Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.058238

ARTICLE

MMDistill: Multi-Modal BEV Distillation Framework for Multi-View 3D Object Detection

Tianzhe Jiao, Yuming Chen, Zhe Zhang, Chaopeng Guo and Jie Song*

Software College, Northeastern University, Shenyang, 110819, China

*Corresponding Author: Jie Song. Email: songjie@mail.neu.edu.cn

Received: 08 September 2024 Accepted: 11 November 2024 Published: 19 December 2024

ABSTRACT

Multi-modal 3D object detection has achieved remarkable progress, but it is often limited in practical industrial production because of its high cost and low efficiency. The multi-view camera-based method provides a feasible solution due to its low cost. However, camera data lacks geometric depth, and only using camera data to obtain high accuracy is challenging. This paper proposes a multi-modal Bird-Eye-View (BEV) distillation framework (MMDistill) to make a trade-off between them. MMDistill is a carefully crafted two-stage distillation framework based on teacher and student models for learning cross-modal knowledge and generating multi-modal features. It can improve the performance of unimodal detectors without introducing additional costs during inference. Specifically, our method can effectively solve the cross-gap caused by the heterogeneity between data. Furthermore, we further propose a Light Detection and Ranging (LiDAR)-guided geometric compensation module, which can assist the student model in obtaining effective geometric features and reduce the gap between different modalities. Our proposed method generally requires fewer computational resources and faster inference speed than traditional multi-modal models. This advancement enables multi-modal technology to be applied more widely in practical scenarios. Through experiments, we validate the effectiveness and superiority of MMDistill on the nuScenes dataset, achieving an improvement of 4.1% mean Average Precision (mAP) and 4.6% NuScenes Detection Score (NDS) over the baseline detector. In addition, we also present detailed ablation studies to validate our method.

KEYWORDS

3D object detection; multi-modal; knowledge distillation; deep learning; remote sensing

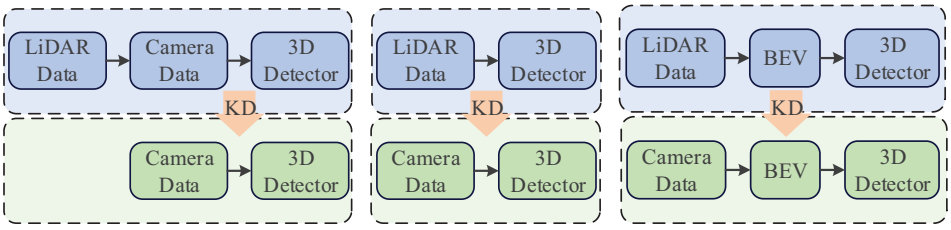
1 Introduction

3D object detection is an essential task in the computer vision field, used for identifying and locating objects of interest in images or video frames [1,2]. The multi-modal 3D object detection method combines data from sensors (such as LiDAR, camera, and radar) to improve detection accuracy and robustness. It provides a more comprehensive environmental perception capability, which can achieve more stable and reliable object detection in various complex environments. However, multi-modal methods typically require large amounts of computing resources (such as high-performance computing devices, software resources, and data resources) to process and fuse multi-modal data

for improving real-time performance and system responsiveness, which poses challenges for practical deployment with multi-modal methods [3].

The existing research attempts to design lightweight neural networks to reduce multi-modal models' parameter count and computational complexity. Common lightweight methods include Pruning, Quantization, Neural Architecture Search (NAS), Group Convolution, and Hybrid strategy [4]. The methods above mainly reduce model computational consumption by removing unimportant connections in the neural network and decreasing the accuracy of model parameters. However, lightweight networks must extract high-quality features to ensure the model's performance while maintaining low computational overhead. Consequently, lightweight models are often optimized for specific tasks and data, which results in poor universality and portability. In addition, recent work avoids the high costs of multi-modal models by using multi-view camera methods to predict depth information in place of LiDAR data to provide 3D geometric features [5]. Although this method has potential advantages in improving the completeness of spatial information, accurately inferring 3D geometric information from camera data alone remains challenging. This results in lower accuracy and limited ceiling performance compared to LiDAR-based methods.

Knowledge Distillation (KD) is a powerful method to improve model accuracy without increasing model complexity. This method effectively balances the high computational of multi-modal models with the low accuracy of unimodal models. KD utilizes a pre-trained complex teacher model to transfer its latent knowledge to a lightweight student model, enabling the latter to approach the former's performance while achieving higher computational efficiency. In light of this, recent works explore the projection of LiDAR points onto the image plane to serve as input for the teacher model [6]. With the same model architecture, as shown in Fig. 1a, the student model can naturally perform feature imitation with the teacher model. However, the conversion between modalities can lead to the loss of data features. It limits the effective information learned from the LiDAR-based teacher. Another work utilizes 3D information extracted from the voxel space to directly guide the 2D branch in cross-modal knowledge learning [7], as shown in Fig. 1b. However, directly performing knowledge distillation between different modalities often results in limited performance improvements due to significant modal heterogeneity. Therefore, to alleviate the issues above, we must propose an effective solution to solve cross-modal gaps.



(a) Simulated distillation (b) Cross-modal distillation (c) BEV distillation

Figure 1: Comparison of BEV distillation in our MMDistill with previous distillation frameworks

Cross-modal distillation primarily encounters two challenges: 1) Modal Heterogeneity. LiDAR provides sparse 3D point cloud data, whereas cameras capture dense 2D image data. These inherent differences in data representation make it difficult to align data directly. 2) Feature Disparity. LiDAR offers more precise spatial information, whereas cameras provide richer color and texture information. Thus, effectively performing cross-modal knowledge distillation across these two distinct modalities is challenging. This paper addresses the above challenges by designing a multi-modal knowledge

distillation framework, namely MMDistill. This framework contains a LiDAR-based teacher model and a camera-based student model. Unlike the methods above, we consistently map all features to the BEV space, preserving geometric depth and semantic information simultaneously, as shown in Fig. 1c. Features from different modalities can naturally align with minimal information loss using a shared BEV representation. In the BEV distillation process, we utilize soft target techniques to enhance the student model's ability to learn BEV features from the teacher model. Note that directly transferring knowledge from LiDAR features to image features is challenging. To address this challenge, we further introduce a LiDAR-guided geometric compensation module to assist student models in generating effective geometric features. The geometric compensation module consists of three components: dense feature extractor, context-awareness modular, and depth predictor. MMDistill primarily utilizes cameras, with LiDAR as a supplementary component, reducing reliance on expensive sensors and decreasing maintenance costs. Furthermore, this paper optimizes computation by utilizing lightweight feature extraction networks and high-performance convolution techniques, reducing computational costs. Through cross-modal distillation, we integrate multi-modal information into an unimodal model, replacing traditional multi-modal models that require various modalities as input. This way, we can effectively capture multi-modal features without increasing model inference time, enhancing model accuracy and achieving high cost-effectiveness. The main contributions of this paper are as follows:

- We propose a novel multi-modal distillation framework (MMDistill) for 3D object detection. It can integrate multi-modal information into an unimodal model to improve the performance of unimodal detectors without introducing additional costs during inference.
- We introduce a LiDAR-guided geometric compensation module to reduce the gap between different modalities and capture better geometric features to improve the accuracy of the camera-based detector.
- Through experiments and ablation studies, this paper validates the effectiveness and superiority of MMDistill on the nuScenes benchmark. Our best model reaches 56.3% NDS on the nuScenes dataset.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. Section 3 introduces our proposed MMDistill in detail. Experimental settings and results are presented in Section 4. Finally, Section 5 presents the conclusion of this paper.

