recent years, camera-based methods have emerged as a popular choice due to their low cost, ease of deployment and the ability to capture rich semantic information [5–8]. However, these methods lack explicit depth information and are highly sensitive to variations in illumination and weather. Compared to camera-based methods, LiDAR-based ones offer accurate geometric information and remain reliable under varying lighting conditions [9–11]. But they are usually limited by sparse point clouds and poor semantic information. Given their complementary characteristics, multi-modal fusion has become the mainstream approach in current research. By integrating

the distinct advantages of different modalities, multi-modal fusion methods achieve superior performance in 3D object detection [12–15].

Despite substantial progress in multi-modal 3D object detection, accurate alignment of heterogeneous modality features remains a core challenge. Direct fusion of misaligned features typically leads to a degradation in detection performance. To align the features from different modalities, current methods propose to map the features into a common space through geometric projection [16–19], shared BEV representations [14], and attention-based alignment [20–22]. In MVP [15] and MVX-Net [16], the image features are segmented and projected onto the LiDAR point cloud for alignment. BEVFusion [13] constructs a unified Bird's-Eye-View (BEV) space, where the features from heterogeneous modalities are aligned within a shared spatial domain. AutoAlign [20] and TransFusion [14] employ a dual Transformer decoder architecture. The first optimizes input-dependent, category-aware queries with spatial cross-attention constraints, while the second incorporates an image-guided query initialization module for soft LiDAR associations, thus mitigating feature misalignment effectively. While these strategies have demonstrated effectiveness under ideal conditions, they tend to degrade in complex real-world scenarios, such as vehicle occlusion, adverse weather and sensor malfunctions [9].

To align multi-modal features in complex scenarios, SPaRC [23] projects LiDAR features into the image space to align the features along the vertical dimension. MetaBEV [24] initializes a set of learnable BEV queries and then establishes associations with both the camera and LiDAR modalities through a BEV-Evolving decoder to handle sensor failures and missing modalities. While the above methods attempt to align multi-modal features directly, there are also some research works that choose to align them by recovering the obscure information of the raw data [25–27]. For example, UniM^{2}AE [28], employs random masking on raw PV-space data to enhance multi-modal representation. However, this approach overlooks the geometric discrepancy between PV and BEV spaces, distracting the network from the critical PV to BEV transformation and yielding limited 3D detection gains. Additionally, random masking introduces low-information background regions, resulting in inefficient training. In contrast, our framework employs GT-based foreground masking directly in the unified BEV space, effectively avoiding geometric misalignment. Building on BEV-based foundations, BEV-MAE [29] and vision BEV-MAE [30] apply masking directly to BEV features. However, these methods employ reconstruction to recover masked features, neglecting the complementary relationships between sensors. In contrast, our dual-branch BEV-masked autoencoder ensures bi-directional alignment through a cross-modal guided reconstruction module, fully exploiting the complementary features of multi-modal data.

In this paper, we propose CG-MAE, a dual-branch BEV masked autoencoder framework based on cross-modal guidance, for the modality alignment in the 3D object detection task. Specifically, a BEV masking strategy is designed to decouple the space transformation module and the reconstruction module, improving their concentration and interpretability. To raise the efficiency of model training, the proposed BEV masking strategy utilizes object Ground Truth (GT) as the reference to cover up the foreground objects and generate meaningful masked BEV data for the network. Based on the proposed masking strategy, a cross-model guided reconstruction module is proposed to predict the representations of the obscure objects in the BEV space, lowering the difficulty of the modality alignment during the multi-modal fusion process. Additionally, this paper builds a dual-branch BEV masked autoencoder to ensure the bi-directional alignment, where one modality is masked, and the information from the other modality is used to guide the reconstruction. To validate the effectiveness of the proposed framework, we conducted experiments on the large-scale nuScenes dataset [31] using camera-LiDAR inputs. Experimental results demonstrate that CG-MAE achieves superior performance over existing state-of-the-art multi-modal learning methods. Compared to the baseline method, CG-MAE significantly improves the 3D object detection accuracy.

It is worth noting that the proposed dual-branch BEV masked autoencoder framework combines BEV-space masking, cross-modal reconstruction and bidirectional alignment for robust multi-modal fusion. Its effectiveness stems from the collaboration of these components rather than any single module.

The main contributions of this paper are summarized as follows:

- A cross-modal guided reconstruction module is developed to align different modalities by predicting the representations of the obscured objects in the BEV space. To mimic the object obscuration caused by occlusion or adverse environment, this paper proposes a GT-based foreground masking strategy to cover up the objects like vehicles and pedestrians, thereby encouraging the reconstruction module to focus on the modality alignment in the informative foreground regions.
- Considering that all the modality data can be obscured, this paper builds a dual-branch BEV masked autoencoder to ensure the bi-directional alignment, where one modality is masked, and the information from the other modality is used to guide the reconstruction. Such a dual-branch design simultaneously reconstructs both masked modalities, avoiding the interference of the incomplete data from different modalities.
- Experimental results on the large-scale nuScenes dataset demonstrate that the proposed framework achieves superior performance over existing state-of-the-art multi-modal learning methods. Compared to the baseline method, CG-MAE significantly improves the 3D object detection accuracy.

2 Related Work

This section reviews related work about the multi-modal 3D object detection and multi-modal feature alignment.

2.1 Multi-Modal 3D Object Detection

In autonomous driving, the perception of the surrounding environment is crucial. Object detection aims to identify the classes and the positions of the important objects, becoming the indispensable technique in autonomous driving. Traditional object detection methods like YOLO [32] have achieved great success in various real-world scenarios due to their real-time performance and ease of deployment. Further, Yolo-DKR [33] is designed to enhance the efficiency and accuracy of the detection model through differentiable architecture search and kernel reuse. However, these methods mainly focus on 2D detection. With the rapid advancement of autonomous driving, 3D perception systems have gained significant attention for enabling vehicles to understand a complex environment.

Multi-modal 3D object detection exploits the complementary strengths of different sensors to precisely predict the position, dimension and class of surrounding obstacles. Earlier approaches employed early-level fusion strategies, focusing on the alignment of camera and LiDAR features at the feature level to achieve multi-modal integration. RangeLVDet [34] utilizes a pretrained 2D convolutional neural network to extract features from both images and LiDAR. These features are then fused at the point cloud level before being transformed into a BEV representation for downstream training tasks. SparseFuseDense [35] proposes a RoI fusion strategy called 3D-GAF, which utilizes image features via depth completion to generate pseudo-point clouds and performs grid-based fusion of 3D RoI features within the point cloud pairs. RI-Fusion [36] is designed to align image and point cloud data by converting point clouds into a compact range representation, then utilizing an attention mechanism to fuse these two modalities. To better preserve both geometric structure and semantics, recent work has shifted toward mid-level fusion within a unified 3D or BEV space. MLOD [37] generates 3D proposals within the BEV projection of LiDAR point clouds and subsequently performs multi-modal fusion by combining image features with these localized BEV representations. UA-Fusion [12] employs a probabilistic cross-modal attention mechanism to achieve

adaptive and complementary fusion of multi-modal data. BEVFusion [13] performs fusion within a unified BEV space, effectively preserving both geometric and semantic information. Similarly, TransFusion [14] utilizes an attention mechanism to establish soft associations between images and LiDAR points for fusion. Furthermore, MapFusion [38] leverages the complementary information of multi-modal BEV features to enhance the overall performance of downstream tasks. UniTR [39] adopts a unified modeling approach with shared parameters to map cross-modal data into a unified representation for effective fusion. Although the methods mentioned above achieve great progresses on multi-modal fusion, they face a significant challenge that the inherent differences in heterogeneous data generate vulnerable feature alignment between different modalities, leading to a performance degradation in real-world applications.

2.2 Multi-Modal Feature Alignment

In multi-modal learning, feature alignment across modalities aims to establish semantic and geometric correspondences among cross-modal features, facilitating the fusion of complementary information from different sensors.

