SimDistill: Simulated Multi-Modal Distillation for BEV 3D Object Detection

Haimei Zhao[1], Qiming Zhang[1], Shanshan Zhao[1] Zhe Chen[2] Jing Zhang[1], Dacheng Tao[1]

[1]School of Computer Science, The University of Sydney, Australia,
[2]School of Computing, Engineering and Mathematical Sciences, La Trobe University, Australia
hzha7798@uni.sydney.edu.au, qzha2506@uni.sydney.edu.au, sshan.zhao00@gmail.com, zhe.chen@latrobe.edu.au,
jing.zhang1@sydney.edu.au, dacheng.tao@gmail.com

Abstract

Multi-view camera-based 3D object detection has become popular due to its low cost, but accurately inferring 3D geometry solely from camera data remains challenging and may lead to inferior performance. Although distilling precise 3D geometry knowledge from LiDAR data could help tackle this challenge, the benefits of LiDAR information could be greatly hindered by the significant modality gap between different sensory modalities. To address this issue, we propose a **Si**mulated **m**ulti-modal **Distill**ation (**SimDistill**) method by carefully crafting the model architecture and distillation strategy. Specifically, we devise multi-modal architectures for both teacher and student models, including a LiDAR-camera fusion-based teacher and a simulated fusion-based student. Owing to the "identical" architecture design, the student can mimic the teacher to generate multi-modal features with merely multi-view images as input, where a geometry compensation module is introduced to bridge the modality gap. Furthermore, we propose a comprehensive multi-modal distillation scheme that supports intra-modal, cross-modal, and multi-modal fusion distillation simultaneously in the Bird's-eye-view space. Incorporating them together, our SimDistill can learn better feature representations for 3D object detection while maintaining a cost-effective camera-only deployment. Extensive experiments validate the effectiveness and superiority of SimDistill over state-of-the-art methods, achieving an improvement of 4.8% mAP and 4.1% NDS over the baseline detector. The source code will be released at https://github.com/ViTAE-Transformer/SimDistill.

Introduction

3D object detection is a pivotal technique with extensive applications in fields such as autonomous driving, robotics, and virtual/augmented reality (Zhang and Tao 2020). In recent years, camera-based 3D object detection methods, which infer objects' 3D locations from multi-view images (Huang et al. 2021; Li et al. 2023b), have attracted great attention from both academia and industry because of the high perceptual ability of dense color and texture information with low deployment cost. However, due to the lack of accurate 3D geometry reasoning ability, their detection performance falls largely behind LiDAR-based methods, which poses a challenge to the practical deployment of camera-based methods.

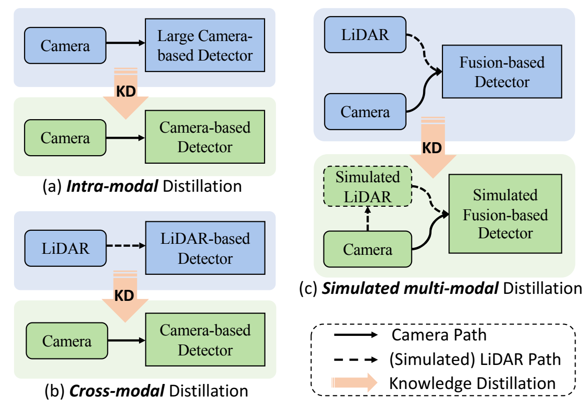

Figure 1: Comparison of our SimDistill with previous distillation frameworks. (a) Intra-modal distillation between camera-only teacher and student models cannot learn accurate 3D information due to the limited capacity of the teacher model for inferring 3D geometry. (b) Cross-modal distillation between the LiDAR teacher and Camera student enables learning useful 3D information from the teacher but suffers from the large cross-modal gap. (c) Our simulated multi-modal distillation enables effective knowledge distillation within/between modalities and fully takes advantage of complementary information from different modalities.