2 Related Work

Multi-View Camera-Based 3D Object Detection. Multi-view camera-based methods achieve high-accuracy 3D object detection by fusing image data from various perspectives, offering more comprehensive scene coverage and significantly enhancing detection accuracy and robustness. Recent research utilizes multi-view camera parameters to generate geometric encoders as positional information to aid 3D object detection, effectively improving the accuracy and robustness of detection models [8]. Additionally, many works identify BEV representation as an ideal feature space for multi-view perception due to its capability to effectively resolve scale ambiguity and occlusion challenges. Qin et al. [9] introduce a unified BEV fusion method that minimizes information loss found in traditional methods, fully utilizing obtained features to improve 3D object detection accuracy.

Considering the critical role of depth in camera-based 3D detection, Li et al. [5] propose the BEVDepth, which introduces a camera-aware depth estimation module to enhance depth prediction capability through the joint optimization of depth estimation and multi-view features, thereby improving model efficiency while maintaining accuracy. Li et al. [10] further introduce spatial and temporal

information, proposing a novel framework called BEVFormer, which achieves more comprehensive and efficient 3D object detection by effectively integrating information from different time frames and perspectives. However, cameras are more affected by environmental factors than LiDAR, especially in low-light and high-glare scenarios, where they may fail to capture sufficient image information, making effective recognition of the surrounding environment challenging. Additionally, as the distance to the object increases, the number of pixels representing the object in the image decreases, and the error in depth estimation correspondingly increases, which can affect the accuracy and reliability of detection. Due to the lack of accurate depth information, camera-based methods still have a significant performance gap compared to their counterparts based on LiDAR and multi-modal fusion [1].

Fusion-Based 3D Object Detection. 3D object detection based on multi-modal fusion has enormous potential, and there has been a significant increase in the number of studies related to multi-modal research recently. This method effectively integrates data from various scenarios, especially cameras and LiDAR, significantly enhancing 3D object detection performance in complex environments, thereby becoming mainstream research. It has been demonstrated that multi-modal learning can more accurately represent potential space than unimodal learning [11]. An early and representative method is Multi-View 3D (MV3D) [12], which systematically combines point clouds and images for 3D object detection. Similarly, Liu et al. [13] propose BEVFusion, which unifies multi-view camera data and point cloud into a BEV representation, excelling in handling distant and occluded objects in complex scenarios. In order to improve the responsiveness of the perception system to environmental changes and dynamic objects, Chen et al. [14] further introduce FUTR3D, which is based on the Transformer architecture and integrates multi-modal features through the Future Cross-Attention mechanism.

Current mainstream fusion-based methods can be divided into input-level and feature-level methods categorized by the stage of data fusion [15]. These methods capture more comprehensive and detailed features, often providing higher detection accuracy and robustness than decision-level fusion methods. However, multi-modal fusion methods typically require the design of complex neural network models, increasing system development complexity. It can become a bottleneck in resource-constrained environments. Therefore, weighing the advantages and disadvantages of multi-modal fusion methods and making selections based on specific requirements is essential.

Knowledge Distillation-Based 3D Object Detection. Knowledge distillation is widely used in various fields, including computer vision and natural language processing, to transfer knowledge learned from large and complex models (teacher models) to smaller and more efficient models (student models), making it suitable for resource-constrained environments or real-time applications. Distillation methods for object detection can be categorized into global feature distillation and local feature distillation. Global feature extraction methods rely entirely on the teacher model, where the student model mimics the full feature maps from the intermediate layers of the teacher model [16]. Local feature distillation methods aim to guide the student model in learning specific local features that are more important or beneficial for the detection task, such as cars and people [17]. Additionally, Yang et al. [18] combine global and local feature distillation methods to construct a feature extractor, further optimizing the distillation process for object detection.

Recent advances attempt to utilize teacher models from other modalities to guide student model training. Wang et al. [19] use a multi-view camera-based student detector to mimic features of a pre-trained LiDAR-based teacher detector, enhancing its representation learning capability and generalizing knowledge transfer to multi-scale layers with temporal fusion. Similarly, Zhao et al. [20] utilize camera features to simulate LiDAR features and then employ a multi-modal model to achieve cross-modal distillation. Unlike the methods above, Klingner et al. [21] propose a comprehensive

knowledge distillation framework across different modalities, tasks, and stages by employing multiple teacher models, optimizing the accuracy of the student detector from multiple perspectives. Although these cross-modal distillation techniques highlight the potential of transferring knowledge from robust LiDAR teachers to camera-based students, most existing methods are limited to specific tasks, and cross-modal gaps still need to be addressed. Thus, we design a unified multi-modal distillation framework specifically aimed at overcoming the problem of cross-modal gaps.

3 Methodology

In this section, we introduce our proposed MMDistill in detail. We first illustrate the general overview of the framework in Fig. 2, which utilizes a more powerful LiDAR-based teacher model to guide the multi-view camera-based student model for learning multi-modal features. MMDistill consists of three modules: LiDAR-guided geometric compensation, BEV feature-based distillation, and response-based distillation. MMDistill follows the common knowledge distillation paradigm and has two steps. In the first step, we design a LiDAR-guided geometric compensation module to facilitate depth prediction capability by utilizing explicit depth supervision. The second step proposes a two-stage distillation framework to resolve cross-modal gaps, learning useful features from the LiDAR-based method. We will introduce each step in the following sections.