Several approaches employ projection techniques to achieve feature alignment across different modalities. In PI-RCNN [17] and Seg-VoxelNet [18], images are passed through a segmentation network to generate pixel-wise semantic predictions. These semantic labels are then appended to the 3D point clouds based on point-to-pixel correspondence. Building upon this, MVP [15] utilizes image segmentation techniques to align 2D segmentation masks with 3D point cloud features via projection matrices. Similarly, MVX-Net [16] voxelizes the raw point cloud and projects image features from corresponding pixels onto these voxels to achieve feature alignment between the two modalities. Both ContFuse [19] and BEVFusion [13] adopt a unified BEV representation to align features from heterogeneous modalities. ContFuse [19] retrieves the K-nearest neighbors for each query point in the BEV plane, projects them into 3D space, and subsequently re-projects them onto the image plane. The sampled image features are then concatenated with continuous 3D offsets and processed by an MLP to align object features within the BEV space. Meanwhile, BEVFusion [13] transforms both camera and LiDAR features into a shared BEV space and utilizes a fully convolutional module to mitigate spatial misalignment between the different modalities. Alternatively, some approaches utilize attention mechanisms for alignment by mapping voxel or point features to queries (Q) and the complementary modal features to keys (K) and values (V) for attention computation. This strategy is exemplified by AutoAlign [20] and AutoAlignV2 [21], where a deformable cross-attention mechanism is employed to allow voxels to perceive the entire image space for precise multi-modal alignment. Furthermore, TransFusion [14] utilizes dual Transformer decoders to learn soft associations between camera and LiDAR representations, facilitating feature alignment without strict geometric projections. While existing approaches have shown promise in feature alignment, they struggle to maintain robustness in challenging real-world scenarios. Factors such as vehicle occlusion, sensor failure and severe lighting variations often lead to feature degradation, severely compromising the alignment accuracy.

For complex scenarios, such as vehicle targets under occlusion, sensor failures and adverse weather or illumination conditions, FAOD [27] is proposed to integrate cross-modal attention with occlusion handling and feature completion. To address the sensor reliability problem, MetaBEV [24] employs cross-attention mechanisms to resolve modal alignment challenges during sensor signal loss. Furthermore, SPaRC [23] leverages adaptive LiDAR-based spatial alignment and attention mechanisms to enhance robustness under adverse weather and low-visibility conditions. Masked reconstruction approaches have emerged as a viable solution for alignment challenges in occluded environments. Vision BEV-MAE [30] performs masking in the PV space to simulate occlusion conditions and then reconstructs high-quality BEV features, while BEV-MAE [29] performs masking on point clouds in the BEV space to encourage the model to learn

critical information through reconstruction. However, both are single-modal masking and reconstruction approaches, which do not fully exploit the complementary advantages of multi-modal features. In M-BEV [8], a unified masked BEV perception framework is proposed to enhance the perception robustness in complex scenarios by performing random masking and reconstruction on camera views. UniM^{2}AE [28] masks inputs from both camera and LiDAR modalities and projects them into a coherent 3D volumetric space for BEV transformation. In this paper, CG-MAE employs cross-modal reconstruction strategy that utilizes complete modal features to guide the reconstruction of masked ones, enabling the model to fully learn the complementary advantages among multi-modal features. To address this issue, the dual-branch BEV masked autoencoder is trained to model the alignment relationship between the features from different modalities and predict the representations of the obscure objects under the cross-modal guidance.

3 Methodology

3.1 Overall Framework

The proposed CG-MAE framework is illustrated in Fig. 1. Build on BEVFusion [13], this paper inserts the proposed dual-branch BEV masked autoencoder between the BEV transformation modules and the BEV fusion module. The GT-based foreground masking strategy is designed to synthesize masked BEV feature maps for autoencoder training. The corresponding cross-modal reconstruction loss supervises the dual-branch reconstruction and encourages the autoencoder to recover the complete BEV feature maps from the masked ones. After obtaining the BEV feature maps from the modality-specific BEV transformation modules, the masking strategy randomly covers up the foreground regions of the obtained BEV feature maps. Given the masked feature maps, the dual-branch BEV masked autoencoder aims to recover the missing information under the supervision of the cross-modal reconstruction loss. In real-world scenarios, the vehicle occlusion, adverse weather and sensor malfunctions are likely to cause key object obscuration for one of the modalities, thereby degrading the feature alignment during the multi-modal feature fusion process. To address this issue, the dual-branch BEV masked autoencoder is trained to model the alignment relationship between the features from different modalities and predict the representations of the obscure objects under the cross-modal guidance.

3.2 GT-Based Foreground Masking Strategy

Inspired by Masked Autoencoders (MAE) [40] current research work typically masks the inputs and reconstructs them to enhance the feature representation ability of networks. In the context of multi-modal 3D object detection, there are image-based [41], LiDAR-based [25] and BEV-based [29,30] masking strategies. While the image-based and LiDAR-based strategies mainly mask the input images and point clouds separately, as shown in Fig. 2a,b, the BEV-based strategy focuses on randomly masking the BEV feature maps (Fig. 2c). In the PV space, masking destroys the geometric information of features. More importantly, the PV-to-BEV transformation heavily relies on dense local textures, and applying masking before projection disrupts these critical cues, leading to irreversible geometric distortion and additional transformation errors. In contrast, masking in BEV space effectively avoids these issues. More importantly, performing reconstruction in PV space diverts the alignment module's attention from modeling spatial transformations, resulting in degraded BEV features. By performing masking in BEV space, the spatial transformation and reconstruction modules are effectively decoupled, improving their concentration and interpretability. Since the masking strategy in the raw data tends to distract the alignment network's attention from the space transformation, this paper chooses the BEV masking strategy to decouple the space transformation module and the reconstruction module, improving their concentration and interpretability. Considering the random masking is likely to introduce heavy background noise into the training data, this

paper utilizes object GT as the reference to cover up the foreground objects and generate meaningful masked BEV data for model training, as shown in Fig. 2d.

