Adaptive-Smooth LiDAR-Camera Knowledge Distillation with Heterogeneous Fusion for Multi-View 3D Object Detection

Rui Zhao[1], Shuoyao Wang[1*], Xinhu Zheng[2], Shijian Gao[2,3]
[1] Shenzhen University, Shenzhen
[2] The Hong Kong University of Science and Technology, Guangzhou
[3] Guangdong Provincial Key Laboratory of Future Networks of Intelligence, The Chinese University of Hong Kong, Shenzhen
zhaorui2022@email.szu.edu.cn, sywang@szu.edu.cn, xinhuzheng@hkust-gz.edu.cn, shijiangao@hkust-gz.edu.cn

Abstract

Multi-view 3D object detection has garnered increasing attention, particularly due to its success in autonomous driving systems. Although multi-view systems possess rich semantic information, their spatial-geometric reasoning capabilities remain limited. Recent studies employ simulated point cloud generation mechanisms to facilitate LiDAR-camera multi-modal knowledge distillation, achieving formal structural consistency. Despite advancements, these methods still face two main issues: i) alignment challenges caused by discrepancies between LiDAR and camera data, and ii) prediction errors from simulated point clouds that compromise the semantic information extracted from images during fusion. To address these problems, we propose adaptive-smooth distillation to optimize alignment granularity based on feature discrepancies for improved LiDAR-camera knowledge distillation. Specifically, this work considers both LIDAR-to-camera cross-modal distillation and LiDAR-camera fusion to simulated point cloud-camera fusion multi-modal distillation. Then, we introduce a heterogeneous fusion module to strategically bias the fusion process toward the extracted camera features, thereby enhancing the robustness of the fusion feature. Additionally, soft-weighted response distillation is proposed to facilitate the student model to selectively mimic the high-quality output of the teacher model. Extensive experiments have demonstrated the superiority of our method, achieving statistically significant improvements of 4.9% in mean Average Precision (mAP) and 4.5% in NuScenes Detection Score (NDS) over the benchmark.

1 Introduction

3D object detection provides richer spatial information than 2D, with applications in autonomous driving, robotics, and industrial automation (Song et al. 2024; Mao et al. 2023; Wang et al. 2023a; Ma et al. 2023). Although LiDAR-based and multi-modal 3D object detection has achieved notable progress, persistent challenges remain in high latency, high cost, and hardware degradation (Huang et al. 2025; Qi et al. 2024; Qian, Lai, and Li 2022). Multi-view 3D object detection has recently garnered significant attention due to its low cost, low loss, and high resolution (Yang et al. 2025; Liu et al. 2023a; Wang et al. 2025). Although the BEV-based approach offers improved robustness and scalability for autonomous driving in complex environments (Wang et al. 2024; Zhang et al. 2023a; Zhu et al. 2023), depth estimation errors from multi-view approaches become amplified during viewpoint transformation. Overall, the 3D object detection performance achieved by multi-view based methods still lags behind the LiDAR-based and multi-modal approaches (Li et al. 2023a; Zhang, Hou, and Yuan 2024).

To bridge this gap, several studies have explored distillation methods to transfer precise 3D spatial information from the LiDAR point cloud to multi-view camera models,

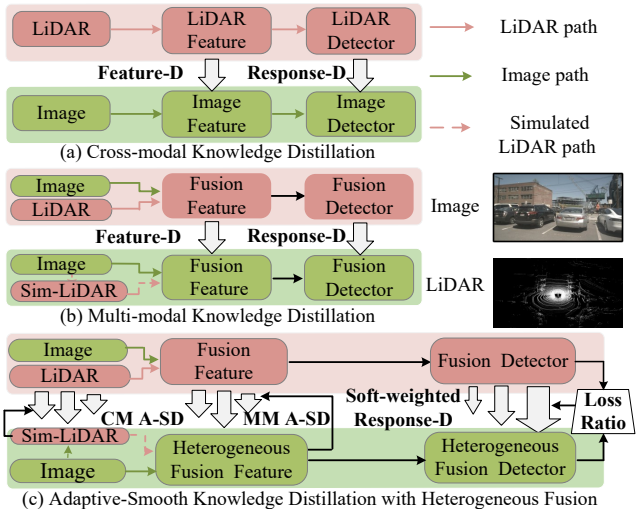

Figure 1: Comparison of different distillation frameworks: (a) LiDAR-to-camera cross-modal distillation transfers 3D geometric awareness but faces structural alignment gaps. (Feature-D and Response-D stand for Feature Distillation and Response Distillation, respectively.) (b) Multi-modal distillation aligns teacher and student models using simulated point clouds, yet struggles with input modality disparities. (c) Our adaptive-smooth distillation integrates heterogeneous fusion to effectively reduce side effect of input discrepancies and enhance fine-grained knowledge transfer.(CM A-SD: Cross-Modal Adaptive-Smooth Distillation; MM A-SD: Multi-modal Adaptive-Smooth Distillation.)

*Corresponding Author

known as cross-modal distillation (Kim et al. 2024; Jang et al. 2023; Wang et al. 2023b). Illustrated in Figure 1(a), BEV feature distillation and response distillation are commonly employed as fundamental strategies to facilitate the student model to mimic both the feature and the output of the LiDAR teacher model (Marvin et al. 2023; Ji et al. 2024b; Chen et al. 2023). Nevertheless, these approaches overlook the challenges posed by the inherent structural differences between teacher and student models. As depicted in Figure 1(b), SimDistill (Zhao et al. 2024a) simulates point-cloud features using image features to ensure structural consistency between teacher and student models. However, rigid feature alignment may lead to overfitting to noise and gradient instability in the student model when substantial teacher-student discrepancies arise from heterogeneous input data. Furthermore, noise inherent in simulated point clouds could impair multi-modal fusion feature, thereby compromising subsequent distillation processes and model convergence.