To address this issue, researchers attempt to impose LiDAR data to provide accurate 3D geometry information. Some multi-view camera-based methods (Li et al. 2023b,a) generate ground truth depth from LiDAR point cloud and use it as the supervisory signal for depth estimation to help transform image features to the Bird's-eye-view (BEV) space (Zhao et al. 2022) accurately. Except for directly using LiDAR as supervision during training, some recent work employs LiDAR information by applying the knowledge distillation (KD) technique (Gou et al. 2021) to improve the detection performance of camera-based methods. KD-based 3D object detection methods usually leverage the informative features or predictions of a well-trained teacher model to facilitate the learning of the student model. One straightforward approach is intra-modal distillation (Li et al. 2022a; Zhang et al. 2022) between a large teacher model and a small student model, as shown in Figure 1 **(a)**, which conducts distillation within the image modality. However, the ceiling performance of the model can be limited since the teacher model infers 3D geometry solely from image data as well. Another approach is cross-modal distillation, as shown in Figure 1 **(b)**, which utilizes LiDAR data as the input of teacher models and transfers 3D knowledge to camera-based students (Chong et al. 2021; Chen et al. 2023; Li et al. 2022b). The student is usually forced to learn and mimic the output of a LiDAR-based teacher in different representation spaces, including monocular view features (Chong et al. 2021), BEV features (Chen et al. 2023), and voxel features (Li et al. 2022b). Nevertheless, performing knowledge distillation directly between different modalities might face significant cross-modal gaps and struggle in aligning features learned by distinct architectures of teacher and student models, resulting in limited performance improvements.

In this paper, we address this challenge from the perspective of architecture design and multi-modal knowledge distillation scheme, presenting a **Si**mulated **m**ulti-modal **Distill**ation (**SimDistill**) method for 3D object detection. It encourages the student to ***simulate*** multi-modal representation with solely image modality as input thereby advancing the representation learning for 3D object detection. For the architecture, we design a LiDAR-camera fusion-based teacher and a simulated multi-modal student. The student model not only involves a camera path but also introduces an additional simulated LiDAR path parallel to the camera counterpart, as shown in Figure 1 **(c)**. Different from other distillation methods in Figure 1 **(a)** and **(b)**, our student model possesses two knowledge-transferring paths to learn complementary information from the corresponding two branches of the teacher model. Despite the simulation nature, our student shares a nearly "identical" pipeline as the teacher to produce the camera feature, LiDAR feature, fusion feature, and detection predictions. The resulting aligned learning workflow greatly mitigates the cross-modal gap and benefits multi-modal knowledge distillation.

Built upon this architecture, we propose a new simulated multi-modal distillation scheme that supports intra-modal (IMD), cross-modal (CMD), and multi-modal fusion distillation (MMD) simultaneously. We adopt the widely used MSE loss on the corresponding feature representations distillation in the unified BEV space and an additional quality-aware prediction distillation (Hong, Dai, and Ding 2022). It is noteworthy that directly transferring knowledge from the LiDAR feature to the simulated LiDAR feature is challenging due to the cross-modal gap. To approach this challenge, we devise a geometry compensation module in CMD to help it attend more to the valuable surrounding context from the learned locations to conduct geometry remediation and distill more informative features from object regions. Equipping the proposed model with the distillation scheme, our SimDistill could effectively learn better feature representations for 3D object detection while enjoying cost-effective camera-only deployment.

The main contribution of this paper is threefold. **Firstly**, we propose a unique multi-modal distillation framework for BEV 3D object detection, including a LiDAR-camera fusion-based teacher and a carefully crafted simulated multi-modal student. By ensuring that the teacher and student models share nearly the same workflows, we effectively reduce the modality gap in knowledge distillation. **Secondly**, we present a novel simulated multi-modal distillation scheme that supports intra-modal, cross-modal, and multi-modal fusion distillation simultaneously, which is a universal strategy and can be easily adapted to different models. **Thirdly**, comprehensive experiments and ablation studies on the nuScenes benchmark validate the effectiveness of SimDistill and its superiority over existing state-of-the-art methods, improving the mAP and NDS of the baseline detector by 4.8% and 4.1%, respectively.