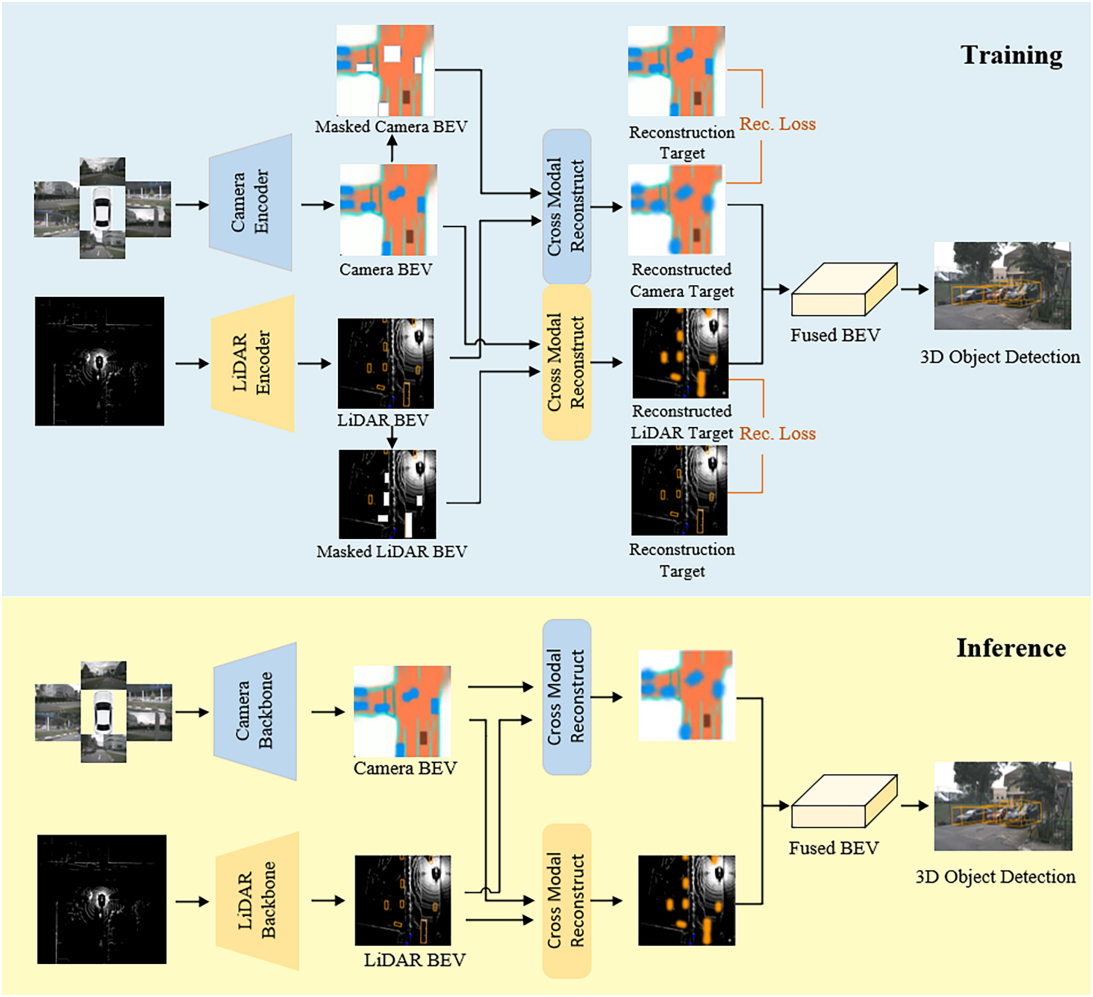


Figure 1: The GC-MAE framework architecture, illustrating training and inference phases. In training, modality-specific encoders transform multi-view images and LiDAR point clouds into the BEV space, followed by GT-based foreground masking on feature maps. A cross-modal guided reconstruction module leverages intact features from one modality to recover masked regions of the other, facilitating deep feature alignment for 3D detection. During inference, the masking operation is deactivated; complete multi-modal BEV features are directly fed into the reconstruction module for deep feature alignment, ensuring robust fusion for the final 3D detection output.

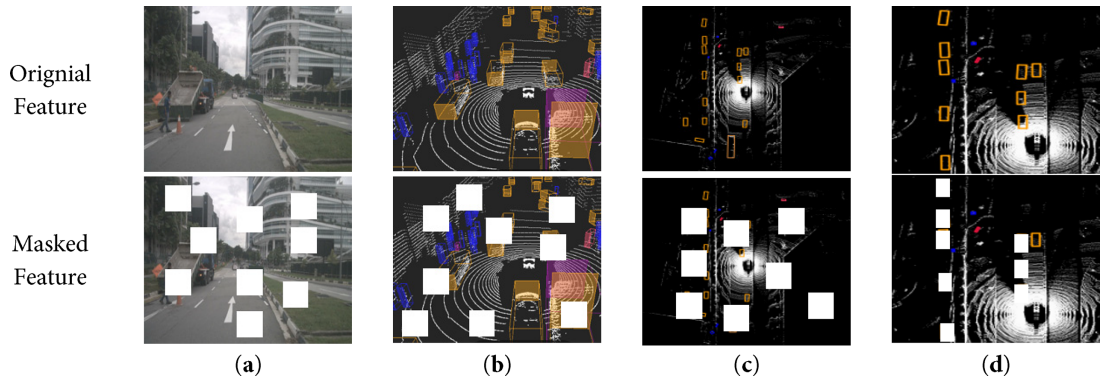


Figure 2: Visualization of different masking strategies. The top row displays the original features, while the bottom row presents the corresponding masked features. Specifically, (**a**) and (**b**) illustrate random block masking applied to the camera image and LiDAR point cloud, respectively. (**c**) depicts random masking within the BEV feature space. (**d**) shows the proposed GT-based foreground masking strategy.

Given the BEV feature maps, we extract the set of GT 3D bounding boxes from the dataset, which is denoted as Eq. (1).

$$\mathbf{b}_i = (\mathbf{c}_i, \mathbf{s}_i, \theta_i)_{i=1,\ldots,N}, \tag{1}$$

where $\mathbf{c}_i = (x_i, y_i, z_i)^\top \in \mathbb{R}^3$ denotes the center coordinates of the i-th object in the ego vehicle coordinate system. The element $\mathbf{s}_i \in \mathbb{R}^3$ represents its dimensions (length, width, height). The yaw angle $\theta_i \in (-\pi, \pi)$ is a movement around the z-axis. N stands for the number of the objects around the ego. All geometric parameters are measured in meters (m) and radians (rad). To construct a binary mask on a BEV feature map of size $H \times W$, each 3D bounding box is first projected onto the BEV plane. We adopt an axis-aligned bounding box strategy for this projection, effectively ignoring the yaw angle θ_i to obtain its circumscribed rectangle on the horizontal plane. Specifically, the projected BEV region of the i-th object in the physical coordinate system is defined by its extents as Eq. (2).

$$\begin{aligned} x_i^{\min} &= x_i - \frac{w_i}{2}, x_i^{\max} = x_i + \frac{w_i}{2}, \\ y_i^{\min} &= y_i - \frac{l_i}{2}, y_i^{\max} = y_i + \frac{l_i}{2}. \end{aligned} \tag{2}$$

These local physical boundaries (measured in meters) need to be mapped to the coordinate system of the BEV feature map. We define $[x_{\min}, x_{\max}]$ and $[y_{\min}, y_{\max}]$ denoting the predefined global perception ranges of the BEV space along the x and y axis, respectively. Spatial normalization is applied followed by a scaling operation to transform the physical boundaries into continuous coordinates. The transformation is formulated as Eq. (3).

$$\begin{aligned} u_i^{\min} &= \frac{x_i^{\min} - x_{\min}}{x_{\max} - x_{\min}} \cdot W, u_i^{\max} = \frac{x_i^{\max} - x_{\min}}{x_{\max} - x_{\min}} \cdot W, \\ v_i^{\min} &= \frac{y_i^{\min} - y_{\min}}{y_{\max} - y_{\min}} \cdot H, v_i^{\max} = \frac{y_i^{\max} - y_{\min}}{y_{\max} - y_{\min}} \cdot H, \end{aligned} \tag{3}$$

where u and v represent the projected continuous coordinates along the width (W, corresponding to the grid columns) and height (H, corresponding to the grid rows) of the feature map. This linear transformation

essentially calculates the relative position ratio of the object within the entire physical perception range and then scales it to the $H \times W$ grid dimensions. Finally, the foreground binary mask is generated based on the indices as Eq. (4).

$$\mathbf{M}[v,u] = \begin{cases} 1, & \text{if} \exists i \in \{1,\ldots,N\} \text{ such that } u_i^{\min} \le u < u_i^{\max} \text{ and } v_i^{\min} \le v < v_i^{\max}, \\ 0, & otherwise. \end{cases} \tag{4}$$

The proposed masking strategy accurately identifies object regions on the BEV plane and effectively masks critical foreground areas. Although using ground-truth annotations during training provides hard inputs to the reconstruction network for training, GT-based masking forces the network to recover the erased objects by relying only on information from the complementary modality rather than simple local interpolation. By learning to solve this challenging reconstruction task, the network acquires highly generalized cross-modal feature representations. These learned representations directly enable the model to handle unseen scenarios and severe occlusions effectively during inference, enhancing the generalization ability of our method.

3.3 Dual-Branch Cross-Modal Guided BEV Reconstruction

Existing MAE-based BEV reconstruction methods typically focus on single modality, neglecting the inherent alignment between the modalities. This paper employs the window-based cross-attention mechanism to implement the cross-modal guided BEV reconstruction. The mechanism confines the cross-modal interaction within local spatial windows. Within each local window, the masked Query tokens dynamically aggregate the unmasked Key/Value tokens from the complementary modality. Supervised by the reconstruction loss, the cross-attention acts as a precise semantic translator. It forces the network to discover and align the corresponding semantic representations within the exact same physical neighborhood, effectively preserving semantic consistency without distortion. The geometric consistency is fundamentally preserved because the cross-attention operates within a shared BEV space. Before the attention computation, we inject 2D sinusoidal positional embeddings into both the masked Query tokens and the unmasked Key/Value tokens. This forces the attention weights to strictly respect physical spatial constraints, ensuring the geometric alignment. The proposed dual-branch reconstruction module performs masked reconstruction for both LiDAR and camera BEV features, as shown in Fig. 3. We take the camera branch as an example to describe the proposed method in the following.