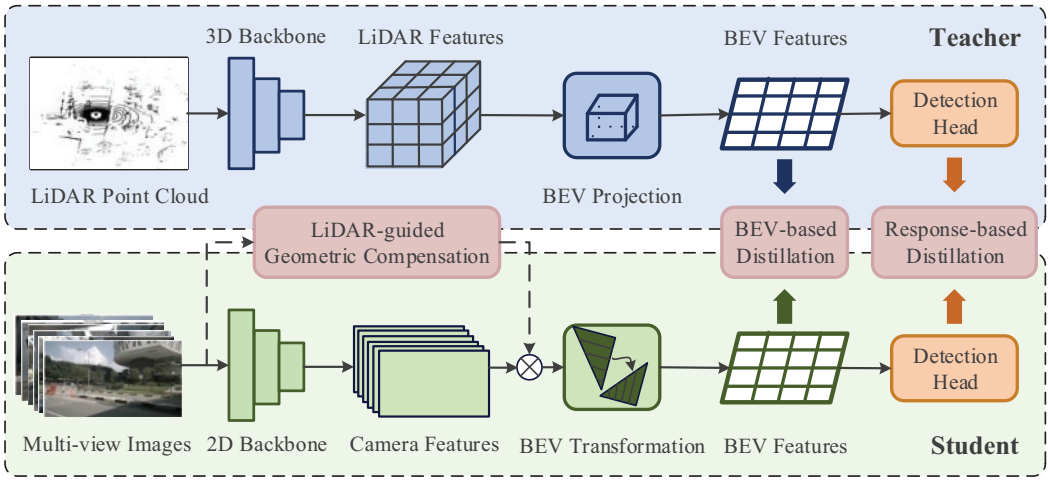


Figure 2: Framework of MMDistill

3.1 LiDAR-Guided Geometric Compensation

For the student, we adopt multi-view camera-based models, which fundamentally rely on visual data and can be inadequate in accurately capturing depth and spatial dimensions. Thus, camera-based 3D object detection systems often face challenges in accurately discerning object distances and shapes, potentially leading to less reliable performance in specific scenarios, such as low lighting or adverse weather. To this end, we design a LiDAR-guided geometric compensation module for the student model to address the challenges above.

Algorithm 1: Single point cloud to depth image conversion

Input: Point cloud P, Image height h, Image width w, Rotation matrix R, Rotation vector r, Translation vector t, Camera intrinsic matrix K, Distortion coefficients d
Output: Depth image D_l

Initialize an empty image array D_l of size (h, w) with dtype uint16
Initialize counters *n_point* and *n_proj* to 0
For each point in P:
    If the y-coordinate of the point is greater than 0:
        Convert the point to a numpy array q
        Project the 3D point to 2D using cv2.projectPoints with r, t, K, and d
        Squeeze the projected point to remove extra dimensions, resulting in *p_2D*
        Compute the transformed coordinates *p_3D* using the rotation matrix R and q
        Add the translation vector t to *p_3D* to get *p_3D_trans*
        Compute the depth value z as the third coordinate of *p_3D_trans* multiplied by 256
        Compute the pixel coordinates u and v from *p_2D*
        If u and v are within the image dimensions ($0 <= u < w$ and $0 <= v < h$):
            Increment *n_proj*
            If the pixel at (v, u) in D_l is 0:
                Set the pixel value at (v, u) in D_l to z
                Increment *n_point*
            ENDIF
    ENDIF
END
Return the depth image D_l

Depth Supervision. As shown in Fig. 2, MMDistill utilizes explicit depth supervision to train the geometric compensation module. Due to the difficulty of monocular depth estimation, we transform point cloud data P into ground-truth D_l to supervise the depth prediction D_p. Algorithm 1 outlines a process for converting point cloud P to a depth map. It starts by initializing an empty depth map array D_l and some counters. For each point in the point cloud, if its y-coordinate is positive, it is converted into a NumPy array and projected from 3D to 2D using OpenCV's projectPoints function. The point is then transformed with a rotation matrix and translation vector to compute its depth value. The 2D coordinates are checked to judge whether they fall within the image dimensions. When the pixel position has not been set in the depth map, the depth value is assigned to the pixel. Finally, the algorithm concludes by returning the completed depth map D_l.

Geometric Compensation Module. The geometric compensation module consists of three components: a dense feature extractor, context-awareness modular, and depth predictor, as shown in Fig. 3. In the feature extraction stage, we apply standard deep convolutional neural networks as the feature extractor. Note that the number of max-pooling layers and the value of striding can affect the spatial resolution of the feature maps. Thus, we combine traditional convolution with dilated convolution as the basic convolution module of deep convolutional neural networks. In this way, we can enlarge the receptive field of filters and keep spatial resolution without increasing the number of parameters.

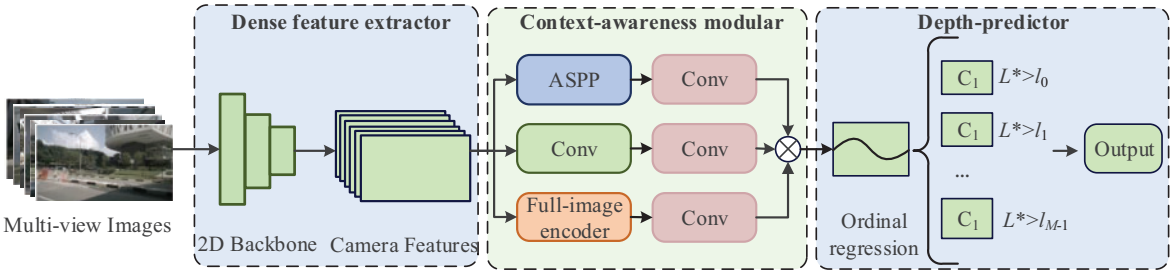


Figure 3: The architecture of LiDAR-guided geometric compensation module

We initially utilize the transformed depth maps as ground truth labels to perform supervised pre-training on the LiDAR-guided module. Subsequently, we integrate the pre-trained LiDAR-guided module into the student model as depicted in Fig. 2. Specifically, a set of multi-view images is processed concurrently through a 2D backbone network to extract camera feature information and through the LiDAR-guided module to capture depth information. The camera feature and depth information are then fused along the channel dimension. The fused feature information is input into the BEV transformation to complete feature mapping. During the subsequent two-stage distillation, the LiDAR-guided module is treated as a part of the student model for weight updates, thereby addressing the challenge of directly transferring knowledge from LiDAR features to image features and ensuring effective knowledge transfer.

In context-awareness modular, we utilize an atrous spatial pyramid pooling (ASPP) to extract features from multiple receptive fields with different sizes. The dilation rates of dilated convolutional are 6, 12, and 18, respectively. Next, the second branch is a convolutional layer with a kernel size of 1×1, which can learn complex cross-channel interactions. The third branch is a full-image encoder, which captures global contextual information. Finally, we quantize the depth interval into discrete depth values. This way, we can directly transform the regression problem into a classification problem and utilize an ordinal loss to learn weight. Eq. (1) describes the ordinal loss $L(\mathrm{X}, \Theta)$.

$$L(\mathrm{X}, \Theta) = -\frac{1}{w \times h} \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} \Phi(i, j, \mathrm{X}, \Theta) \tag{1}$$

where X denotes the feature map. w and h are the width and height of the feature map. $\Theta = (\theta_0, \theta_1, ..., \theta_{2M-1})$ is the weight vectors. $\Phi(i, j, \mathrm{X}, \Theta)$ is the pixel-wise ordinal loss, as shown in Eq. (2).

$$\Phi(i, j, \mathrm{X}, \Theta) = \sum_{m=0}^{l_{(i,j)}-1} \log\left(Q^m_{(i,j)}\right) + \sum_{m=l_{(i,j)}}^{M-1} \left(1 - \log\left(Q^m_{(i,j)}\right)\right) \tag{2}$$

$$Q^m_{(i,j)} = P\left(\hat{l}_{(i,j)} > m | \mathrm{X}, \Theta\right) \tag{3}$$

where $l_{(i,j)} \in \{0, 1, \ldots, M-1\}$ is the discrete label at spatial location (i, j). $\hat{l}_{(i,j)}$ is the estimated discrete value at spatial location (i, j).