Related Work

Camera-based 3D Object Detection Monocular 3D object detection methods have been widely studied and made great progress (Simonelli et al. 2019; Reading et al. 2021; Wang et al. 2021b; Lu et al. 2021; Ma et al. 2021; Huang et al. 2022a) on the KITTI (Geiger, Lenz, and Urtasun 2012) benchmark. However, with the release of large-scale datasets with multi-view cameras such as nuScenes (Caesar et al. 2020) and Waymo (Sun et al. 2020), there is growing attention for accurate 3D object detection in these more challenging scenes. Recent works adopt the Bird's-eye view (BEV) representation as an ideal feature space for multi-view perception due to its excellent ability to address scale-ambiguity and occlusion issues (Huang et al. 2021; Huang and Huang 2022; Li et al. 2022c). Various methods have been proposed to transform perspective image features to the BEV space, such as the lifting operation from LSS (Philion and Fidler 2020) used by BEVDet (Huang et al. 2021) and the cross-attention mechanism-based grid queries used by BEVFormer (Li et al. 2022c). The camera-based BEVDet approach has been further improved by imposing depth supervision (Li et al. 2023b,a; Wang et al. 2022; Chu et al. 2023) and temporal aggregation (Huang and Huang 2022; Park et al. 2022), resulting in better performance. However, there is still a significant performance gap compared to LiDAR-based and fusion-based counterparts.

Fusion-based 3D Object Detection LiDAR differs from cameras in its ability to capture precise geometric and structural information. However, the data it produces is sparse and irregular, with a large volume. Some methods use PointNet (Qi et al. 2017a) directly on the raw point cloud (Qi et al. 2017b; Shi, Wang, and Li 2019; Chen et al. 2022) to learn 3D features, while others voxelize the point cloud into pillars (Lang et al. 2019; Wang et al. 2020; Yin, Zhou, and Krahenbuhl 2021) or voxels (Zhou and Tuzel 2018; Yan, Mao, and Li 2018) before extracting features using SparseConvNet (Graham, Engelcke, and Van Der Maaten 2018). State-of-the-art techniques (Yin, Zhou, and Krahenbuhl 2021; Bai et al. 2022) typically transform 3D features into the BEV representation to simplify operations in 3D space, and then feed the resultant features to subsequent detection heads.

Due to their distinct strengths in perceiving, both cameras and LiDAR are integrated into sensor fusion methods to enhance the performance of perception systems. Existing fusion-based approaches can be categorized as input-level methods (Vora et al. 2020; Wang et al. 2021a; Xu et al. 2021) and feature-level methods (Bai et al. 2022; Liang et al. 2022; Liu et al. 2023; Yan et al. 2023), depending on the stage at which information from different sensors is combined. Recently, it has been shown that BEV space is an ideal space for multi-modal fusion, resulting in outstanding performance (Liang et al. 2022; Liu et al. 2023). These methods follow a simple yet effective pipeline that involves extracting features from both modalities, transforming features into the BEV space, fusing multi-modal features using fusion modules, and conducting subsequent detection, largely improving the performance.