Given the camera BEV feature map $B_C \in \mathbb{R}^{H\times W\times C_C}$ and the LiDAR BEV feature map $B_L \in \mathbb{R}^{H\times W\times C_L}$, we first apply the GT-based foreground masking strategy to mask B_C to mitigate the potential occlusion for the camera modality inputs. The masked feature maps, represented as $MASK(B_C)$, are then treated as a sequence of patches and transformed into token representations (Patch.Emb). To preserve the essential spatial structure, 2D positional encodings E_{pos}^C are subsequently added to the token sequence. This entire process is formulated as Eq. (5).

$$B_{C_t}^M = \text{Patch.Emb}(\text{MASK}(B_C)) + E_{pos}^C, \tag{5}$$

where $B_{C_t}^M$ denotes the camera BEV tokens after the masking and embedding operations. E_{pos}^C denotes the positional encoding operation. By adding positional embeddings to these tokens, precise spatial information is provided, enabling the masked regions to be distinguished based on their coordinates.

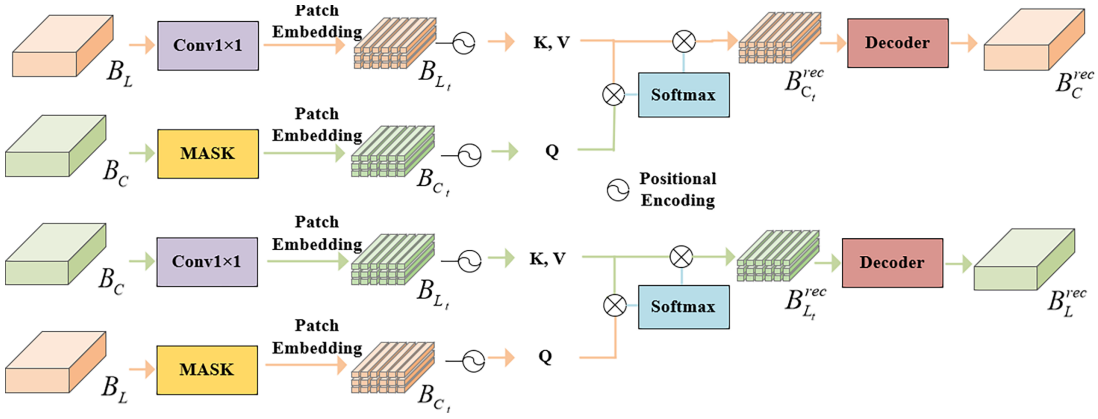


Figure 3: The detailed architecture of the cross-modal guided reconstruction module. This module facilitates bidirectional feature recovery between LiDAR and Camera modalities in the BEV space. Specifically, to reconstruct masked modality, the complementary modality undergoes 1×1 convolution and patch embedding to serve as the Key (K) and Value (V), while the masked target modality provides the Query (Q). Through a cross-attention mechanism and subsequent decoding, the module generates reconstructed feature B_C^{rec} and B_L^{rec}, achieving deep inter-modal alignment and maintaining cross-modal feature consistency.

The reconstruction of the masked camera feature map is guided by the complete LiDAR feature map B_L.A1 $\times$ 1 convolution layer is first applied to change the channel dimension of B_L from C_L to C_C, achieving cross-modal channel alignment and ensuring compatibility with the subsequent attention mechanism. Then, we tokenize the LiDAR feature map and add the 2D sinusoidal positional embeddings as Eq. (6).

$$B_{L_t} = \text{Patch.Emb}\left(\text{Conv}_{1\times1}(B_L)\right) + E^L_{\text{pos}}, \tag{6}$$

where B_{L_t} denotes the tokens of the LiDAR BEV feature map which is used for guided reconstruction.

Given a BEV feature map of spatial size $H \times W$, applying a standard Global Cross-Attention mechanism incurs a quadratic computational complexity with respect to the spatial resolution, denoted as $O\left((\text{HW})^2\right)$. Due to the inherently large spatial dimensions of BEV features (e.g., 256×704), such a global formulation results in prohibitive memory consumption, posing a severe risk of Out-of-Memory (OOM) errors. To circumvent this computational bottleneck and ensure both the stability of the overall model and its feasibility for practical deployment, we adopt a window-based cross-attention mechanism. This approach partitions the high-resolution BEV space into non-overlapping local windows and strictly restricts the attention computation within each window. Consequently, the computational complexity is significantly reduced to $O\left(\text{HWM}^2\right)$, ensuring the efficiency and stability of our dual-branch cross-modal reconstruction architecture. Specifically, we utilize the masked camera tokens as queries (Q), and the complete LiDAR tokens as keys (K) and values (V). The cross attention is computed as Eq. (7).

$$B^{rec}_{C_t} \& = CrossAttn(\boldsymbol{Q} = B^M_{C_t}, \boldsymbol{K} = B_{L_t}, \boldsymbol{V} = B_{L_t}), \tag{7}$$

where $B^{rec}_{C_t}$ stands for the tokens for the reconstructed camera BEV feature map. The operator $CrossAttn$ is the window-based attention as Eq. (8).

$$CrossAttn(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{8}$$

Here, Q, K, and V denote the query, key, and value matrices, respectively. The parameter d_k represents the dimension of the key vectors, serving as a scaling factor to regulate the output distribution of the Softmax function and prevent the gradients from vanishing during training. The window-based attention mechanism is employed in this paper for lowering the computational cost.

Subsequently, these updated tokens $B_{C_t}^{rec}$ are fed into a decoder to derive the ultimate reconstructed camera features B_C^{rec}. Specifically, we adopt a simple 3×3 convolutional layer as the decoder, which effectively aggregates local spatial contexts and refines the reconstructed feature representations.

Similarly, the reconstruction process for the masked LiDAR feature map is formulated as Eq. (9).

$$B_{L_t}^{rec} \& = CrossAttn(\boldsymbol{Q} = B_{L_t}^{M}, \boldsymbol{K} = B_{C_t}, \boldsymbol{V} = B_{C_t}). \tag{9}$$

The ultimate reconstructed LiDAR feature map B_L^{rec} is obtained by feeding $B_{L_t}^{rec}$ into the reconstruction decoder with the same architecture of the one for the camera feature map.

3.4 Loss Function

To supervise the dual-branch reconstruction and facilitate robust cross-modal alignment, we formulate the cross-modal reconstruction loss as Eq. (10).

$$\mathcal{L}_{cross-recon} = \mathcal{L}_{recon}(B_C^{rec}, B_C) + \mathcal{L}_{recon}(B_L^{rec}, B_L), \tag{10}$$

where $\mathcal{L}_{recon}$ is defined as a normalized Mean Squared Error (MSE) which is represented as Eq. (11).

$$\mathcal{L}_{recon}(F, \tilde{F}) = \frac{1}{\sum_{i,j} M_{i,j}} \sum_{i,j} \left(M_{i,j} \cdot \frac{F_{i,j} - \tilde{F}_{i,j}}{\sigma} \right)^2, \tag{11}$$

where $M \in \{0,1\}$ is a binary mask which ensures that the loss is computed exclusively within the masked regions, preventing the network from trivially copying the unmasked background. $F_{i,j}$ and $\tilde{F}_{i,j}$ denote the reconstructed and original target features at spatial position (i, j), respectively. The normalization by the target features' standard deviation σstabilizes training convergence across modalities with varying distributions.

By integrating the cross-modal reconstruction loss into the overall 3D detection training objective, the total loss $\mathcal{L}_{\text{total}}$ is formulated as Eq. (12).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det 3d}} + \alpha \mathcal{L}_{cross-recon}, \tag{12}$$

where α stands for the weight to balance the 3D detection loss $\mathcal{L}_{\text{det 3d}}$ [14] and the cross-modal reconstruction loss $\mathcal{L}_{cross-recon}$. Empirically, we set $\alpha = 0.1$ in experiments.