In this paper, we propose an adaptive-smooth distillation framework incorporating heterogeneous fusion and soft-weighted response distillation, enabling the student model to mimic the feature and output of the teacher model in a soft, stable, and selective manner. As shown in Figure 1(c), to reduce the negative impact of differences in input modality during distillation, we introduce an adaptive-smooth strategy to align the features in a soft way. This method dynamically calibrates alignment constraints, relaxing them for highly dissimilar cross-modal samples to mitigate over-regularization, while intensifying them for similar instances to enable high-precision knowledge transfer. The key is to maintain the student model's mimicry flexibility via soft learning. To address unreliable simulated point cloud features, the student model employs a heterogeneous fusion mechanism that prioritizes extracted image features over simulated point clouds during feature fusion. Furthermore, directly mimicking all outputs from the teacher model may introduce unnecessary noise and thus potentially mislead the student model. To tackle the output noise from the teacher model, we introduce a soft-weighted knowledge distillation approach with classification and regression joint-quality scores to suppress noise while enhancing gradient allocation to high-confidence predictions. In summary, we present the key contributions as follows:

- We propose a novel LiDAR-camera distillation framework that mitigates the negative effects of the discrepancies in the predicted features and the extracted representations while also effectively alleviating the interference caused by unreliable simulated point cloud features.
- We introduce an adaptive-smoothing distillation framework that modulates alignment granularity through the assessment of modal discrepancies, thus effectively mitigating both misalignment and feature degradation during knowledge transfer.
- We present a heterogeneous fusion mechanism that alleviates noise affects by biasing toward camera features, alongside response distillation to selectively mimic high-quality teacher outputs. Our method improves mAP by 4.9% and NDS by 4.5% on the nuScenes benchmark.

2 Related Work

Multi-View 3D Object Detection With the ongoing advancement of autonomous driving, multi-view 3D object detection is gaining increasing attention (Lin et al. 2022; Liu et al. 2022; Zhao et al. 2024b; Wen et al. 2023). The BEV representation systematically mitigates occlusion and scale distortion inherent in monocular 3D perception through multi-view geometric consistency, emerging as the de facto paradigm for robust environmental understanding in 3D object detection.

The BEV-based approach can currently be categorized into two main types: depth estimation methods and query-based methods. For depth estimation based methods, such as LSS (Philion and Fidler 2020), which first design a depth estimation network to predict the depth of each pixel before lifting it to the 3D space. Building on this foundation, BEVDet (Huang et al. 2021) integrates inputs from multiple cameras to create a unified Bird's Eye View (BEV) space, thereby eliminating the need for complex 2D-to-3D post-processing. In addition, BEVDepth (Li et al. 2023b) improves the BEV features by explicitly estimating the depth of the detection loss. Query-based methods generally work by learning a set of fixed or learnable queries (Song et al. 2024; Chen et al. 2022). In contrast to purely attention-based approaches, BEVFormer (Li et al. 2022) introduces a geometry-attention hybrid framework, where explicit BEV coordinate encoding and implicit cross-view attention jointly optimize the trade-off between prior-based reasoning and data-driven perception. Additionally, to reduce redundant computation, SparseBEV (Liu et al. 2023a) employs scale-adaptive attention for multi-scale feature interaction and adaptive spatio-temporal sampling for motion-aware feature selection.

Knowledge Distillation for Multi-View Object Detection Knowledge distillation enhances the performance of student models by transferring knowledge from large, high-performance teacher models to smaller, less complex student models (Shen et al. 2024; Li et al. 2024). Regarding 3D multi-view object detection, cross-modal distillation has emerged as a prevalent strategy, which leverages LiDAR point cloud as the teacher model to transfer spatial reasoning capabilities to the multi-view student model, thereby enhancing the camera model's ability to infer 3D structures and geometry information (Huang et al. 2023; Jang et al. 2023; Hong, Dai, and Ding 2022). As a pioneering framework, BEVDistill (Chen et al. 2023) emphasizes the foreground regions by generating Gaussian masks based on the center of the GT boxes, and it also combines confidence scores and the IoU of the predicted boxes to select high-value instances. In addition, UniDistill (Zhou et al. 2023) employs a triple distillation strategy (i.e., feature, relation, and Response Distillation) to suppress the background while preserving semantic consistency in the sparse foreground feature. Next, by dividing the regions, DistillBEV (Wang et al. 2023b) designs the regional distillation to reduce interference between the background and the imbalanced samples. In light of these methods, SimDistill (Zhao et al. 2024a) adopts a LiDAR-camera multi-modal model as the teacher and achieves structural alignment between the student and

teacher models through the simulated point cloud.

3 Methodology

3.1 Overview

The overall architecture of our proposed method is shown in Fig. 2. The image branch extracts image BEV features from the multi-view images input, while the LiDAR branch extracts LiDAR BEV features from the LiDAR point cloud input. Then, the BEV features are fused and fed to the detection head to produce the detection output. In the student model, due to the absence of LiDAR data, the LiDAR branch is replaced with a LiDAR BEV feature prediction module, following SimDistill (Zhao et al. 2024a). In particular, we consider three key components in this paper: adaptive-smooth distillation, heterogeneous fusion, and soft-weighted response distillation. We first propose the adaptive-smooth distillation method to adaptively adjust the alignment granularity to reduce the negative impact of the input heterogeneity. Then, we also introduce a heterogeneous fusion module that dynamically enhances the fusion weight of camera features to improve robustness of the fusion feature. Finally, we employ a soft-weighted response distillation approach (including regression and classification distillation), which quantifies output discrepancies between teacher and student models to selectively calibrate the distillation weights. In essence, the primary objective is to improve the spatial reasoning abilities of the student model. And it will be achieved through the above strategies while preserving its inherent semantic richness.

3.2 Adaptive-Smooth Distillation

Although the student model has achieved a multi-modal fusion by generating the simulated point cloud to ensure structural consistency with the teacher model, there is an inherent discrepancy in their input modalities. Specifically, the student model relies solely on camera data, while the teacher model utilizes both camera and LiDAR data. Due to significant disparities in input modalities, employing loss functions such as MSE to make the student model mimic all features of the teacher model and enforce strict alignment within the same feature space may result in the loss of modality-specific information.

Building on this, we propose an adaptive-smooth distillation method that addresses the constraints of rigid alignment through elastic feature correlation between two modals. The method implements an adaptive distillation mechanism that allows a degree of spatial misalignment when teacher-student discrepancy thresholds are exceeded, while enabling precise optimization when they fall below.