Knowledge Distillation in 3D Object Detection Knowledge distillation presents a promising avenue for empowering compact models (*i.e.*, students) with effective representations via knowledge transfer from larger models (*i.e.*, teachers). In the context of 3D object detection, prior research (Cho et al. 2023; Zhang et al. 2023, 2022; Yang et al. 2022) has successfully extended knowledge distillation techniques, requiring the student network to emulate features or predictions learned by a teacher model within the same modality. Recent advancements in the area of KD-based 3D object detection have ventured into employing teachers from different modalities (Chong et al. 2021; Li et al. 2022a; Hong, Dai, and Ding 2022; Chen et al. 2023), *i.e.*, leveraging a LiDAR-based teacher. UVTR (Li et al. 2022b) aligns features from both LiDAR and camera in voxel space, facilitating knowledge distillation. BEVDistill (Chen et al. 2023) transforms features into the BEV space for the feature and instance-wise prediction distillation. In a similar vein, TiG-BEV (Huang et al. 2022b) introduces inner-depth supervision and inner-feature distillation to enhance geometry learning in the BEV space. These cross-modal distillation techniques underscore the potential of transferring knowledge from robust LiDAR teachers to camera-based students. Nevertheless, these approaches overlook the prospect of distilling multi-modal knowledge for 3D object detection. Our approach diverges by exploring a multi-modal teacher and designing a nearly identical yet simulated multi-modal architecture alongside tailored distillation schemes to effectively perform multi-modal distillation. While concurrent work Unidistill (Zhou et al. 2023) also embraces a multi-modal teacher, it is designed as a universal knowledge distillation framework to support both single-to-single and fusion-to-single cross-modal distillation. It pays no attention to the architecture discrepancy issue between teacher and student and fails to perform comprehensive multi-modal distillation and overcome the cross-modal gap.

Methodology

In this section, we present the details of how the proposed SimDistill realizes comprehensive multi-modal knowledge distillation for 3D object detection. We first introduce the model architecture, which consists of a multi-modal fusion-based teacher and a simulated multi-modal student. Next, we describe the simulated multi-modal distillation scheme that supports knowledge distillation within and between modalities. Last, we present the training objectives for our method.

Multi-modal Architecture

SimDistill is proposed as a flexible multi-modal distillation method, offering the flexibility to select both the teacher model and the student model from diverse methods. In the subsequent sections, we present a concrete implementation of SimDistill, employing BEVFusion (Liu et al. 2023) as the teacher model and design the student model based on the camera branch of BEVFusion (BEVFusion-C). The architectural layout of SimDistill is depicted in Figure 2. The upper block depicts the configuration of the teacher model, while the lower block represents the student model. In both instances, the LiDAR branch and the camera branch workflows are denoted by red and blue arrows, respectively.

Multi-modal Teacher To encode multi-modal knowledge effectively, we adopt the state-of-the-art fusion-based method, *i.e.*, BEVFusion (Liu et al. 2023) as the teacher model. Its architecture comprises two branches, as depicted in the top part of Figure 2. The LiDAR branch follows the standard pipeline of a LiDAR-based detector (Yan, Mao, and Li 2018; Yin, Zhou, and Krahenbuhl 2021). It uses SparseConvNet (Graham, Engelcke, and Van Der Maaten 2018) En_{3D}^T to extract the 3D features, and obtains the BEV features $F_{L_{bev}}^T$ through vertical dimension reduction ($\mathrm{Flatten}$). On the other hand, the camera branch follows the paradigm of BEVDet (Huang et al. 2021), using a 2D feature extractor En_{2D}^T and an efficient projection $Proj^T$ to transform features from the camera view to the BEV space $F_{C_{bev}}^T$. Both modalities' features are then embedded in a unified BEV space using a fully-convolutional fusion module $fuse^T$, which produces the fused BEV features $F_{U_{bev}}^T$. Finally, a detection head $head^T$ predicts the objects' bounding boxes and classes P^T. This process is formulated as:

$$\begin{aligned} F_{L_{bev}}^T &= \mathrm{Flatten}(En_{3D}^T(L)), \\ F_{C_{bev}}^T &= Proj^T(En_{2D}^T(I)), \\ F_{U_{bev}}^T &= fuse^T(F_{L_{bev}}^T, F_{C_{bev}}^T), \\ P^T &= head^T(F_{U_{bev}}^T), \end{aligned} \tag{1}$$

where L and I denotes LiDAR and image input. T and S in all formulations represent the teacher and student models. The projection $Proj$ will be explained in the following part.

Simulated multi-modal Student For the student model, we adopt BEVFusion-C (Liu et al. 2023) as the basis model. To mimic the multi-modal fusion pipeline of the teacher model, we make a modification to the network, as shown in the bottom part of Figure 2. Specifically, after feature extraction from the 2D encoder En_{2D}^S, we devise an additional simulated LiDAR branch (workflow denoted with red arrows) in parallel to the camera branch (blue arrows in the bottom) to simulate LiDAR features from images, which are supervised by the real LiDAR features from the teacher.