4 Experiments

4.1 Implementation Details

We evaluate the proposed CG-MAE framework on the nuScenes dataset [31]. The nuScenes [31] dataset contains 28,130 training samples and 6019 validation samples, each providing point clouds collected with a 32-beam LiDAR and images captured by 6-view cameras. In this paper, all models are trained on the training set and evaluated on the validation set. The nuScenes [31] dataset evaluates the performance of 3D object detection algorithms using two primary metrics, i.e., mean Average Precision (mAP) and nuScenes Detection Score (NDS) [31]. mAP is calculated as the arithmetic mean of the Average Precision (AP) across

all categories. For the NDS, five True Positive (TP) metrics are first computed, including Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). NDS is the sum of the weighted sum of the five TP metrics.

This paper adopts BEVFusion [13] as the baseline which means that the training configurations are set to same with BEVFusion [13]. The proposed framework is trained on 4 GeForce RTX 4090D GPUs with $batch\_size = 12$.

4.2 Experimental Results

4.2.1 Quantitative Analysis

We compare the proposed CG-MAE with the State-Of-The-Art (SOTA) methods in Table 1 to demonstrate its effectiveness. "C" and "L" represent the modality of the camera and LiDAR separately. The combination "C+L" means that the models are trained by inputting both camera and lidar data. The compared approaches include camera-only ("C") methods: BEVStereo [5], Vision BEV-MAE [30], BEVFormer [7], M^{2}BEV [42], M-BEV [8], and KAN-RCBEVDepth [43]; LiDAR-only ("L") methods: UVTR [44], PillarNext [45], and UniBEV-L [46]; and multi-modal ("C+L") methods: MVP [15], UniBEV [46], GraphAlign [47], PointPainting [48], FUTR3D [49], UniM2AE [28], and BEVFusion* [13]. To ensure a fair comparison, we retrain the baseline BEVFusion [13] under the same experimental settings with CG-MAE. We use the symbol "*" to denote that BEV fusion is retrained in this paper. For other state-of-the-art methods, we adopt the results reported in their original papers. From Table 1, CG-MAE achieves superior performance in multi-modal 3D object detection. Compared with existing methods, CG-MAE attains 67.1% mAP and 70.4% NDS. When compared with the baseline BEVFusion [13], CG-MAE yields improvements of 1.2% mAP and 0.6% NDS, demonstrating its effectiveness for 3D object detection.

Table 1: Performance comparison of different methods for 3D object detection on the nuScenes validation set. mAP and NDS are used as the primary evaluation metrics. C and L denote the camera and LiDAR modalities, respectively. L+C indicates that the model is trained with multi-modal inputs from both camera and LiDAR.

Method	Modality	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVStereo [5]	C	44.3	34.4	65.8	28.2	58.6	52.9	23.3
Vision BEV-MAE [30]	C	49.0	39.1	71.7	27.3	42.3	45.5	19.2
BEVFormer [7]	C	47.9	37.0	72.1	27.9	40.7	43.6	22.0
M^{2}BEV [42]	C	47.0	41.7	64.7	27.5	37.7	83.4	24.5
M-BEV [8]	C	45.8	34.6	–	–	–	–	–
KAN-RCBEVDepth [43]	C	41.5	31.6	70.1	–	63.1	–	–
PillarNext [49]	L	68.4	62.2	–	–	–	–	–
UVTR [44]	L	69.7	63.9	30.2	24.6	35.0	20.7	12.3
UniBEV-L [47]	L	58.9	49.7	–	–	–	–	–
MVP [15]	C+L	65.2	57.8	–	–	–	–	–
GraphAlign [45]	C+L	68.5	62.8	–	–	–	–	–
PointPainting [46]	C+L	68.5	64.2	–	–	–	–	–
UniBEV [47]	C+L	70.1	66.5	26.3	23.8	32.1	31.3	13.4
FUTR3D [48]	C+L	68.3	64.5	28.1	24.7	36.8	25.8	12.4
UniM2AE [28]	C+L	68.1	64.3	–	–	–	–	–
BEVFusion*	C+L	69.8	65.9	28.4	25.2	29.8	25.8	17.8
CG-MAE* (Ours)	C+L	70.4	67.1	27.9	25.3	29.1	25.1	17.2

Note: ↑ denotes that higher values are better, ↓ denotes that lower values are better, *denotes that we retrained the model under the same experimental settings with CG-MAE for a fair comparison.

To validate the model's robustness to sensor degradation and severe occlusions, this section simulates the scenarios of data corruptions during the evaluation phase, on the standard nuScenes dataset. In experiments, we design three specific data corruption scenarios, including view drop, beam reduction and obstacle occlusion. The view drop is simulated by zeroing out the pixel values of random three camera views. The beam reduction is a typical scenario of LiDAR sensor degradation. This paper reduces the original 32 beams of the LiDAR data to 16 beams to achieve the beam reduction. As for the obstacle occlusion, fixed-size zero-value masks are designed to randomly cover up the objects of the input images.

The evaluation results are shown in Table 2. It is observed that CG-MAE significantly outperforms the baseline BEVFusion, demonstrating the model's robustness to sensor degradation and severe occlusions. Specifically, CG-MAE achieves mAPs of 61.3% and 57.4% under the scenarios of camera failure and beam reduction separately. In the obstacle occlusion scenario, CG-MAE exceeds the baseline by 1.9% in mAP. These evaluation results demonstrate that the proposed cross-modal guided reconstruction method effectively aligns and recovers the representation of the obscured objects.

Table 2: The evaluation results when comparing to the baseline BEVFusion on the scenarios of view drop, beam reduction and obstacle occlusion.

Method	View Drop 3 Drops		Beam Reduction 16 Beams		Obstacle Occlusion w Occlusion	
	NDS↑	mAP↑	NDS↑	mAP↑	NDS↑	mAP↑
BEVFusion	63.1	60.5	59.4	56.9	65.8	62.6
CG-MAE	64.2	61.3	60.2	57.4	67.1	64.5

Note: ↑ denotes that higher values are better.

4.2.2 Qualitative Results

Fig. 4 illustrates the 3D detection results of the baseline BEVFusion [13] and the proposed framework on the images from the nuScenes dataset [33]. It is worth noting that the GT 3D detection boxes are also provided in the first column of Fig. 4 as the reference. The cyan dashed circles in Fig. 4 highlight the differences between the different models in challenging scenarios. Specifically, the first row demonstrates the capability to handle severe occlusions (e.g., vehicles heavily obscured by tree trunks). The second and third rows emphasize the effective identification of distant objects under slight occlusion. This demonstrates the key advantage of our cross-modal reconstruction strategy. For severely occluded objects, our network treats them as masked regions and uses complementary modalities to recover missing spatial geometry. The fourth row highlights the successful detection of small-scale and occluded objects situated at the edges of the field of view, proving that our cross-modal guidance effectively enhances weak foreground features. Unlike BEVFusion [13], which passively fuses degraded features and frequently misses these challenging targets, CG-MAE actively reconstructs the missing representations. These results confirm that our dual-branch reconstruction framework effectively refines challenging foreground features, ensuring superior robustness in complex autonomous driving environments.

4.3 Ablation Study

We conducted a series of ablation experiments to evaluate the efficacy of the proposed components within the CG-MAE framework. First, we evaluate the contribution of each module design and analyze the weights of different loss functions. We further investigate model complexity and experimentally study the decoder layers and masking ratio.

(**a**) Ground Truth (**b**) BEVFusion (**c**) CG-MAE

Figure 4: Qualitative results. (**a**) ground truth, (**b**) detection results of BEVFusion, (**c**) detection results of CG-MAE.

4.3.1 Ablation Study on Key Components

To evaluate the contribution of each component proposed in this paper, we conducted an ablation study on the nuScenes [31] validation set. Specifically, we replace the GT-based masking strategy with a random masking strategy to verify the effectiveness of the GT-guided mask design. To validate the superiority of cross-modal feature alignment, we replace the cross-modal guidance with a single-modal reconstruction network. For the dual-branch component, we compare the proposed full model with two single-branch variants, i.e., single-branch camera and single-branch LiDAR. Single-branch camera denotes that masked reconstruction is applied only on camera BEV features, where the LiDAR BEV feature map is directly used in the fusion module. Similarly, Single-branch LiDAR stands for that only LiDAR BEV feature map is masked and reconstructed. The ablation results are shown in Table 3.