It is critical to clarify that this method is effective for both: (1) cross-modal alignment between LiDAR point clouds and simulated point clouds, and (2) multi-modal alignment of fused features. These two applications differ only in their input feature. In the multi-modal context, the input features specifically denote the fused embeddings, while cross-modal distillation refers to the LiDAR or the simulated point cloud features. *To differentiate this mechanism across distillation stages, we term them Multi-Modal Adaptive-Smooth Distillation as MM A-SD and Cross-Modal Adaptive-Smooth Distillation as CM A-SD, respectively.*

Specifically, we first compute the cosine similarity error between features of two modals, as shown in Equation (1):

$$error_i = 1 - \frac{\langle S_i, T_i \rangle}{\|S_i\| \|T_i\|}, \tag{1}$$

where S_i and T_i represent the feature vectors of the i^{th} samples from both the student model and the teacher model, respectively. Then, we calculate the dynamic boundary margin, which is defined as the sum of the mean and the standard deviation of the relevant errors in Equation (2):

$$margin = \mu_{\text{error}} + \frac{1}{2} \cdot \sigma_{\text{error}}, \tag{2}$$

where μ_{error} denotes the mean of the error distributions, σ_{error} represents the standard deviation of the error. This integrates statistical distribution coverage with cognitive uncertainty modeling, combining the center of error distribution and variability into the boundary to encompass core error ranges.

Subsequently, we calculate the alignment loss as described in Equation (3):

$$\mathcal{L}_{align} = \frac{1}{B} \sum_{i=1}^{B} \begin{cases} \frac{1}{2}\|\boldsymbol{e}_i\|^2, & \text{if } \|\boldsymbol{e}_i\| < M, \\ M \cdot \left(\|\boldsymbol{e}_i\| - \frac{1}{2}M\right), & \text{otherwise,} \end{cases} \tag{3}$$

where M denotes the margin, B indicates the batch size and $\boldsymbol{e}_i$ is the cosine similarity error vector for the i-th sample. When the error is within margin, indicating minor feature discrepancies, quadratic loss ensures fine-grained alignment. Conversely, for larger errors exceeding the margin, the loss becomes linear to avoid excessive penalty and allow tolerable misalignment.

After that, we also introduce a contrastive loss to further enhance the distillation process. The contrastive loss helps the student model develop discriminative features and prevents feature collapse. As shown in Equation (4):

$$L_{con} = \frac{1}{N} \sum_{i=1}^{N} \left(\log \sum_{j=1}^{N} e^{S_{ij}/\tau} - \frac{S_{ii}}{\tau} \right), \tag{4}$$

where $S_{i,j}$ means the similarity score between the i^{th} and j^{th} samples from the teacher and student model, when $j \neq i$, is treated as a negative sample pair. The S_{ii} indicates the similarity score of the i^{th} sample to its positive sample pair. And positive samples are defined as the feature pairs of identical input data from the teacher model and the student model. τ is the temperature parameters to control the sharpness of the similarity distribution.

Finally, the adaptive-smooth distillation loss is formulated as a weighted combination of the alignment and contrastive loss components. As shown in Equation (5):

$$\mathcal{L}_{A\text{-}SD} = \alpha \mathcal{L}_{align} + (1 - \alpha)\mathcal{L}_{con}. \tag{5}$$

The α update mechanism is dynamically modulated via an Exponential Moving Average (EMA) to harmonize the