In the camera branch, we adopt the same efficient view projection $Proj_C$ with the one used in the teacher model ($Proj^T$) to transform camera-view features to the corresponding BEV features $F_{C_{bev}}^S$ (Philion and Fidler 2020; Liu et al. 2023). During the feature transformation, the extracted 2D feature $F_{C_{uv}}^S$ is first feed to a light Depth Net ϕ and a

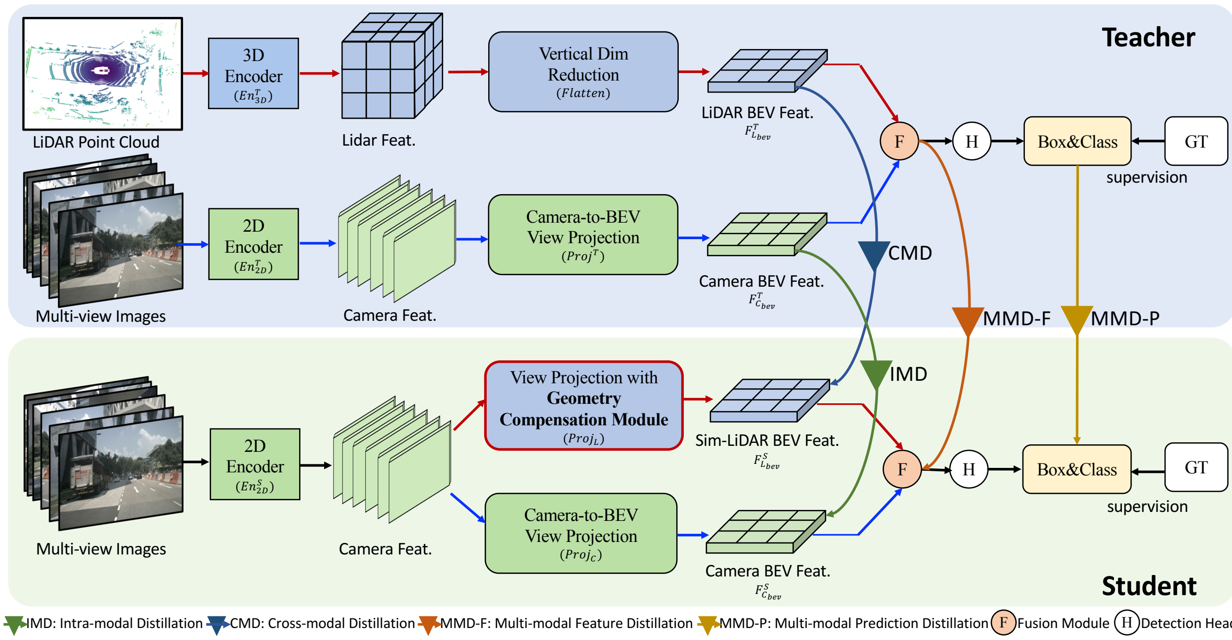

Figure 2: Overall pipeline of SimDistill. It consists of a fusion-based teacher model (top) and a simulated multi-modal student model (bottom). SimDistill supports (1) Intra-Modal Distillation (IMD) between the camera features of the teacher and student; (2) Cross-Modal Distillation (CMD) between the teacher's LiDAR feature and the student's Simulated-LiDAR feature. (3) Multi-Modal fusion Distillation (MMD) between the fusion features (MMD-F) and predictions (MMD-P) of the teacher and student. The workflows of the (simulated) LiDAR and camera branches are denoted by red and blue arrows, respectively.