In Table 3, the results demonstrate that CG-MAE achieves the best performance with 67.1% mAP and 70.4% NDS, marking a significant improvement of +1.2% mAP and +0.6% NDS over the baseline. It is observed that using the GT-based mask strategy in row 6 yields improvements of 0.8% mAP and 0.4% NDS compared with the random mask strategy in row 2. Such a performance gain is attributed to the inherent differences in the objectives. Random masking treats the BEV space uniformly, inevitably masking the areas with low information intensity in the background. In this case, the training of the

detection model suffers from the recovery for the meaningless areas. In contrast, the proposed GT-based masking strategy focuses on the foreground objects. This accurate object-level masking not only mitigates background interference but also forces the network to infer complete structural and semantic information of the detected objects using cross-modal guidance, thereby learning more discriminative and object-aware representations for the 3D detection task. The cross-modal guided reconstruction in row 6 outperforms the single-modal reconstruction in row 3 by 0.9% mAP and 0.4% NDS, demonstrating that the cross-modal module effectively bridges the modality gap between camera and LiDAR data. It enables the model to fully exploit complementary information (rich texture and semantics from camera and robust spatial and structural cues from LiDAR) to learn deep cross-modal geometric and semantic alignment representations. Using information from the complementary modality, the proposed approach reconstructs the missing geometric structures to compensate for incomplete inputs, leading to significant improvements in detection accuracy. As shown in Table 3, the dual-branch design in row 6 achieves improvements of 0.6% mAP and 0.3% NDS over the single-branch camera in row 4 and 0.4% mAP and 0.2% NDS over the single-branch LiDAR in row 5. This is primarily because the performance of single-branch reconstruction is limited by the absolute quality of the guiding modality, and any inherent degradation in the guiding features will cause lower overall reconstruction performance. When adopting the dual-branch architecture to perform bidirectional reconstruction simultaneously, the model achieves the best performance.

Table 3: The ablation study on key components.

GT-Based Masking	Cross-Modal Guid.	Dual-Branch	mAP↑	NDS↑
—	—	—	65.9	69.8
—	√	√	66.3	70.0
√	—	√	66.2	70.0
√	√	Single-Bra. C	66.5	70.1
√	√	Single-Bra. L	66.7	70.2
√	√	√	67.1	70.4

Note: ↑ denotes that higher values are better.

4.3.2 The Effect of Loss Weight

Table 4 investigates the effect of the loss weight parameter α on the final detection performance. In this paper, α is responsible for balancing the auxiliary masked feature reconstruction loss with the primary 3D object detection loss. The experimental results in Table 4 demonstrate a clear inverted-U performance curve, with the model achieving its optimal state at $\alpha = 0.1$, peaking at 70.4 in NDS and 67.1 in mAP while simultaneously minimizing all True Positive error metrics. When the weight is too small, e.g., $\alpha = 0.01$, the network places insufficient emphasis on the cross-modal reconstruction task, failing to fully extract rich contextual semantics. Conversely, when the weight is excessively large ($\alpha = 0.5$), the performance degrades significantly, with NDS dropping to 68.9% and mAAE surging to 20.6%. This is because an overwhelmingly large auxiliary loss dominates the gradient updates, severely interfering with the optimization of the core 3D detection task. Therefore, we set $\alpha = 0.1$ as the default configuration.

Table 4: The effect of the different α.

Loss Weight α	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
0.01	69.1	65.5	28.7	26.3	30.2	26.1	19.7
0.05	70.2	66.4	28.3	25.9	29.6	25.7	17.6
0.1	70.4	67.1	27.9	25.3	29.1	25.1	17.2
0.5	68.9	65.2	29.1	26.5	30.2	27.5	20.6

Note: ↑ denotes that higher values are better, ↓ denotes that lower values are better.

4.3.3 Model Complexity

During the inference phase, all evaluations are performed on a single NVIDIA RTX 4090 D (24 GB) GPU with a batch size of 1. The input images are resized to 256 × 704. Same LiDAR voxelization parameters and the data pre-processing pipeline are utilized to ensure a fair comparison. In latency measurements, we excluded the initial 50 frames for GPU warm-up and record the true forward-pass time. We compare the proposed CG-MAE with the baseline BEVFusion and the related SOTA method UniM^{2}AE. Table 5 presents the comparison of model complexity and detection performance. It is observed from Table 5 that CG-MAE achieves a superior balance between accuracy gains and computational overhead. Although UniM^{2}AE illustrates lower overall computational complexity and a lighter architecture, its performance declines significantly. Specifically, while the proposed CG-MAE increases the training time and introduces a marginal increase in inference latency, it remains highly efficient with a processing speed of 4.9 FPS. Notably, the increase in resource consumption yields significant performance improvements. These results demonstrate that CG-MAE effectively boosts multi-modal 3D detection performance without imposing a prohibitive computational burden, making it suitable for autonomous driving applications.

Table 5: The comparison with BEVFusion and UniM^{2}AE.

Method	Training Time↓	Inference Time↓	Param. ↓	FPS↑	GFLOPS↓	mAP↑	NDS↑
BEVFusion	55 h	186.5 ms	44.5M	5.4	800.9G	65.9	69.8
UniM^{2}AE	56 h	179.7 ms	42.8M	5.6	783.1G	64.3	68.1
CG-MAE	62 h	202.7 ms	46.7M	4.9	884.6G	67.1	70.4

Note: ↑ denotes that higher values are better, ↓ denotes that lower values are better.

4.3.4 Decoder Design

To identify the optimal architecture for BEV feature restoration, we investigate the impact of various decoder designs on 3D detection accuracy. As shown in Table 6, we compare three representative architectures: a lightweight Linear layer, a Transformer block, and a 3 × 3 Convolutional decoder. The Convolutional decoder outperforms the others, achieving the best performance with 70.4% NDS and 67.1% mAP. While the Linear layer provides a minimal-overhead baseline, it fails to capture the intricate spatial structures necessary for high-quality restoration. Although the Transformer block performs excellently in global modeling, its performance is inferior to that of the convolutional decoder. This is because the masked BEV features are relatively sparse. While the Transformer calculates the interactions of the entire feature map, it introduces noise from the background region in the sparse BEV space. In contrast, the 3 × 3 convolutional decoder focuses on effective local context, avoiding interference from background noise. This refined local feature optimization is crucial for accurately locating the occluded objects. Therefore, we choose the convolutional decoder as the default configuration.

Table 6: Analysis of different decoder designs.

Decoder	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
Linear	70.1	66.5	28.5	25.9	29.8	26.2	19.8
Transformer Block	70.3	66.9	28.1	25.5	29.3	25.6	18.5
3 × 3 Conv	70.4	67.1	27.9	25.3	29.1	25.1	17.2

Note: ↑ denotes that higher values are better, ↓ denotes that lower values are better.

4.3.5 Masking Ratio

The masking ratio is an important hyper-parameter. A low ratio is hard to provide enough occlusion information for the reconstruction network to learn inter-modal feature alignment, while an excessively high ratio erases too much spatial context, damaging the visual clues completely. Fig. 5 illustrates the impact of masking ratios ranging from 40% to 80% on detection performance. The results exhibit an inverted U-shaped trend, peaking at a masking ratio of 60% where 67.1% mAP and 70.4% NDS are achieved. When the masking ratio decreases, the masked BEV feature maps are easy to be reconstructed, making the model less robust to the occlusion scenarios. Conversely, increasing the masking ratio results in an abundance of masked regions, causing substantial damage to the context information of the BEV feature maps.