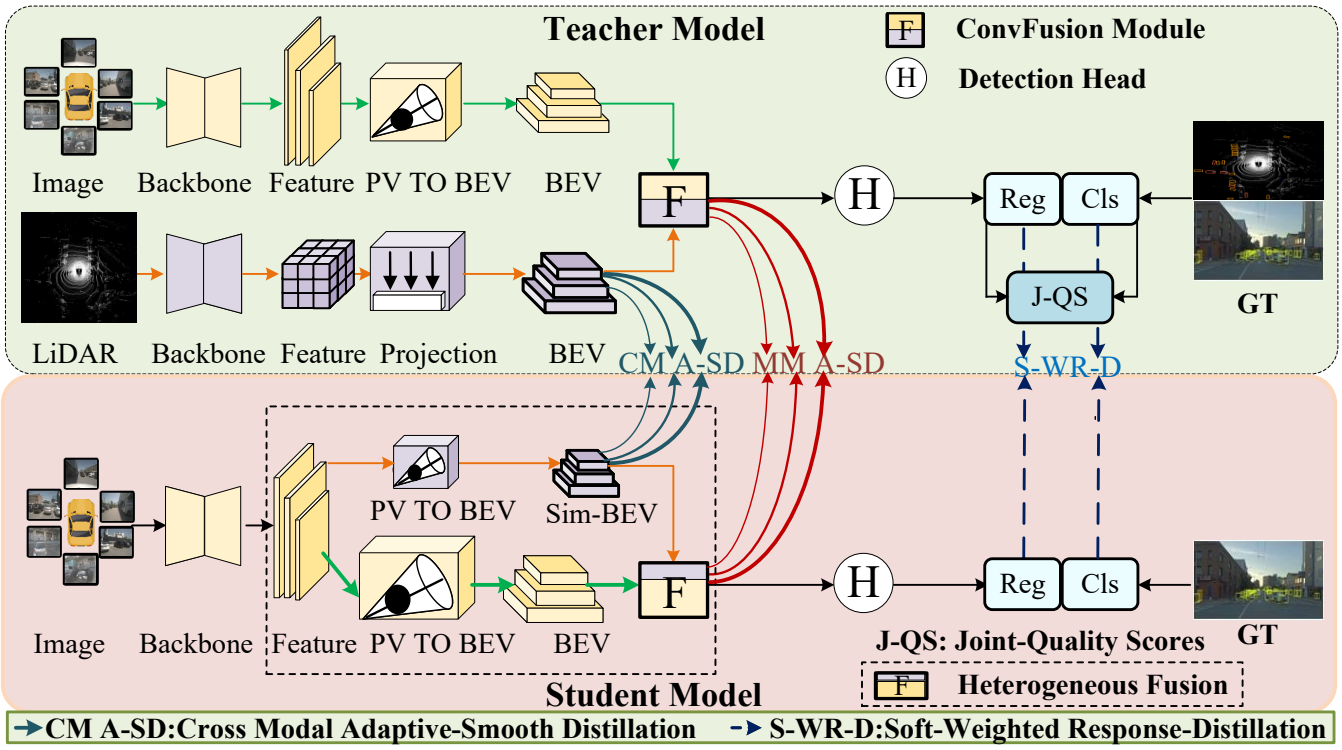

Figure 2: The overall architecture of our proposed method. The teacher model is the LiDAR-camera framework, and the student is the camera-simulated point cloud model. We first propose adaptive-smooth distillation to adaptively adjust alignment granularity, and also introduce a heterogeneous fusion module that dynamically enhances fusion weights between camera features. Finally, we adopt soft-weighted response distillation, a soft balancing mechanism that selectively calibrates distillation weights.

weighting between alignment loss and contrastive loss. And α is updated in Equation (6):

$$\alpha = (1 - \lambda) \cdot \alpha_{init} + \lambda \cdot r, \tag{6}$$

where α_{init} is the initial value. And the default value is 0.5. λ is the smoothing factor and the value is 0.1. The r is the ratio of L_{align} to the sum of L_{align} and L_{con}. This strategy adaptively modulates the weighting coefficients based on the relative magnitudes of task-specific losses, thereby enabling the model to autonomously regulate the inter-task equilibrium throughout training.

3.3 Heterogeneous Fusion Module

To mitigate the side effects of the low-confidence simulated point clouds on multi-modal distillation, we employ heterogeneous fusion during feature fusion in the student model, giving priority to camera features over simulated point cloud features instead of applying equal-weight fusion or channel concatenation fusion.

As depicted in Figure 3, camera features are enhanced by a dual convolution enhance module integrated with a channel attention mechanism, which reinforces visual textures and increases information density. The Simulated LiDAR point cloud features is processed through a module that employs normalization to suppress noise fluctuations and utilizes the Tanh activation function to compress the value range to [0, 1]. This module ensures the stability and predictability of the simulated LiDAR point clouds.

The porcess of the simulated LiDAR point cloud assessment module is shown in Equation (7):

$$L_{conf} = \sigma\left(W_c^2 \cdot R\left(W_c^1 \cdot X_{\text{liDAR}}\right)\right), \tag{7}$$

where W_c^1 and W_c^2 indicate the weights of the first and second convolution layers, σ is the sigmoid activation function and R is the ReLu activation function. This spatially aware simulated LiDAR point cloud feature modeling generates a confidence heatmap to guide the allocation of fusion weights.

Then we calculate the base fusion weight by gate fusion in Equation (8):

$$W_{base} = \sigma\left(W_f^2 \cdot R\left(W_f^1 \cdot A(X'_{cam} \copyright X'_{liDAR})\right)\right), \tag{8}$$

where W_f is the weight matrix of the convolution layer, © denotes the concatenation operation, A is the Average Pooling operation, X'_{cam} and X'_{liDAR} represents preprocessed features of the camera and the simulated LiDAR, respectively. This step is to evaluate the optimal ratio of feature fusion. After that, the average confidence is calculated in Equation (9):

$$\bar{C} = \frac{1}{HW} \sum^{H} \sum^{W} L_{\text{conf}}(h, w), \tag{9}$$

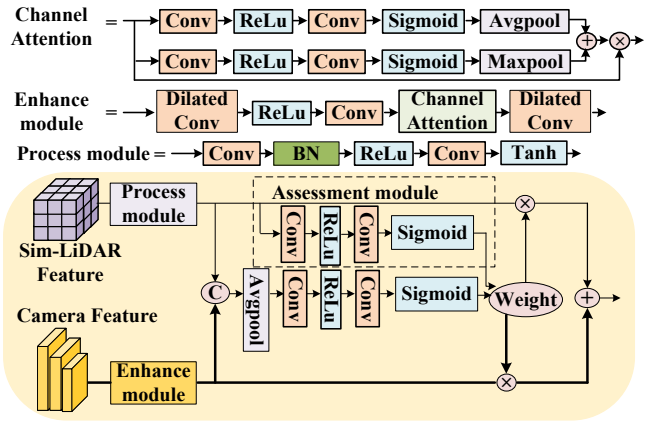

Figure 3: The overall framework of the Heterogeneous Fusion Module.

where $H \times W$ denotes the global resolution. Then the dynamic weight adjustment is performed in Equation (10):

$$W_{cam} = [0.5 + \sigma(\beta)(1 - \bar{C})] \cdot W_{base} \tag{10}$$

where $\sigma(\beta)$ is the learnable camera bias parameter. This ensures that the fusion weight of the image side is greater than 0.5 and less than 1. The total weight, including the simulated LiDAR point cloud and the camera, is set to 1 to avoid conflicting weight assignments. The final fusion feature is shown in Equation (11):

$$Y_{fused} = W_{cam} \odot X'_{cam} + W_{liDAR} \odot X'_{liDAR}, \tag{11}$$

3.4 Soft-weighted Response Distillation

To effectively align the outputs of the student model with the predictions of the teacher model, we propose a filterable output distillation mechanism called soft-weighted response distillation. This work is inspired by Unbiased Teacher v2 (Liu, Ma, and Kira 2022), which incorporates the principle of targeted feature selection. This method strategically avoids mimicking all outputs of the teacher model through dynamic weight allocation, thus mitigating negative transfer of the noise output. Firstly, we consider both classification and regression aspects to evaluate the teacher's output and derive a quality score. Specifically, we obtain the classification confidence at the corresponding position from the heatmap and select its maximum value as the classification score. Due to the high computational complexity of 3D IoU, we use cosine similarity, which is linearly mapped to the range [0, 1] using an affine transformation, to obtain the regression score. As shown in Equation (12)

$$Q_{Reg} = \frac{1}{2}\left(1 + \frac{\sum a_i b_i}{\sqrt{\sum a_i^2} \cdot \sqrt{\sum b_i^2} + \epsilon}\right), \tag{12}$$