3.2 Multi-Modal Distillation

Compared to traditional multi-modal models, multi-modal distillation models generally require fewer computational resources for inference, which reduces deployment costs and enhances inference speed. This advancement enables multi-modal technology to be applied more quickly and widely

in practical scenarios. In order to address the cross-modal gaps, we design a universal two-stage distillation framework, which consists of two parts: BEV feature-based distillation and response-based distillation.

BEV Feature-Based Distillation. In order to conduct the feature distillation, we keep the same architecture for student and teacher models, unifying the image and LiDAR features to the BEV space, as shown in Fig. 2. In this way, the shared BEV representation can naturally align the features from different modalities without much information loss. In conventional training, a student network typically learns by minimizing the cross-entropy loss against true labels. However, these hard targets may overlook crucial information, such as the relative similarity between classes. Furthermore, this strategy may not work well under the cross-modal feature conditions, as there can still be inherent differences between the modalities.

In response to these issues, we adopt a soft targets technique to enhance the student model's ability to learn BEV features from the teacher model. Applying temperature scaling to the model's output BEV features achieves a smoother probability distribution. Temperature scaling distillation can improve the student model's performance by integrating soft targets with true labels, particularly in complex tasks and scenarios. Eq. (4) is the temperature scaling formula.

$$q_i = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)} \tag{4}$$

where z represents the unscaled output of the model, referring to the BEV features in this paper. q denotes the probability distribution obtained after temperature scaling. i and j represent the channel indices of the BEV feature map, respectively. T is the temperature parameter. When $T > 1$, the resulting distribution becomes smoother. At $T = 1$, the distribution remains a standard probability distribution. When $T < 1$, the distribution becomes sharper.

In practical operation, we first train a high-performance LiDAR-based teacher model to obtain the BEV features of LiDAR data. Then, these BEV features are transformed into soft targets by applying the temperature scaling formula. Subsequently, we define the student model and design the loss function. The total loss L_F is defined using the cross-entropy loss L_{CE} with true labels and the Kullback-Leibler divergence (KL divergence) loss L_{KD} with temperature scaling.

$$L_F = \alpha \cdot L_{CE}\left(y_t, F_s\right) + (1-\alpha) \cdot T^2 \cdot L_{KD}\left(F_t, F_s\right) \tag{5}$$

$$L_{CE} = -\sum_i y_t \cdot \log\left(F_s^i\right) = -\sum_i y_t \cdot \log\left(\frac{\exp\left(y_s^i/T\right)}{\sum_j \exp\left(y_s^j/T\right)}\right) \tag{6}$$

$$\begin{aligned} L_{KD} &= \sum_i F_t^i \cdot \log\left(\frac{F_t^i}{F_s^i}\right) \\ &= \sum_i \frac{\exp\left(y_t^i/T\right)}{\sum_j \exp\left(y_t^j/T\right)} \cdot \log\left(\frac{\exp\left(y_t^i/T\right) \cdot \sum_j \exp\left(y_s^j/T\right)}{\sum_j \exp\left(y_t^j/T\right) \cdot \exp\left(y_s^i/T\right)}\right) \end{aligned} \tag{7}$$

where α is the hyper-parameter for balancing the two loss components. y_s and y_t represent the unscaled BEV features from the student and teacher models, respectively. F_s and F_t denote the outputs of the temperature-scaled BEV features from the student and teacher networks, respectively. i and j represent the channel indices of the BEV feature map, respectively. The selection of the temperature parameter T is crucial, as it controls the smoothness of the output. In most studies, the temperature range is between 2 and 10. Therefore, we perform a grid search within this predefined temperature range to

select the suited T. The student model is trained sequentially, observing the impact of different T on performance, ultimately setting T equal to 3.

Response-Based Distillation. Response-based distillation is a head network distillation that enables the student network to achieve comparable performance to the teacher network in classification and localization tasks through effective knowledge transfer. We distill knowledge at the head network by calculating the loss between the outputs of the teacher and the student. We combine KL divergence and Smooth L1 loss to achieve head network distillation. This design allows the student network to gradually absorb the teacher network's capabilities in category differentiation and spatial position prediction, facilitating the efficient transfer of cross-modal knowledge and enabling the student network to achieve optimal performance.

The distillation process in the classification task is implemented using KL divergence, which aims to minimize the difference in probability distributions between the teacher and student networks. L_{class} is the loss of classification task, as shown in Eq. (8).

$$L_{class} = \frac{1}{\mathrm{N}} \sum_{i} \sum_{c} p_t^i(c) \log \left(\frac{p_t^i(c)}{p_s^i(c)} \right) \tag{8}$$

where $p_t^i(c)$ and $p_s^i(c)$ represent the probability outputs for category c of the teacher and student networks for the i-th sample, respectively. N is the total number of samples. By optimizing this loss function, we expect the student network to more accurately reflect the teacher network's weighting in category decisions, thereby providing precise and reliable classification results in complex environments. For the regression task in 3D object detection, we employ Smooth L1 loss to measure the similarity in bounding box coordinates between the teacher and student networks. Smooth L1 loss provides a robust measurement method to handle errors in bounding box prediction. It ensures that student networks' spatial positioning accuracy gradually approaches teacher networks, thereby achieving effective knowledge transfer in spatial positioning, as shown in Eq. (9).

$$L_{reg} = \frac{1}{\mathrm{M}} \sum_{i} \sum_{j} SmoothL1\left(b_t^{i,j}, b_s^{i,j}\right) \tag{9}$$

where $b_t^{i,j}$ and $b_s^{i,j}$ are the bounding box coordinates from the teacher and student networks, respectively. Here, i is used to identify different samples within the dataset, while j is used to identify different coordinates within each sample. M is the total number of these coordinates. Smooth L1 loss is robust to outliers, making it suitable for handling large errors while providing fine granularity when errors are small, ensuring that the student network's spatial localization accuracy gradually approximates that of the teacher network. To achieve a balance between classification and regression tasks, we define the final distillation loss by a weighted combination:

$$L_{distill} = \lambda_c L_{class} + \lambda_r L_{reg} \tag{10}$$

where λ_c and λ_r are the weighting parameters for balancing classification and regression distillation. By systematically adjusting these weights, optimal overall performance can be achieved in different application scenarios.

4 Experiments

In this section, we first introduce the experimental settings and the baseline modal. Next, we show the main results for comparing the state-of-the-art methods on the nuScenes dataset. We compare the performance of our method with multi-modal and unimodal 3D object detection methods and discuss

the model's efficiency. In order to verify the effectiveness of each component, we conduct the ablation study and related analyses.