Context Net ψ to predict the depth distribution and semantic context on each pixel. Then, each 2D feature pixel can be scattered into D discrete points along the camera ray by rescaling the context feature with their corresponding depth probabilities. The resulting 3D feature point cloud is then processed by the efficient BEV pooling operation ρ, to aggregate features in BEV grids and obtain the BEV features:

$$F^S_{C_{bev}} = Proj_C(F^S_{C_{uv}}) = \rho(\psi(F^S_{C_{uv}}) \times \phi(F^S_{C_{uv}})). \quad (2)$$

In the simulated LiDAR branch, to acquire the simulated LiDAR feature $F^S_{L_{bev}}$, the view projection $Proj_L$ is combined with a specifically designed geometry compensation module in both camera-view and BEV spaces, which will be explained later in Eq. (5) of Sec. 3.2.2. It offers the ability to mitigate the geometry misalignment caused by inaccurate depth prediction and modality gap during distillation. After obtaining BEV features from two branches $F^S_{C_{bev}}$ and $F^S_{L_{bev}}$, we use the fusion module $fuse^S$ to acquire the multi-modal fusion features $F^S_{U_{bev}}$. And the detection head $head^S$ is exploited to yield the final detection results P^S. Both the fusion module and detection head have the same architecture as the teacher. This process is formulated as:

$$\begin{aligned} F^S_{C_{uv}} &= En^S_{2D}(I), \\ F^S_{L_{bev}} = Proj_L(F^S_{C_{uv}}), \quad F^S_{C_{bev}} &= Proj_C(F^S_{C_{uv}}), \\ F^S_{U_{bev}} &= fuse^S(F^S_{L_{bev}}, F^S_{C_{bev}}), \\ P^S &= head^S(F^S_{U_{bev}}). \end{aligned} \quad (3)$$

Owing to the simulated multi-modal fusion architecture, the student model can learn features from multiple modalities without equipping a real LiDAR. In the next part, we will explain how this architecture facilitates effective knowledge distillation within and between modalities, including intra-modal, cross-modal, and multi-modal fusion distillation.

Multi-modal Distillation

To better utilize the knowledge of different modalities encoded by different branches of the teacher model, we propose a novel simulated multi-modal distillation scheme including Intra-modal Distillation (IMD), Cross-modal Distillation (CMD), and Multi-modal fusion Distillation (MMD).

Intra-modal Distillation Since both the teacher and student models take images as input, a straightforward strategy is to align the image features from the camera branch of both models, which we name intra-modal distillation. Specifically, we leverage the BEV feature of the teacher $F^T_{C_{bev}}$ as the supervisory signal for the learning of the student counterpart $F^S_{C_{bev}}$ via an MSE loss, *i.e.*,

$$\mathcal{L}_{IMD} = \text{MSE}(F^T_{C_{bev}}, F^S_{C_{bev}}). \quad (4)$$

Due to the same modality in IMD, the student model can be trained directly through the above distillation objective to gain useful visual domain knowledge to facilitate 3D object detection performance. However, relying on images alone may not provide enough geometry-related information to help detect target objects. To address this limitation,

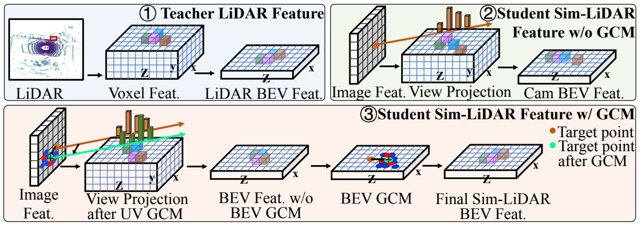

Figure 3: Illustration of Geometry Compensation Module (GCM). The colorful voxels denote learned features of the target object. Best viewed with zoom-in.

we implement cross-modal distillation on the proposed simulated LiDAR branch in the student model, enabling it to gain knowledge from the LiDAR modality.

Cross-modal Distillation CMD aims to align the LiDAR BEV features of the teacher and the simulated LiDAR BEV features of the student. However, due to geometry misalignment and modal difference, directly applying the distillation loss between features generated from different modalities may lead to an incorrect mimic of the noisy features and inaccurate 3D geometry representation. Therefore, we propose a geometry compensation module to address the geometry misalignment and handle the modal difference.