where a_i and b_i represent the i^{th} elements of the feature vectors for the bounding boxes a and b, respectively. And the linear transformation equation is $S_{map} = (S_{cos} + 1)/2$. S_{cos} is the value of cosine similarity. The total quality score is defined in Equation (13):

$$Q = (\mathrm{Q}_{Cls})^{\kappa} \times (\mathrm{Q}_{Reg})^{(1-\kappa)}, \tag{13}$$

where κ is the weighting factor, the default value is 0.4. Then, we calculate the soft weight, as shown in Equation (14) and Equation (15):

$$\alpha_{heatmap} = \sigma\left(\beta \cdot (r_{heatmap} - \gamma)\right), \tag{14}$$

$$\alpha_{bbox} = \sigma\left(\beta \cdot (r_{bbox} - \gamma)\right), \tag{15}$$

where $r_{heatmap}$ and r_{bbox} are the heatmap loss ratio and regression loss ratio, i.e., the ratio of the student model's output loss to that of the teacher model. β is the parameter to control the steepness of the sigmoid function, γ is the shift parameter, and σ is the sigmoid function.

Then, following SimDistill (Zhao et al. 2024a), the soft-weighted response loss $\mathcal{L}_{\text{S-WRD}}$ is the sum of the regression loss L_{reg_d}, i.e., the SmoothL1 loss, and the classification loss L_{cls_d}, which is the quality focal loss (i.e., QFL). As shown in Equation (16):

$$\begin{aligned}\mathcal{L}_{S\text{-}WRD} &= \mathcal{L}_{reg_d} + \mathcal{L}_{cls_d} \\ &= \text{SmoothL1}(B^T, B^S) \cdot \alpha_{bbox} \cdot Q \\ &\quad + \text{QFL}(C^T, C^S) \cdot \alpha_{heatmap} \cdot Q,\end{aligned} \tag{16}$$

where B^T and B^S represent the regression boxes of the teacher and student model. Similarly, C^T and C^S denote the classification prediction by the teacher and student model. The distillation coefficient α dynamically adjusts based on relative performance: increasing when the teacher model outperforms the student and decreasing when the student is better. This mechanism enables the student model to prioritize high-quality outputs from the teacher model while mitigating the noise of low-quality outputs.

3.5 Total Loss

The total loss of our model is defined as in Equation (17):

$$\mathcal{L} = \lambda_1 \mathcal{L}_{MM\,A\text{-}SD} + \lambda_2 \mathcal{L}_{CM\,A\text{-}SD} + \lambda_3 \mathcal{L}_{S\text{-}WRD} + \mathcal{L}_{Det}, \tag{17}$$

where $\{\lambda_i\}_{i=1}^{3}$ are the weighting factors. The formulation includes adaptive-smooth distillation loss, soft-weighted response distillation loss, and the original detection loss.

4 Experiments

4.1 Experiment Setting

Dataset and Metrics. As in previous work, we train and evaluate our method on the large-scale autonomous driving dataset nuScenes(Caesar et al. 2020), which has ten classes for the 3D object detection task. The dataset provides synchronized LiDAR and camera data in 1,000 urban driving scenarios, with a standardized split of 700 training scenarios, 150 validation scenarios, and 150 test scenarios. We use the official evaluation metric, mAP and NDS with five other metrics to measure translation, scale, orientation, velocity, and attribute-related errors, i.e., mATE, mASE, mAOE, mAVE, and mAAE.

Implementation Detail. We conduct the experiments using 8 NVIDIA GeForce RTX A100 GPUs in PyTorch. We implement BEVFusion (Liu et al. 2023c) as the teacher model; the image size is resized to 256×704 and the voxel size of