4.1 Setup

Baseline Modal. We select BEVFormer [10] as the baseline model, specifically designed for autonomous driving applications. BEVFormer can efficiently extract complex spatial features using the transformer architecture and supports multitasking while providing real-time performance and robust reliability. To enhance the detector's accuracy, we replace its backbone network with EfficientNet [22]. As shown in Table 1, EfficientNet-B3 outperforms other backbone networks in both mAP and NDS, achieving 0.423 and 0.525, respectively. Despite having fewer parameters, it significantly improves performance over ResNet-101 and ResNeXt-101, reducing parameters by 72 M while increasing accuracy by 2%. This efficiency is attributed to the well-balanced architecture of EfficientNet-B3.

Table 1: Evaluation of backbone networks used in student model

Model	#Params	mAP↑	NDS↑
ResNet-101	44.5 M	0.396	0.481
ResNeXt-101	84 M	0.421	0.518
DenseNet-264	14 M	0.406	0.509
EfficientNet-B3 (Ours)	12 M	0.423	0.525

Our teacher model is primarily based on Object-Dynamic Graph CNN (Object-DGCNN) [23], a LiDAR-based detector renowned for its spatial feature extraction capabilities. This model uses convolutional neural networks to extract features on a grid and employs PointPillars [24] to map sparse point clouds to the BEV space, which is crucial for providing comprehensive geometric information. We directly utilize the optimal model from the nuScenes dataset provided in Reference [23] as a reliable source of knowledge for the student model. Knowledge transfer between the teacher and student models is achieved through BEV feature distillation and response distillation. Specifically, we align the features of different modalities via BEV feature distillation to minimize information loss and employ response distillation to enable the student model to mimic the output of the teacher model better, thereby improving detection accuracy.

Dataset and Evaluation Metrics. We utilize the nuScenes dataset [25] to evaluate our distillation framework, one of the most popular datasets for 3D object detection. The nuScenes has a fully autonomous vehicle sensor suite, including six cameras, five radars, and one LiDAR, establishing a one-to-one correspondence between each 3D point and the 2D image plane. It comprises 700 scenes for training, 150 for validation, and 150 for testing, including 3D bounding boxes for ten classes: Car, Truck, Bus, Trailer, Construction vehicle, Pedestrian, Motorcycle, Bicycle, Traffic cone, and Barrier. We adopt the official evaluation metrics to report our work. Mean Average Precision (mAP) and NuScenes Detection Score (NDS) are our main evaluation metrics. We also use other officially released true positive (TP) metrics in our experiments, including mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE).

$$\text{NDS} = \frac{1}{10}\left(5 \times mAP + \sum_{TP \in \mathbf{TP}} (1 - \min(1, TP))\right) \tag{11}$$

Implementation Details. We construct our codebase based on MMDetection3D[1], implemented with PyTorch using 8 NVIDIA A100 GPUs. Our model takes images with a size of 900 × 1600 and LiDAR point cloud with a voxel size of (0.1, 0.1, 0.2 m) as input. The teacher model utilizes PointPillars to facilitate efficient feature extraction, whereas the student model utilizes EfficientNet as its backbone network. The default values of the hyper-parameters are set as $\{\alpha, \lambda_c, \lambda_r\} = \{0.5, 1, 1\}$. In the distillation phase, the batch size is set to 1 per Graphic Processing Unit (GPU), employing AdamW [26] as the optimizer with an initial learning rate 2e-4. We conduct training for 24 epochs utilizing a cyclic policy, achieving the optimal model. Within BEVFormer, the grid size of the BEV plane is set to 128 × 128. We adopt ResNet101 as the feature extractor within the geometric compensation module.

During dataset processing, we use the dataset processing scripts provided by the open-source framework MMDetection3D. With the help of the converter module in MMDetection3D, we read the metadata and sensor data of the dataset, including point cloud and image data. We perform coordinate system transformations, filtering, and downsampling for point cloud data to ensure data quality and consistency. Image data undergo cropping, scaling, and color space transformation to meet the input requirements of the model. Annotation data are converted to the target format, involving coordinate transformation and category label mapping. All processed data are stored in a specified output directory to facilitate subsequent model training and evaluation.

4.2 State-of-the-Art Comparisons

In our experiments, Table 2 compares the performance of various state-of-the-art methods on the nuScenes validation set. Our model MMDistill demonstrates outstanding performance across multiple key metrics, with a score of 0.457 on mAP, achieving the best results. Compared to our baseline method BEVFormer [10], MMDistill improves mAP by 4.1%. This improvement is mainly attributed to our proposed multi-modal distillation framework, which effectively combines the advantages of LiDAR and camera data, enabling the student model to learn richer geometric information without increasing inference costs. By unifying all features into the BEV space, we can retain geometric depth and semantic information simultaneously, leading to effective cross-modal knowledge fusion. Additionally, MMDistill achieves 0.563 on the NDS, demonstrating its superiority in overall detection performance.

Table 2: Comparison of SOTA works on the nuScenes validation set

Methods	Backbone	Image Size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet-Base [27]	SwinT	640 × 1600	0.393	0.472	0.608	**0.259**	0.366	0.822	0.191
FCOS3D [28]	ResNet-101	900 × 1600	0.343	0.415	0.725	0.263	0.422	1.292	**0.153**
BEVDepth [5]	ResNet-101	512 × 1408	0.412	0.535	0.565	0.266	0.358	0.331	0.190
STS [29]	ResNet-101	512 × 1408	0.431	0.542	0.525	0.262	0.380	0.369	0.204
DETR3D [30]	ResNet-101	900 × 1600	0.349	0.434	0.716	0.268	0.379	0.842	0.200
PETR [31]	ResNet-101	512 × 1408	0.357	0.421	0.710	0.270	0.490	0.885	0.224
UVTR-L2CS [7]	ResNet-101	256 × 704	0.392	0.488	0.720	0.268	**0.354**	0.534	0.206
BEVFormer [10]	ResNet-101	900 × 1600	0.416	0.517	0.673	0.274	0.372	0.394	0.198
TiG-BEV [32]	ResNet-101	512 × 1408	0.440	0.544	0.570	0.267	0.392	0.331	0.201
MMDistill	ResNet-101	900 × 1600	0.451	0.559	0.512	0.269	0.368	0.330	0.187
MMDistill	EfficientNet	900 × 1600	**0.457**	**0.563**	**0.497**	0.260	0.374	**0.327**	0.196

[1] "MMDetection3D." Available: https://github.com/open-mmlab/mmdetection3d, accessed on 09 June 2024.

The geometric compensation module aids the student model in better focusing on critical information in the environment, thereby extracting effective geometric features. The introduction of this module effectively reduces the gap between modalities, facilitating smoother knowledge transfer from LiDAR features to image features. In contrast, other methods, such as BEVDet-Base [27], perform well in certain metrics but still fall short of MMDistill in overall performance. Through novel framework design, MMDistill achieves performance improvement in multi-view 3D object detection tasks, demonstrating its potential and advantages in practical applications.