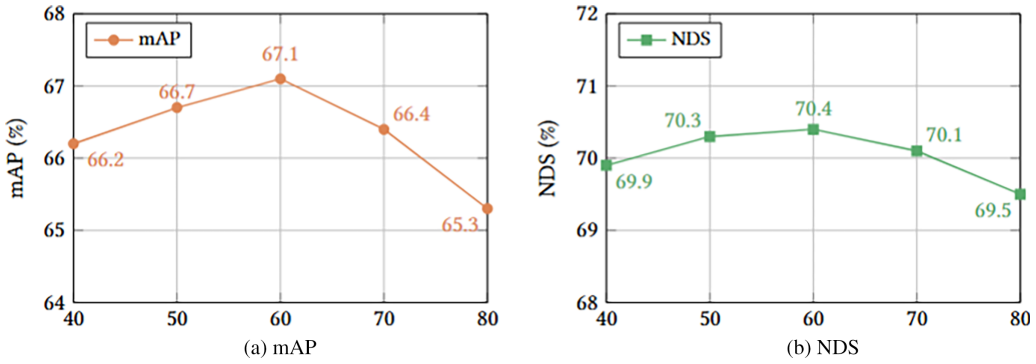
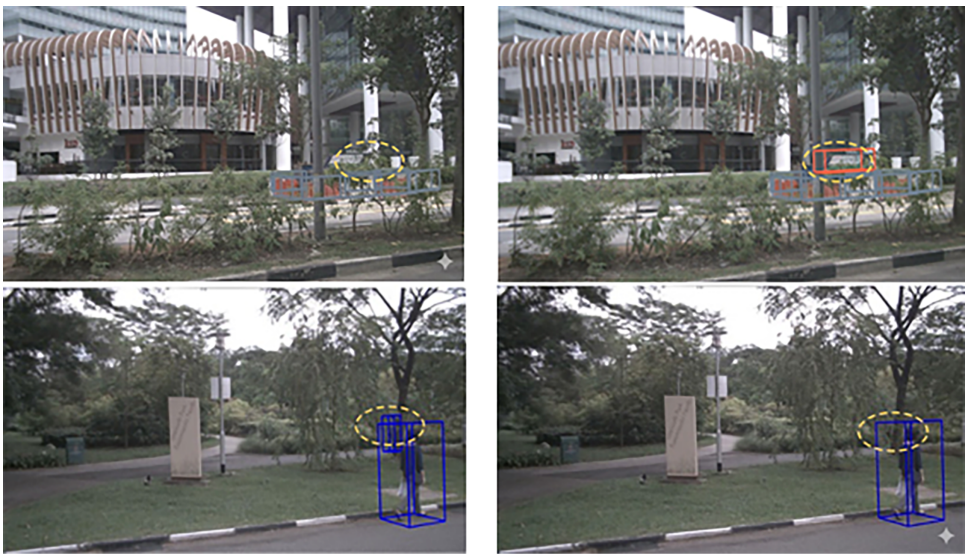


Figure 5: The effect of the masking ratio. (**a**) mAP, (**b**) NDS.

4.4 Failure Case

Fig. 6 illustrates the limitations of CG-MAE in handling specific extreme conditions, resulting in both false positives and missed detections. In the yellow elliptical region of the first row, CG-MAE detects an occluded trash bin as a vehicle. This false positive reveals that cross-modal semantic confusion may occur when background plants, such as trees cause severe occlusion. In the second row, CG-MAE has a missed detection when identifying a pedestrian behind a tree. This is because the features of small targets (such as distant and occluded pedestrians) are inherently sparse. During reconstruction, if the complementary modal sensor also fails to perceive them, it becomes difficult to recover the missed objects.



(a) Ground Truth (b) CG-MAE

Figure 6: Failure cases: (**a**) ground truth, (**b**) detection results of CG-MAE.

5 Conclusion and Future Work

This paper has proposed a dual-branch BEV masked autoencoder framework based on cross-modal guidance for 3D object detection in autonomous driving. The GT-based foreground masking strategy is designed to mimic the object obscuration caused by occlusion or adverse environments. To reconstruct the masked BEV feature maps, a cross-modal guided reconstruction network is developed to predict the representations of the obscured objects in the BEV space. Considering that all the modality data can be obscured, this paper builds a dual-branch BEV masked autoencoder to ensure the bi-directional alignment. Compared to the baseline method, the proposed framework has achieved promising performance gains on the nuScenes dataset.

Despite the great performance on multi-modal 3D object detection, the proposed method still exhibits limitations under some conditions. As the presented results in failure cases, the cross-modal guided reconstruction may suffer from semantic confusion in scenarios with severe occlusion and complex backgrounds. Moreover, the proposed method focuses on the 3D detection task, ignoring other tasks in autonomous driving systems.

To tackle these limitations, our future work will integrate temporal information across continuous frames to capture dynamic motion cues. By aggregating historical features to the present, the network tends to mitigate the adverse impacts of occlusions and sensor noise on overall detection accuracy. We also plan to extend the proposed framework to multi-task learning to obtain a comprehensive perception for the surrounding environment of the ego vehicle. Furthermore, we aim to explore the scalability of the dual-branch cross-modal guided reconstruction mechanism across additional sensor modalities, such as radar-camera fusion, to further enhance perception resilience in extreme autonomous driving environments.

Acknowledgement: Not applicable.

Funding Statement: This paper is supported by the National Natural Science Foundation of China (No. 62301497); the Science and Technology Research Program of Henan (No. 252102211024); the Key Research and Development Program

of Henan (No. 231111212000); and the Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2021-04244.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Shouyi Yang and Song Wang; Methodology, Shouyi Yang, Junchen Huo and Song Wang; Software, Junchen Huo; Validation, Junchen Huo; Formal analysis, Shouyi Yang; Investigation, Shouyi Yang and Junchen Huo; Resources, Song Wang; Data curation, Junchen Huo; Writing—original draft, Junchen Huo; Writing—review and editing, Song Wang, Enqing Chen, Yingqiang Ding and Shouyi Yang; Visualization, Junchen Huo; Supervision, Song Wang, Enqing Chen, Yingqiang Ding and Shouyi Yang; Project administration, Shouyi Yang and Song Wang; Funding acquisition, Shouyi Yang and Song Wang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study include the publicly available nuScenes dataset, which can be accessed via its official website: https://www.nuscenes.org. This dataset was originally introduced in the work of Caesar et al.

Ethics Approval: Not applicable.

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

References

1. Shi P, Yang L, Dong X, Qi H, Yang A. Research progress on multi-modal fusion object detection algorithms for autonomous driving: a review. Comput Mater Contin. 2025;83(3):3877–917. doi:10.32604/cmc.2025.063205.
2. Shi P, Liu Z, Qi H, Yang A. MFF-net: multimodal feature fusion network for 3D object detection. Comput Mater Contin. 2023;75(3):5615–37. doi:10.32604/cmc.2023.037794.
3. Mao J, Shi S, Wang X, Li H. 3D object detection for autonomous driving: a comprehensive survey. Int J Comput Vis. 2023;131(8):1909–63. doi:10.1007/s11263-023-01790-1.
4. Jiao T, Chen Y, Zhang Z, Guo C, Song J. MMDistill: multi-modal BEV distillation framework for multi-view 3D object detection. Comput Mater Contin. 2024;81(3):4307–25. doi:10.32604/cmc.2024.058238.
5. Li Y, Bao H, Ge Z, Yang J, Sun J, Li Z. BEVStereo: enhancing depth estimation in multi-view 3D object detection with temporal stereo. Proc AAAI Conf Artif Intell. 2023;37(2):1486–94. doi:10.1609/aaai.v37i2.25234.
6. Li Y, Ge Z, Yu G, Yang J, Wang Z, Shi Y, et al. BEVDepth: acquisition of reliable depth for multi-view 3D object detection. Proc AAAI Conf Artif Intell. 2023;37(2):1477–85. doi:10.1609/aaai.v37i2.25233.
7. Li Z, Wang W, Li H, Xie E, Sima C, Lu T, et al. BEVFormer: learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers. IEEE Trans Pattern Anal Mach Intell. 2024;47(3):2020–36. doi:10.1109/TPAMI.2024.3515454.
8. Chen S, Ma Y, Qiao Y, Wang Y. M-BEV: masked BEV perception for robust autonomous driving. Proc AAAI Conf Artif Intell. 2024;38(2):1183–91. doi:10.1609/aaai.v38i2.27880.
9. Alaba SY, Ball JE. A survey on deep-learning-based LiDAR 3D object detection for autonomous driving. Sensors. 2022;22(24):9577. doi:10.3390/s22249577.
10. Lang AH, Vora S, Caesar H, Zhou L, Yang J, Beijbom O. PointPillars: fast encoders for object detection from point clouds. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 12689–97.
11. 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); 2021 Jun 20–25; Nashville, TN, USA. p. 11779–88.
12. Shao Z, Wang H, Cai Y, Chen L, Li Y. UA-fusion: uncertainty-aware multimodal data fusion framework for 3-D object detection of autonomous vehicles. IEEE Trans Instrum Meas. 2025;74:2514916. doi:10.1109/TIM.2025.3548184.