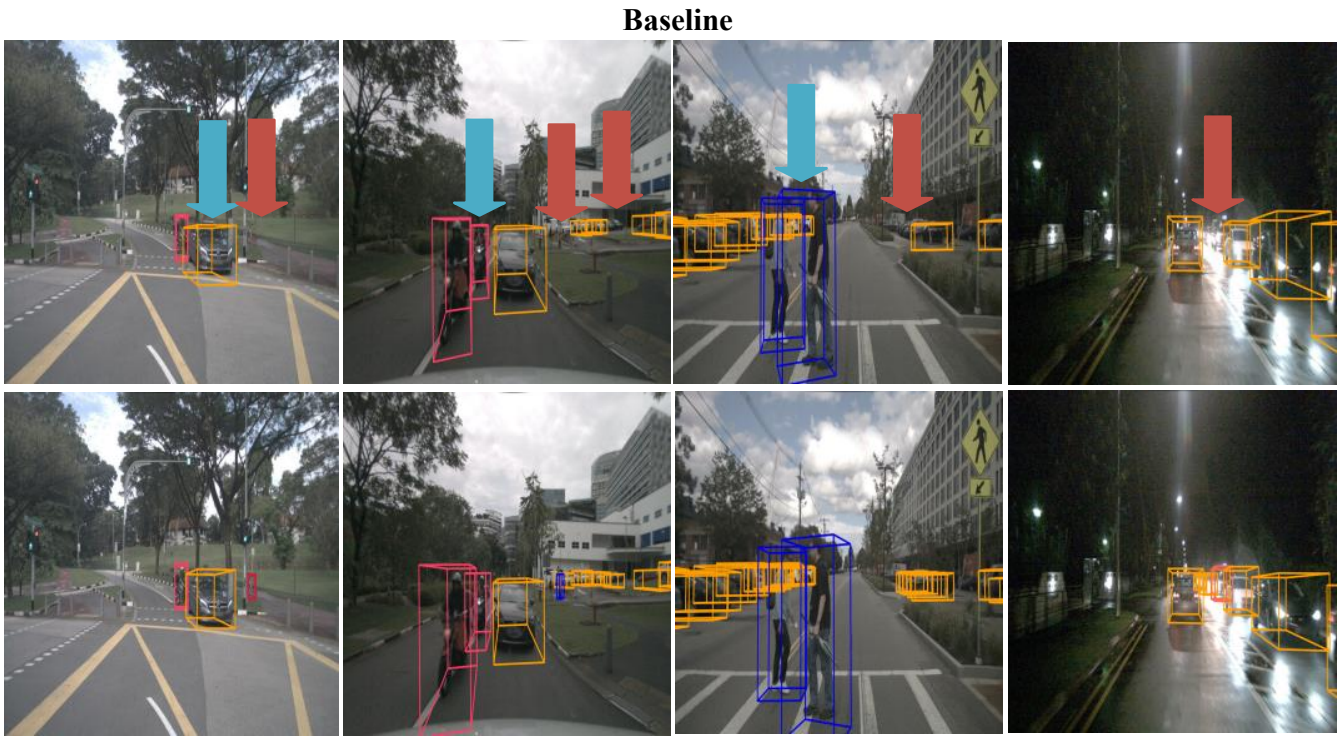

Figure 4: The qualitative comparison of detection results. The top row is the baseline model, which exhibits frequent missed detections and imprecise bounding box regression, whereas our method is the bottom row And our models can significantly reduce missed errors and improves localization accuracy against the baseline model. The red arrows indicate missed detection and blue arrows indicate low-accuracy bounding boxes.

the point cloud is set to (0.075m, 0.075m, 0.2m). The LiDAR backbone is VoxelNet, and the camera backbones are Swin-Transformer and ResNet50. The optimization process utilizes the AdamW algorithm, with learning rate governed by a cosine annealing scheduler.The initial learning rate is configured as 0.0002 to ensure stable convergence during the training phase. Throughout the distillation pipeline, the teacher's weights remain fixed while training the student network for 25 epochs with a batch size of 24, maintaining identical architectural configurations and data augmentations to SimDistill (Zhao et al. 2024a) for our framework.

4.2 Main Results

A comprehensive evaluation of our proposed method was performed on the nuScenes validation set. As shown in Table 1, we classify the benchmarks into the following three categories: non-distillation-based techniques, cross-modal distillation-based approaches, and the baseline methods are employed in our experimental framework. Firstly, we employ BEVfusion-C (Liu et al. 2023c) as the benchmark model in our framework. When employing ResNet50 as the backbone, our method achieves superior mAP compared to all the above approaches, demonstrating improvements of 0.7%, 0.3%, and 4.9% over GeoBEV(Zhang et al. 2025), LabelDistill (Kim et al. 2024), and SimDistill (Zhao et al. 2024a), respectively. Despite inherent limitations of BEVfusion-C (Liu et al. 2023c), the NDS of our method marginally underperforms the BEVDepth-based method (e.g., LabelDistill (Kim et al. 2024), et al.), while demonstrably surpassing BEVfusion-C (Liu et al. 2023c) and SimDistill (Zhao et al. 2024a) by 7.1% and 4.5%, respectively. Similarly, the SwinT backbone architecture yields superior performance, achieving an improvement in mAP and NDS of 3.0% and 5.1% relative to the baseline SimDistill (Zhao et al. 2024a) method. We also add BEVDepth as a benchmark, achieving a 3.7% mAP improvement over LabelDistill (Kim et al. 2024) and an increase of 0.7% in NDS for FSD-BEV(Jiang et al. 2024).

Ablation Studies We conducted comprehensive experiments to evaluate the effectiveness of the individual modules proposed. The experiment is performed on the NuScenes validation set using the SwinT backbone, following the configurations specified in Table 1. As shown in Table 2, MM A-SD and CM A-SD indicate multi-modal and cross-modal adaptive-smooth distillation, HFM means heterogeneous fusion module, and S-WRD represents soft-weighted response distillation. We can see that all proposed methodologies show efficacy, with A-SD (that is, the combination of MM A-SD and CM A-SD) achieving the most substantial performance enhancements of 2.6% NDS, and S-WRD achiev-

Methods	Backbone	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
BEVDet* (Huang et al. 2021)	ResNet50	29.8	37.9	72.5	27.9	58.9	86.0	24.5
BEVDet4D* (Chen et al. 2023)	ResNet50	32.2	45.7	70.3	27.8	49.5	35.4	20.6
PETRv2* (Liu et al. 2023b)	ResNet50	34.9	45.6	70.0	27.5	58.0	43.7	18.7
BEVDepth* (Li et al. 2023b)	ResNet50	35.1	47.5	63.9	26.7	47.9	42.8	19.8
BEVStereo* (Li et al. 2023a)	ResNet50	37.2	50.0	59.8	27.0	43.8	36.7	19.0
FB-BEV* (Li et al. 2023c)	ResNet50	37.8	49.8	62.0	27.3	44.4	37.4	20.0
SA-BEV* (Zhang et al. 2023b)	ResNet50	38.7	51.2	61.3	26.6	35.2	38.2	19.9
PolarBEVU* (Hou et al. 2025)	ResNet50	40.1	49.7	56.1	28.5	59.4	38.3	21.3
SOLOFusion* (Park et al. 2023)	ResNet50	40.6	49.7	60.9	28.4	65.0	31.5	20.4
GeoBEV* (Zhang et al. 2025)	ResNet50	41.5	53.5	53.3	26.5	41.9	29.8	21.4
UniDistill+ (Zhou et al. 2023)	ResNet50	26.5	37.8	-	-	-	-	-
BEVDistill+ (Chen et al. 2023)	ResNet50	33.0	45.2	-	-	-	-	-
X_3KD+ (Klingner et al. 2023)	ResNet50	39.0	50.5	61.5	26.9	47.1	34.5	20.3
DistillBEV+ (Wang et al. 2023b)	ResNet50	40.3	51.0	62.3	26.6	46.4	35.7	20.7
VeXKD+ (Ji et al. 2024a)	ResNet50	41.2	47.7	-	-	-	-	-
FSD-BEV+ (Jiang et al. 2024)	ResNet50	41.2	53.8	52.7	25.6	36.3	33.0	20.7
LabelDistill+ (Kim et al. 2024)	ResNet50	41.9	52.8	58.2	25.8	41.3	34.6	22.0
BEVFusion-C(Liu et al. 2023c)	SwinT	35.6	41.2	66.8	27.3	56.1	89.6	25.9
SimDistill (Zhao et al. 2024a)	ResNet50	37.3	43.8	53.1	28.1	61.4	81.2	27.9
Ours	ResNet50	42.2	48.3	49.2	27.0	57.1	72.2	22.8
SimDistill (Zhao et al. 2024a)	SwinT	40.4	45.3	52.6	27.5	60.7	80.5	27.3
Ours	SwinT	43.4	50.4	48.3	27.2	**48.2**	67.7	21.0
Ours-BEVDepth	SwinT	**45.6**	**54.5**	**41.5**	**25.5**	51.0	**34.3**	**19.5**

Table 1: Comparison on the nuScenes validation set.* represents the non-distillation-based methods, + means the distillation-based methods, the others indicate the baseline model. The image size is 256×704 of all. And the temporal frame is only one.