In order to comprehensively evaluate the robustness and generalization ability of our method, we further conduct experiments with MMDistill on the public KITTI dataset [33]. We aim to explore our method's performance in different environments and conditions to verify its ability to adapt to various challenges. As shown in Table 3, MMDistill demonstrates excellent performance in the car and pedestrian tasks compared to other methods, achieving optimal results. In hard-level tasks, it achieves detection accuracies of 77.85%, 45.97%, and 61.76% for the car, pedestrian, and cyclist categories, respectively. These experimental results indicate that our method is adaptable and robust in other scenarios.

Table 3: Comparison of 3D detection results on the KITTI test set. We evaluate our model for all three classes using the 3D AP under 40 recall thresholds (R40). "L" represents LiDAR and "C" represents Red, Green, Blue (RGB)

Methods	Modality	Car			Pedestrian			Cyclist		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PV-RCNN++ [34]	L	87.72	81.29	76.78	47.50	40.31	38.15	80.34	67.46	60.38
SIEV-Net [35]	L	85.21	76.18	70.60	54.00	44.80	41.11	78.75	59.99	52.37
HMFI [36]	L+C	88.90	81.93	77.30	50.88	42.65	39.78	**84.02**	**70.37**	**62.57**
3D-CVF [37]	L+C	89.20	80.05	73.11	–	–	–	–	–	–
VoPiFNet [38]	L+C	88.81	80.97	76.74	54.65	48.36	44.98	77.64	64.10	58.00
MMDistill	C	**91.07**	**82.74**	**77.85**	**55.76**	**49.51**	**45.78**	82.98	69.57	61.76

As shown in Table 4, MMDistill demonstrates superior efficiency and performance, showing significant advantages in computational resource requirements and inference speed. The model optimizes computational efficiency through a two-stage distillation framework, integrating multi-modal information into an unimodal model, enhancing the student model's learning capacity without increasing inference costs. MMDistill achieves 9.2 Frames Per Second (FPS), exhibiting a notable real-time performance advantage over other methods, and contains only 52.2 M parameters. MMDistill demonstrates performance advantages across multiple dimensions. MMDistill shows lower computational complexity during inference, benefiting from using only the student model with a lightweight feature extractor. Unlike traditional multi-modal methods that often employ multi-branch structures, MMDistill's design helps reduce the consumption of computational resources. In addition, by cleverly integrating depthwise separable convolutions with conventional convolutions in the feature extraction part of the depth compensation module, the model achieves effective feature compensation while maintaining a low computational cost. The full-image encoding method adopted by MMDistill results in a lower parameter count than traditional methods, further enhancing computational efficiency. Compared with multi-modal fusion methods such as Painted PointPillars and3D-CVF, MMDistill shows outstanding advantages, exceeding them by more than three times in FPS and significantly

reducing parameter count. MMDistill effectively reduces dependency on expensive and computation-intensive LiDAR data through this lightweight architectural design combined with learning guidance from a LiDAR-based teacher model. This enables MMDistill to maintain competitive detection accuracy and demonstrate strong deployment potential in real-world industrial applications where cost and speed are critical.

Table 4: Comparison of traditional multi-modal fusion methods on the nuScenes validation set. "L" represents LiDAR, "R" represents Radar, and "C" represents RGB

Methods	Modalities	FPS	#Params	mAP	NDS
Painted PointPillars+ [39]	L+C	2.7	79.8M	0.464	0.581
3D-CVF [37]	L+C	4.9	71.6M	0.421	0.497
RCM-Fusion [40]	R+C	4.3	87.6M	0.443	0.529
HyDRa [41]	R+C	6.3	90.3M	0.536	0.617
MMDistill	C	9.2	52.2M	0.457	0.563

4.3 Ablation Study

In this section, we evaluate the effectiveness of each component within the 3D object detection framework through the ablation study. We design three components to transfer reliable information from the teacher model: geometric compensation, BEV distillation, and response distillation. Each component individually shows significant improvements in mAP and NDS compared to the baseline. As shown in Table 5, the baseline model achieves an mAP of 0.423 and an NDS of 0.525. Upon integrating the geometric compensation module, mAP and NDS increase to 0.432 and 0.533, respectively. This improvement is attributed to the module's effective reduction of the modality gap, enabling the camera-based student model to capture more accurate geometric features. Incorporating BEV distillation into the framework significantly enhances performance, with mAP increasing by 20% and NDS by 16%. This improvement underscores the effectiveness of BEV distillation in transferring multi-modal knowledge, thereby enriching the feature representation of the student model. MMDistill also combines response distillation to achieve optimal performance, reaching an mAP of 0.457 and an NDS of 0.563. The addition of response distillation further enhances the student model's capacity to mimic the teacher model's output, resulting in more accurate object detection.

Table 5: Ablation study of each component on the nuScenes validation set

Baseline	Geometric compensation	BEV distillation	Response distillation	mAP↑	NDS↑
√				0.423	0.525
√	√			0.432	0.533
√	√	√		0.452	0.549
√	√	√	√	0.457	0.563

The baseline detector primarily relies on single-modality data input. It lacks the integration of multi-modal information, which results in poor performance when addressing issues like scale ambiguity and occlusion, leading to decreased accuracy. MMDistill introduces several innovative mechanisms

to enhance detection performance and overcome the baseline detector's limitations. Specifically, MMDistill employs a LiDAR-guided geometric compensation module to improve the model's depth prediction capability. This module effectively reduces the gap between different modalities, enabling the camera-based student model to capture more precise geometric features. Additionally, MMDistill integrates multi-modal information into an unimodal model through the designed BEV feature distillation and response distillation methods, replacing traditional multi-modal models. This method allows for improved model detection accuracy without increasing inference time.

In the LiDAR-guided geometric compensation module, full-image encoders are vital in enhancing performance. We evaluate the impact of different image encoding methods, as shown in Table 6. Initially, without a full-image encoder, our metrics were mAP at 0.451 and NDS at 0.554. After adopting the fc-fashion method, the performance improved, with mAP and NDS increasing to 0.455 and 0.559, respectively. However, this method introduces a large number of parameters, complicating training and increasing memory usage. To address this, we introduce an average pooling layer to reduce the spatial dimension, followed by a fully connected layer to obtain feature vectors of dimension c. We then use a 1×1 convolutional layer as a cross-channel parameter pooling structure. Finally, the feature vectors were replicated across the spatial dimension to maintain consistent image comprehension across all positions. This method improves all performance metrics, raising mAP and NDS to 0.457 and 0.563, respectively, demonstrating higher detection accuracy and significantly reducing errors. These results indicate that our method improves image encoding by more effectively capturing image features, thus enhancing overall performance.