Geometry Compensation Module (GCM) A crucial process in the multi-view camera-based detection method is the view projection operation, which transforms camera-view (UV) features into the BEV space. Inaccurate geometry inference in this process leads to geometry misalignment between features learned from images and LiDAR, exacerbating the modality gap. Therefore, we propose to conduct geometry compensation before and after the view projection in the simulated LiDAR branch to learn more accurate geometry features in both UV and BEV space.

Deformable Convolutions and Deformable Attentions are known to be effective in enabling neural networks to model spatial transformations and account for geometric deformations or misalignments (Dai et al. 2017; Zhu et al. 2020). Therefore, we adopt deformable self-attention layers to construct GCM, as shown in Figure 3. For geometry compensation in the UV space, we first generate a uniform grid of points Q_{uv} as query points for each 2D camera feature $F^S_{C_{uv}}$. Then, we learn offsets based on each point $q_{(u,v)} \in Q_{uv}$ to generate a set of most related points $\mathcal{P}_{uv}$ around it. These learned points $\mathcal{P}_{uv}$ are taken as reference points and keys used to sample the value features from the 2D camera features $F^S_{C_{uv}}$. With the optimization signals gradually improving attentive locations, the module facilitates the model to compensate for geometric transformations in the x-y plane. We apply standard multi-head attention, learning individual offsets for each head, which captures abundant information and improves feature representations for subsequent context learning, depth estimation, and 3D geometry inference. Similarly, we employ a BEV geometry compensation module after transforming the camera-view features to BEV features $F^S_{C_{uv-bev}}$, which is responsible for correcting the key feature locations in the x-z plane. By doing so, the geometry compensation in the two complementary 2D views can comprehensively improve the feature representation. Overall, the view projection with GCM used in the simulated LiDAR branch is formulated here, with reference to Eq. (2):

$$\begin{aligned} F^S_{L_{bev}} &= Proj_L(F^S_{C_{uv}}) \\ &= GC_{bev}(\rho(\psi(GC_{uv}(F^S_{C_{uv}})) \times \phi(GC_{uv}(F^S_{C_{uv}})))), \end{aligned} \tag{5}$$

where $GC_{uv}(F^S_{C_{uv}}) = \mathrm{DeformAttn}(\mathrm{Q_{uv}}, \mathcal{P}_{\mathrm{uv}}, F^S_{C_{uv}})$ and $GC_{bev}(F^S_{C_{uv-bev}}) = \mathrm{DeformAttn}(\mathrm{Q_{bev}}, \mathcal{P}_{\mathrm{bev}}, F^S_{C_{uv-bev}})$, denoting the UV Geometry Compensation and BEV Geometry Compensation, respectively. Q_{bev} and $\mathcal{P}_{bev}$ are query and reference points generated for BEV features $F^S_{C_{uv-bev}}$.

To get the final simulated LiDAR feature for distillation, we also implement a simple yet effective object-aware mask $\mathcal{M}$ to select the most informative features at the end of GCM. We generate masks in the BEV space from the ground truth center points and bounding boxes using a heatmap-like approach like BEVDisitll (Chen et al. 2023). Therefore, the CMD loss is formulated as:

$$\mathcal{L}_{CMD} = \mathrm{MSE}(\mathcal{M} \odot F^T_{L_{bev}}, \mathcal{M} \odot F^S_{L_{bev}}) \tag{6}$$

where $\odot$ is Hadamard product. The object-aware mask is a technique we utilize together with GCM to improve the ability to overcome the cross-modal gap in CMD and we refrain from attributing it as our original contribution.