CM A-SD	MM A-SD	HFM	S-WRD	mAP↑	NDS↑
				40.4	45.3
✓				41.5	46.3
	✓			40.9	46.9
		✓		41.2	46.3
			✓	41.9	47.0
✓	✓			41.8	47.9
✓	✓	✓		42.7	49.1
✓	✓	✓	✓	**43.4**	**50.4**

Table 2: Ablation Study on Different Components

value	mAP↑	NDS↑	mATE ↓
$\lambda = 0.1, \gamma = 0.5$	**41.8**	47.9	**51.1**
$\lambda = 0.2, \gamma = 0.5$	41.0	47.3	51.9
$\lambda = 0.3, \gamma = 0.5$	41.5	**48.0**	51.5
$\lambda = 0.4, \gamma = 0.5$	40.9	47.1	52.1
$\gamma = 0.5, \lambda = 0.1$	**41.9**	47.0	**50.4**
$\gamma = 0.6, \lambda = 0.1$	41.5	46.5	50.9
$\gamma = 0.7, \lambda = 0.1$	41.2	**47.2**	51.3

Table 3: The sensitive analysis of the λ and γ

ing the most increase in 1.5% mAP. The integration of all methodologies yields optimal performance, with mAP and NDS increasing by 3.0% and 5.1%, respectively.

Qualitative Results Figure 4 compares qualitative detection results between our method and baselines. The experimental visualization demonstrate that our method effectively reduces both missed and erroneous detections while improving the accuracy of the bounding boxes regression, thus validating the efficacy of the proposed approach.

Sensitive Analysis This paper primarily explores principles that extend beyond traditional alignment and fusion methods by refining loss and fusion strategies. Our findings indicate that parameter variations have a limited impact within certain thresholds. As shown in Table 3, varying λ does not significantly affect performance when it falls within the range of 0.1 to 0.4. The same applies to γ.

5 Conclusion

In this paper, we proposed a LiDAR-Camera distillation approach to improve the efficiency of knowledge transfer from multi-modal teacher models to student models. In summary, we introduced an adaptive-smooth distillation framework with heterogeneous fusion, designed to overcome the limitations of existing simulated point cloud generation methods, which often suffer from feature misalignment, and the simulated feature may degrade fused feature quality. Finally, we employed soft-weighted response distillation to prioritize high-quality teacher outputs over uniform knowledge transfer. Extensive experimentation on the challenging large-scale NuScenes benchmark substantiates the efficacy of our approach. In future work, we aim to enhance the quality of simulated LiDAR point clouds by using the cross-attention and contrastive distillation method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project number 62571336, Grant U24A20252), Shenzhen University 2035 Program for Excellent Research (Grant No. 2024C010), the Nansha Key Science and Technology Project 2023ZD006, and also was partially supported by Grant No. 2022B1212010001-OF08.

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11618–11628.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022. Graph-DETR3D: Rethinking Overlapping Regions for Multi-View 3D Object Detection. In *International Conference on Multimedia*, 5999–6008.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2023. BEVDistill: Cross-Modal BEV Distillation for Multi-View 3D Object Detection. In *Proceedings of the International Conference on Learning Representations*, 1–15.

Hong, Y.; Dai, H.; and Ding, Y. 2022. Cross-Modality Knowledge Distillation Network for Monocular 3D Object Detection. In *European Conference on Computer Vision*, volume 13662, 87–104.

Hou, M.; Lyu, C.; Wang, G.; Ma, B.; Xu, R.; Hu, J.; and Fan, X. 2025. PolarBEVU: Multi-Camera 3D Object Detection in Polar Bird's-Eye View via Unprojection. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–14.

Huang, J.; Huang, G.; Zhu, Z.; and Du, D. 2021. BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View. *CoRR*, abs/2112.11790: 1–12.

Huang, L.; Li, Z.; Sima, C.; Wang, W.; Wang, J.; Qiao, Y.; and Li, H. 2023. Leveraging Vision-Centric Multi-Modal Expertise for 3D Object Detection. In *Advances in Neural Information Processing Systems*, volume 36, 38504–38519.

Huang, L.; Wang, H.; Zeng, J.; Zhang, S.; Cao, L.; Yan, J.; and Li, H. 2025. LiDAR-guided Geometric Pretraining for Vision-Centric 3D Object Detection. *International Journal of Computer Vision*, 2(7): 1–14.

Jang, S.; Jo, D. U.; Hwang, S. J.; Lee, D.; and Ji, D. 2023. STXD: Structural and Temporal Cross-Modal Distillation for Multi-View 3D Object Detection. In *Advances in Neural Information Processing Systems*, volume 36, 29323–29342.

Ji, Y.; Chen, Y.; Yang, L.; Ding, R.; Yang, M.; and Zheng, X. 2024a. VeXKD: The Versatile Integration of Cross-Modal Fusion and Knowledge Distillation for 3D Perception. In *Advances in Neural Information Processing Systems*.

Ji, Y.; Chen, Y.; Yang, L.; Rui, D.; Yang, M.; and Zheng, X. 2024b. VeXKD: The Versatile Integration of Cross-Modal Fusion and Knowledge Distillation for 3D Perception. In *Advances in Neural Information Processing Systems*, volume 37, 125608–125634.

Jiang, Z.; Zhang, J.; Zhang, Y.; Liu, Q.; Hu, Z.; Wang, B.; and Wang, Y. 2024. FSD-BEV: Foreground Self-distillation for Multi-view 3D Object Detection. In *European Conference on Computer Vision*, volume 15066, 110–126.