Table 6: Ablation study of different image encoder

Methods	Backbone	Image Size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
W/O Full-image encoder	EfficientNet	900 × 1600	0.451	0.554	0.511	0.269	0.387	0.342	0.207
Fc-fashion	EfficientNet	900 × 1600	0.455	0.559	0.507	0.264	0.381	0.334	0.203
Ours	EfficientNet	900 × 1600	0.457	0.563	0.497	0.260	0.374	0.327	0.196

In addition, to verify the impact of the loss function on model performance, we validate the performance differences caused by using different training constraint strategies during the BEV feature-based distillation stage. Table 7 shows the performance impact of using these different training constraint strategies on the model. In the BEV feature-based distillation stage, the choice of training constraint strategies significantly influences model performance. The traditional constraint strategy uses the teacher model's output as a true label to guide training, which can lead to reduced model robustness due to over-reliance on the teacher model's output. On the other hand, training solely with temperature-scaled outputs may lead to the student model being unable to learn the teacher model's complete knowledge thoroughly. Thus, our proposed method ingeniously combines traditional and temperature scaling strategies, enabling the model to absorb the teacher model's knowledge while maintaining robustness during training. This strategy effectively addresses the limitations of both traditional and scaling strategies, enhancing model performance in object detection tasks. Experimental results demonstrate that our method excels across various evaluation metrics, confirming its effectiveness and superiority.

The experimental results show the performance gains attributed to each component of MMDistill. Integrating the geometric compensation module and BEV distillation can significantly enhance the student's learning capability from the teacher model, effectively narrowing the cross-modal gap. This results in a more robust and accurate camera-based 3D object detection system. The primary advantage of MMDistill is its ability to improve detection accuracy without requiring additional inference costs. By utilizing knowledge distillation techniques, our framework successfully integrates multi-modal information into an unimodal model, providing a cost-effective solution for 3D object detection in practical applications.

Table 7: Ablation study of different training constraint strategies. "L_T" indicates using only the temperature scaling strategy, while "L_N" means not using it

Loss functions	Backbone	Image size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
L_T	EfficientNet	900 × 1600	0.453	0.559	0.501	0.262	0.377	0.331	0.201
L_N	EfficientNet	900 × 1600	0.446	0.554	0.505	0.265	0.380	0.335	0.207
Ours	EfficientNet	900 × 1600	0.457	0.563	0.497	0.260	0.374	0.327	0.196

4.4 Qualitative Results

Figs. 4 and 5 show the visualization results of our proposed model MMDistill and baseline model BEVFormer in 3D object detection tasks. In the figures, it is clear that our method can detect instances that the baseline model fails to identify. Furthermore, Fig. 5 shows the detection results of five scenes from the BEV view. Our model significantly improves the predictive ability of the rotation angle of the bounding box for an object, which means more accurate object localization and direction recognition. These instances qualitatively validate the effectiveness of our method in 3D object detection tasks.

Figure 4: (Continued)

Figure 4: Comparison of detection results between baseline and our method on nuScences dataset. Green arrows mark the objects missed by the baseline but detected by MMDistill

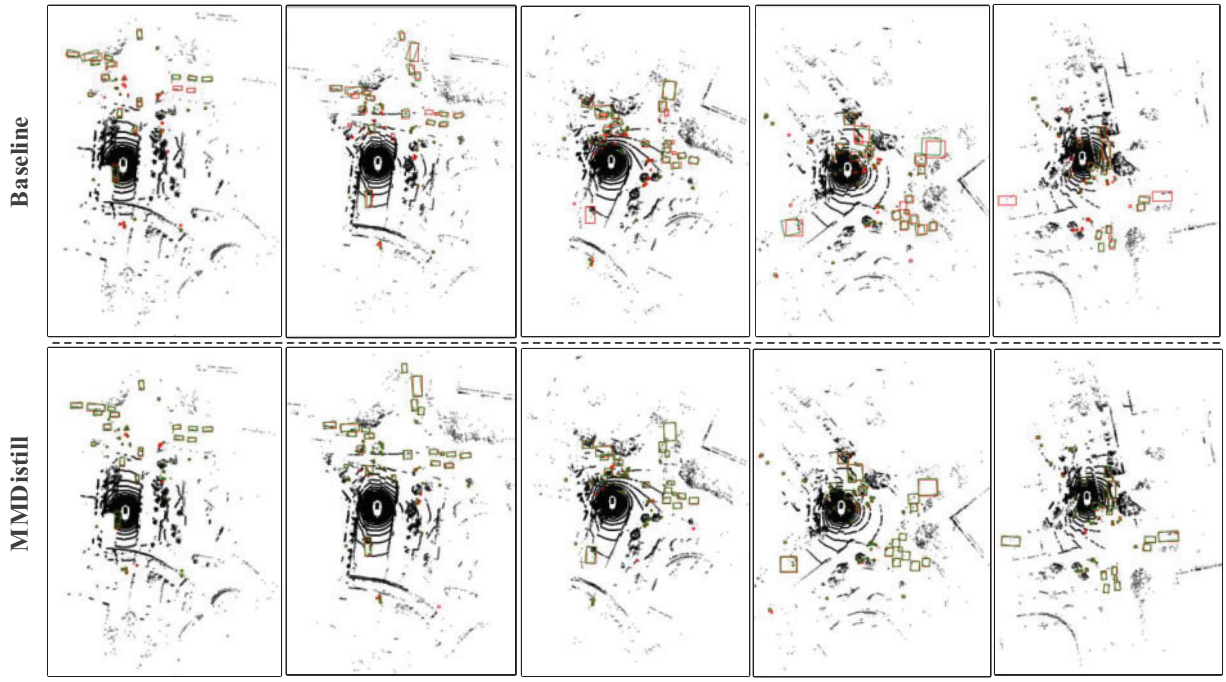


Figure 5: Comparison of predictive ability for the rotation angle of the bounding box. The red and green boxes denote the ground truth and detection results, respectively. A higher degree of alignment between the green and red boxes indicates a better predictive ability for the rotation angle of the bounding box for an object

5 Conclusion

This paper proposes a multi-modal BEV distillation framework named MMDistill, which achieves performance improvement in multi-view 3D object detection tasks, demonstrating its potential and advantages in practical applications. To alleviate the high cost of multi-modal fusion models, we design a two-stage distillation method that utilizes a more powerful LiDAR-based teacher model to guide the multi-view camera-based student model. It learns cross-modal knowledge to generate

effective multi-modal features. In this way, we can improve the performance of multi-view detectors without introducing additional costs during inference. Compared to multi-modal models, MMDistill offers competitive accuracy while reducing deployment costs. We evaluate our method on the nuScenes dataset and conduct an ablation study to demonstrate the effectiveness of our proposed method. Our future work will explore more efficient fusion strategies, lighter model architectures, and implementation deployment and optimization across a wider range of practical application scenarios.

Acknowledgement: The authors are grateful to all the editors and anonymous reviewers for their comments and suggestions and thank all the members who have contributed to this work with us.

Funding Statement: This paper is supported by the National Natural Science Foundation of China (Grant No. 62302086), the Natural Science Foundation of Liaoning Province (Grant No. 2023-MSBA-070) and the Fundamental Research Funds for the Central Universities (Grant No. N2317005).