13. Liu Z, Tang H, Amini A, Yang X, Mao H, Rus DL, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation. In: Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29–Jun 2; London, UK. p. 2774–81. doi:10.1109/ICRA48891.2023.10160968.
14. Bai X, Hu Z, Zhu X, Huang Q, Chen Y, Fu H, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 1080–9.
15. Yin T, Zhou X, Krähenbühl P. Multimodal virtual point 3D detection. Adv Neural Inf Process Syst. 2021;34:16494–507.
16. Sindagi VA, Zhou Y, Tuzel O. MVX-net: multimodal VoxelNet for 3D object detection. In: Proceedings of the 2019 International Conference on Robotics and Automation (ICRA); 2019 May 20–24; Montreal, QC, Canada. p. 7276–82.
17. Xie L, Xiang C, Yu Z, Xu G, Yang Z, Cai D, et al. PI-RCNN: an efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module. Proc AAAI Conf Artif Intell. 2020;34(7):12460–7. doi:10.1609/aaai.v34i07.6933.
18. Dou J, Xue J, Fang J. SEG-VoxelNet for 3D vehicle detection from RGB and LiDAR data. In: Proceedings of the 2019 International Conference on Robotics and Automation (ICRA); 2019 May 20–24; Montreal, QC, Canada. p. 4362–8.
19. 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; 2018 Sep 8–14; Munich, Germany. p. 641–56.
20. Chen Z, Li Z, Zhang S, Fang L, Jiang Q, Zhao F, et al. AutoAlign: pixel-instance feature aggregation for multi-modal 3D object detection. arXiv:2201.06493. 2022.
21. Chen Z, Li Z, Zhang S, Fang L, Jiang Q, Zhao F. Deformable feature aggregation for dynamic multi-modal 3D object detection. In: Proceedings of the 17th European Conference on Computer Vision, ECCV 2022; 2022 Oct 23–27; Tel Aviv, Israel. p. 628–44.
22. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of the 9th International Conference on Learning Representations (ICLR); 2021 May 3–7; Vienna, Austria.
23. Wolters P, Gilg J, Teepe T, Herzog F, Fent F, Rigoll G. SpaRC: sparse radar-camera fusion for 3D object detection. arXiv:2411.19860. 2024.
24. Ge C, Chen J, Xie E, Wang Z, Hong L, Lu H, et al. MetaBEV: solving sensor failures for 3D detection and map segmentation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 8687–97.
25. Chen A, Zhang K, Zhang R, Wang Z, Lu Y, Guo Y, et al. PiMAE: point cloud and image interactive masked autoencoders for 3D object detection. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 5291–301.
26. Wang B, Zhao Q, Tao C, Sun Y, Yan C. IAE-BEV: instance-adaptive enhancement for BEV-based multi-view 3D object detection. IEEE Robot Autom Lett. 2025;10(10):10610–7. doi:10.1109/LRA.2025.3604760.
27. Li Z, Singh B. Robust occluded object detection in multimodal autonomous driving: a fusion-aware learning framework. Electronics. 2026;15(1):245. doi:10.3390/electronics15010245.
28. Zou J, Huang T, Yang G, Guo Z, Luo T, Feng CM, et al. UniM^{2}AE: multi-modal masked autoencoders with unified 3D representation for 3D perception in autonomous driving. In: Proceedings of the 18th European Conference on Computer Vision, ECCV 2024; 2024 Sep 29–Oct 4; Milan, Italy. p. 296–313.
29. Lin Z, Wang Y, Qi S, Dong N, Yang MH. BEV-MAE: bird's eye view masked autoencoders for point cloud pre-training in autonomous driving scenarios. Proc AAAI Conf Artif Intell. 2024;38(4):3531–9. doi:10.1609/aaai.v38i4.28141.
30. Zhang Z, Cao Y, Zhang H, Chen N, Ren C, Li Y. Vision BEV-MAE: masked bird's eye view autoencoders for camera-based 3D object detection. In: Proceedings of the 2024 China Automation Congress (CAC); 2024 Nov 1–3; Qingdao, China. p. 4409–14.

31. Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, et al. nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11618–28.
32. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88.
33. Xue Y, Yao C, Wahib M, Gabbouj M. YOLO-DKR: differentiable architecture search based on kernel reusing for object detection. Inf Sci. 2025;713(15):122180. doi:10.1016/j.ins.2025.122180.
34. Zhang Z, Liang Z, Zhang M, Zhao X, Li H, Yang M, et al. RangeLVDet: boosting 3D object detection in LIDAR with range image and RGB image. IEEE Sens J. 2022;22(2):1391–403. doi:10.1109/JSEN.2021.3127626.
35. Wu X, Peng L, Yang H, Xie L, Huang C, Deng C, et al. Sparse fuse dense: towards high quality 3D detection with depth completion. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5408–17.
36. Zhang X, Wang L, Zhang G, Lan T, Zhang H, Zhao L, et al. RI-fusion: 3D object detection using enhanced point features with range-image fusion for autonomous driving. IEEE Trans Instrum Meas. 2023;72:5004213. doi:10.1109/TIM.2022.3224525.
37. Deng J, Czarnecki K. MLOD: a multi-view 3D object detection based on robust feature fusion method. In: Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC); 2019 Oct 27–30; Auckland, New Zealand. p. 279–84.
38. Hao X, Diao Y, Wei M, Yang Y, Hao P, Yin R, et al. MapFusion: a novel BEV feature fusion network for multi-modal map construction. Inf Fusion. 2025;119(3):103018. doi:10.1016/j.inffus.2025.103018.
39. Wang H, Tang H, Shi S, Li A, Li Z, Schiele B, et al. UniTR: a unified and efficient multi-modal transformer for bird's-eye-view representation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 6769–79.
40. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 15979–88.
41. Zou J, Liao B, Zhang Q, Liu W, Wang X. MIM4D: masked modeling with multi-view video for autonomous driving representation learning. Int J Comput Vis. 2025;133(9):6074–87.
42. Xie E, Yu Z, Zhou D, Philion J, Anandkumar A, Fidler S, et al. M^2BEV: multi-camera joint 3D detection and segmentation with unified birds-eye view representation. arXiv:2204.05088. 2022.
43. Lai Z, Ren M, Sheng S, Liu C, Zhang Z. KAN-RCBEVDepth: integrating multi-modal sensor data for robust 3D object detection in autonomous driving. In: Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN); 2025 Jun 30–Jul 5; Rome, Italy. p. 1–8.
44. Li Y, Chen Y, Qi X, Li Z, Sun J, Jia J. Unifying voxel-based representation with transformer for 3D object detection. Adv Neural Inf Process Syst. 2022;35:18442–55. doi:10.52202/068431-1340.
45. Li J, Luo C, Yang X. PillarNeXt: rethinking network designs for 3D object detection in LiDAR point clouds. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 17567–76.
46. Wang S, Caesar H, Nan L, Kooij JFP. UniBEV: multi-modal 3D object detection with uniform BEV encoders for robustness against missing sensor modalities. In: Proceedings of the 2024 IEEE Intelligent Vehicles Symposium (IV); 2024 Jun 2–5; Jeju Island, Republic of Korea. p. 2776–83.
47. Song Z, Wei H, Bai L, Yang L, Jia C. GraphAlign: enhancing accurate feature alignment by graph matching for multi-modal 3D object detection. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 3335–46.

48. Vora S, Lang AH, Helou B, Beijbom O. PointPainting: sequential fusion for 3D object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 4603–11.
49. Chen X, Zhang T, Wang Y, Wang Y, Zhao H. FUTR3D: a unified sensor fusion framework for 3D detection. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2023 Jun 17–24; Vancouver, BC, Canada. p. 172–81.