Kim, S.; Kim, Y.; Hwang, S.; Jeong, H.; and Kum, D. 2024. LabelDistill: Label-Guided Cross-Modal Knowledge Distillation for Camera-Based 3D Object Detection. In *European Conference on Computer Vision*, 19–37.

Klingner, M.; Borse, S.; Kumar, V. R.; Rezaei, B.; Narayanan, V.; Yogamani, S. K.; and Porikli, F. 2023. X^3KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13343–13353.

Li, W.; Shao, S.; Qiu, Z.; and Song, A. 2024. Multi-perspective analysis on data augmentation in knowledge distillation. *Neurocomputing*, 583: 127516.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2023a. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1486–1494.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023b. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1477–1485.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision*, volume 13669, 1–18.

Li, Z.; Yu, Z.; Wang, W.; Anandkumar, A.; Lu, T.; and Álvarez, J. M. 2023c. FB-BEV: BEV Representation from Forward-Backward View Transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6896–6905.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2022. Sparse4D: Multi-view 3D Object Detection with Sparse Spatial-Temporal Fusion. *CoRR*, abs/2211.10581: 431–456.

Liu, H.; Teng, Y.; Lu, T.; Wang, H.; and Wang, L. 2023a. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18580–18590.

Liu, Y.; Ma, C.; and Kira, Z. 2022. Unbiased Teacher v2: Semi-supervised Object Detection for Anchor-free and Anchor-based Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9809–9818.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, volume 38, 531–548.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023b. PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3239–3249.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023c. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *Proceedings of the The International Conference on Robotics and Automation*, 2774–2781.

Ma, X.; Ouyang, W.; Simonelli, A.; and Ricci, E. 2023. 3d object detection from images for autonomous driving: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3537–3556.

Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8): 1909–1963.

Marvin, K.; Borse, S.; Kumar, V. R.; Rezaei, B.; Narayanan, V.; Yogamani, S.; and Porikli, F. 2023. X³KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition)*, 13343–13353.

Park, J.; Xu, C.; Yang, S.; Keutzer, K.; Kitani, K. M.; Tomizuka, M.; and Zhan, W. 2023. Time Will Tell: New Outlooks and A Baseline for Temporal Multi-View 3D Object Detection. In *Proceedings of the International Conference on Learning Representations*, 800–815.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, 194–210.

Qi, Z.; Wang, J.; Wu, X.; and Zhao, H. 2024. Ocbev: Object-centric bev transformer for multi-view 3d object detection. In *International Conference on 3D Vision*, 1188–1197.

Qian, R.; Lai, X.; and Li, X. 2022. 3D object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796.

Shen, C.; Huang, Y.; Zhu, H.; Fan, J.; and Zhang, G. 2024. Student-Oriented Teacher Knowledge Refinement for Knowledge Distillation. In *International Conference on Multimedia*, 4543–4552.

Song, Z.; Liu, L.; Jia, F.; Luo, Y.; Jia, C.; Zhang, G.; Yang, L.; and Wang, L. 2024. Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook. *IEEE Transactions on Intelligent Transportation Systems*, 25(11): 15407–15436.

Wang, C.; Qin, Y.; Kang, Z.; Ma, N.; and Zhang, R. 2024. Toward Accurate Camera-based 3D Object Detection via Cascade Depth Estimation and Calibration. In *Proceedings of the The International Conference on Robotics and Automation*, 2006–2012.

Wang, H.; Chen, X.; Yuan, Q.; and Liu, P. 2025. A review of 3D object detection based on autonomous driving. *Vis. Comput.*, 41(3): 1757–1775.

Wang, L.; Zhang, X.; Song, Z.; Bi, J.; Zhang, G.; Wei, H.; Tang, L.; Yang, L.; Li, J.; Jia, C.; et al. 2023a. Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. *IEEE Transactions on Intelligent Vehicles*, 8(7): 3781–3798.

Wang, Z.; Li, D.; Luo, C.; Xie, C.; and Yang, X. 2023b. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, volume 39, 8637–8646.

Wen, Z.; Xu, H.; Liu, C.; Guo, T.; Hu, J.; He, X.; Wang, F.; Lou, S.; and Fan, H. 2023. OccluBEV: Occlusion-Aware Spatiotemporal Modeling for Multi-View 3D Object Detection. In *International Conference on Multimedia*, 4074–4083.

Yang, L.; Tang, T.; Li, J.; Yuan, K.; Wu, K.; Chen, P.; Wang, L.; Huang, Y.; Li, L.; Zhang, X.; et al. 2025. BEVHeight++: Toward robust visual centric 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(2): 580–590.

Zhang, J.; Zhang, Y.; Liu, Q.; and Wang, Y. 2023a. Sa-bev: Generating semantic-aware bird's-eye-view feature for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3348–3357.

Zhang, J.; Zhang, Y.; Liu, Q.; and Wang, Y. 2023b. SA-BEV: Generating Semantic-Aware Bird's-Eye-View Feature for Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3325–3334.

Zhang, J.; Zhang, Y.; Qi, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2025. GeoBEV: Learning Geometric BEV Representation for Multi-view 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9960–9968.

Zhang, Y.; Hou, J.; and Yuan, Y. 2024. A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks. *International Journal of Computer Vision*, 132(5): 1592–1624.

Zhao, H.; Zhang, Q.; Zhao, S.; Chen, Z.; Zhang, J.; and Tao, D. 2024a. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7460–7468.

Zhao, X.; Zhang, X.; Yang, D.; Sun, M.; Li, M.; Wang, S.; and Zhang, L. 2024b. MaskBEV: Towards a Unified Framework for BEV Detection and Map Segmentation. In *International Conference on Multimedia*, 2652–2661.

Zhou, S.; Liu, W.; Hu, C.; Zhou, S.; and Ma, C. 2023. Unidistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5116–5125.

Zhu, Z.; Zhang, Y.; Chen, H.; Dong, Y.; Zhao, S.; Ding, W.; Zhong, J.; and Zheng, S. 2023. Understanding the robustness of 3D object detection with bird's-eye-view representations in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21600–21610.