Author Contributions: The contribution of each author is as follows: Study conception and design: Tianzhe Jiao, Jie Song; Algorithm implementation and framework design: Tianzhe Jiao, Yuming Chen; Technical assistance and results analysis: Chaopeng Guo, Zhe Zhang; Draft manuscript preparation: Tianzhe Jiao, Yuming Chen; Review and editing: Jie Song. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials used in this review are derived from publicly accessible databases and previously published studies cited throughout the text.

Ethics Approval: Not applicable.

Conflicts of Interest: The data and materials used in this review are derived from publicly accessible databases and previously published studies cited throughout the text.

References

[1] Y. Ai *et al.*, "LiDAR-camera fusion in perspective view for 3D object detection in surface mine," *IEEE Trans. Intell. Veh.*, vol. 9, no. 2, pp. 3721–3730, Feb. 2024. doi: 10.1109/TIV.2023.3343377.

[2] P. Shi, Z. Liu, H. Qi, and A. Yang, "MFF-Net: Multimodal feature fusion network for 3D object detection," *Comput. Mater. Contin.*, vol. 79, no. 3, pp. 5615–5637, Mar. 2023. doi: 10.32604/cmc.2023.037794.

[3] Z. Li *et al.*, "When object detection meets knowledge distillation: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10555–10579, Mar. 2023. doi: 10.1109/TPAMI.2023.3257546.

[4] W. Song, W. Ma, M. Zhang, Y. Zhang, and X. Zhao, "Lightweight diffusion models: A survey," *Artif. Intell. Rev.*, vol. 57, no. 6, pp. 1–51, May 2024. doi: 10.1007/s10462-024-10800-8.

[5] Y. Li *et al.*, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, Feb. 2023, pp. 1–9.

[6] Z. Chong *et al.*, "MonoDistill: Learning spatial features for monocular 3D object detection," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–17.

[7] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun and J. Jia, "Unifying voxel-based representation with transformer for 3D object detection," in *Proc. Adv. Neural Inform. Process. Syst. 35 (NeurIPS 2022)*, New Orleans, LA, USA, Nov. 2022, pp. 1–14.

[8] D. Chen, J. Li, V. Guizilini, R. Ambrus, and A. Gaidon, "Viewpoint equivariance for multi-view 3D object detection," in *Proc. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 9213–9222.

[9] Z. Qin, J. Chen, C. Chen, X. Chen, and X. Li, "UniFusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view," in *Proc. 2023 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 8656–8665.
[10] Z. Li *et al.*, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 1–18.
[11] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao and L. Huang, "What makes multi-modal learning better than single (provably)," in *Proc. Neural Inform. Process. Syst.*, Dec. 2021, pp. 10944–10956.
[12] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1907–1915.
[13] Z. Liu *et al.*, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Automat.*, London, UK, May 2023, pp. 2774–2781.
[14] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "FUTR3D: A unified sensor fusion framework for 3D detection," in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, pp. 172–181.
[15] L. Wang *et al.*, "Multi-modal 3D object detection in autonomous driving: A survey and taxonomy," *IEEE Trans. Intell. Veh.*, vol. 8, no. 7, pp. 3781–3798, Jul. 2023. doi: 10.1109/TIV.2023.3264658.
[16] T. Ma, W. Tian, and Y. Xie, "Multi-level knowledge distillation for low resolution object detection and facial expression recognition," *Knowl.-Based Syst.*, vol. 240, pp. 1–15, Mar. 2022. doi: 10.1016/j.knosys.2022.108136.
[17] S. Xu *et al.*, "IDA-DET: An information discrepancy aware distillation for 1-bit detectors," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 346–361.
[18] Z. Yang *et al.*, "Focal and global knowledge distillation for detectors," in *Proc. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 4643–4652.
[19] Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, "DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Paris, France, Oct. 2023, pp. 8603–8612.
[20] H. Zhao, Q. Zhang, S. Zhao, Z. Chen, J. Zhang and D. Tao, "SimDistill: Simulated multi-modal distillation for BEV 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, Feb. 2024, pp. 7460–7468.
[21] M. Klingner, S. Borse, V. R. Kumar, B. Rezaei, V. Narayanan and S. Yogamani, "X3KD: Knowledge distillation across modalities, tasks and stages for multi-camera 3D object detection," in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 13343–13353.
[22] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, Jun. 2019, pp. 6105–6114.
[23] Y. Wang and J. M. Solomon, "Object DGCNN: 3D object detection using dynamic graphs," in *Proc. Neural Inform. Process. Syst.*, Dec. 2021, pp. 20745–20758.
[24] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 12697–12705.
[25] H. Caesar *et al.*, "nuScenes: A multi-modal dataset for autonomous driving," in *Proc. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 11618–11628.
[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, New Orleans, LA, USA, May 2019, pp. 1–8.
[27] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.
[28] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. 2021 IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 913–922.

[29] Z. Wang *et al.*, "STS: Surround-view temporal stereo for multi-view 3D detection," 2022, *arXiv:2208.10145*.
[30] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, Auckland, New Zealand, Dec. 2022, pp. 1–12.
[31] Y. Liu, T. Wang, X. Zhang, and J. SunPETR, "PETR: Position embedding transformation for multi-view 3D object detection," in *Proc. Euro. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 531–548.
[32] P. Huang *et al.*, "TiG-BEV: Multi-view BEV 3D object detection via target inner-geometry learning," 2022, *arXiv:2212.13979*.
[33] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 3354–3361.
[34] S. Shi *et al.*, "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection," *Int. J. Comput. Vis.*, vol. 131, no. 2, pp. 531–551, Feb. 2023. doi: 10.1007/s11263-022-01710-9.
[35] Z. Li, Y. Yao, Z. Quan, J. Xie, and W. Yang, "Spatial information enhancement network for 3D object detection from point cloud," *Pattern Recognit.*, vol. 128, no. 1, pp. 1–12, Aug. 2022. doi: 10.1016/j.patcog.2022.108684.
[36] X. Li *et al.*, "Homogeneous multi-modal feature fusion and interaction for 3D object detection," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany, Oct. 2022, pp. 691–707.
[37] J. H. Yoo, Y. Kim, J. S. Kim, and J. W. Choi, "3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection," in *Proc. Euro. Conf. Comput. Vis.*, Glasgow, UK, Aug. 2020, pp. 720–736.
[38] C. Wang, H. Chen, Y. Chen, P. Hsiao, and L. Fu, "VoPiFNet: Voxel-pixel fusion network for multi-class 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 8527–8537, Aug. 2024. doi: 10.1109/TITS.2024.3392783.
[39] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 4603–4611.
[40] J. Kim, M. Seong, G. Bang, D. Kum, and J. W. Choi, "RCM-Fusion: Radar-camera multi-level fusion for 3D object detection," in *Proc. 2024 IEEE Int. Conf. Robot. Automat. (ICRA)*, Yokohama, Japan, May 2024, pp. 18236–18242.
[41] P. Wolters *et al.*, "Unleashing hyDRa: Hybrid fusion, depth consistency and radar for unified 3D perception," 2024, *arXiv:2403.07746*.