Multi-modal fusion Distillation In light of the aligned architecture and workflow with the teacher model, the student model also produces multi-modal fusion features as well as detection predictions. To make the fused feature and predictions consistent with those in the teacher model, we devise multi-modal distillation in both feature level (MMD-F) and prediction level (MMD-P). Owing to the proximity of the fusion module and the detection head, MMD-F is expected to distill highly useful multi-modal knowledge that directly contributes to the detection. It is implemented by aligning the fusion feature of the teacher model and the simulated fusion feature of the student model:

$$\mathcal{L}_{MMD-F} = \mathrm{MSE}(F^T_{U_{bev}}, F^S_{U_{bev}}). \tag{7}$$

After the fusion module, the fused feature in the student model is fed into the detector to output the detection results in the same way as the teacher model. Thus, we also employ MMD-P by taking the predictions from the teacher model as soft labels. We adopt the quality-aware prediction distillation loss L_{MMD-P} (Hong, Dai, and Ding 2022), which consists of the classification loss $\mathcal{L}_{cls}$ for object categories and the regression loss $\mathcal{L}_{reg}$ for 3D bounding boxes:

$$\begin{aligned} \mathcal{L}_{MMD-P} &= \mathcal{L}_{reg} + \mathcal{L}_{cls}, \\ &= \mathrm{SmoothL1}(P^T_B, P^S_B) \cdot s + \mathrm{QFL}(P^T_C, P^S_C) \cdot s, \end{aligned} \tag{8}$$

where P^T_B and P^T_C (resp. P^S_B and P^S_C) denote the predicted bounding boxes and categories by the teacher model (resp. the student model). $\mathrm{QFL}(\cdot)$ denotes the quality focal loss (Li

Methods	Modality	Backbone	Image Size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFusion (Liang et al. 2022)	LC	VoxelNet SwinT	448 × 800	67.9	71.0	-	-	-	-	-
BEVFusion (Liu et al. 2023)	LC	VoxelNet SwinT	256 × 704	68.5	71.4	28.6	25.3	30.0	25.4	18.6
FCOS3D (Wang et al. 2021b)	C	R101	900 × 1600	29.5	37.2	80.6	26.8	51.1	113.1	17.0
BEVDet (Huang et al. 2021)	C	R50	256 × 704	29.8	37.9	72.5	27.9	58.9	86.0	24.5
PETR (Liu et al. 2022)	C	R50	384 × 1056	31.3	38.1	76.8	27.8	56.4	92.3	22.5
DETR3D (Huang and Huang 2022)	C	R101	900 × 1600	34.9	43.4	71.6	26.8	37.9	84.2	20.0
Set2Set (Li et al. 2022b)	C*	R50	900 × 1600	33.1	41.0	-	-	-	-	-
MonoDistill (Chong et al. 2021)	C*	R50	900 × 1600	36.4	42.9	-	-	-	-	-
UVTR (Li et al. 2022b)	C*	R50	900 × 1600	36.2	43.1	-	-	-	-	-
TiG-BEV (Huang et al. 2022b)	C*	R50	256 × 704	33.1	41.1	67.8	27.1	58.9	78.4	21.8
UniDistill (Zhou et al. 2023)	C*	R50	256 × 704	26.5	37.8	-	-	-	-	-
BEVDistill (Chen et al. 2023)	C*	SwinT	256 × 704	36.3	43.6	64.2	27.4	57.6	87.8	28.2
BEVFusion-C (Liu et al. 2023)	C	SwinT	256 × 704	35.6	41.2	66.8	27.3	56.1	89.6	25.9
SimDistill	C*	SwinT	256 × 704	40.4	45.3	52.6	27.5	60.7	80.5	27.3

Table 1: Quantitative comparisons on the nuScenes validation Set. L and C in the second column denote the input modality, *i.e.*, LiDAR and camera, while C* means using LiDAR for knowledge distillation during training.

et al. 2020). s is a quality score used as the loss weight, obtained by measuring the IoU between the predictions and the ground truth to determine the confidence of the soft label.

Discussion It is noteworthy that previous methods have not explored multi-modal fusion distillation due to the absence of a dedicated multi-modal architecture in the student model for aligning fusion features or predictions. Instead, these methods distill information solely by aligning the teacher model's fusion features or predictions to a single-modal student counterpart, which leads to subpar performance due to the modality gap. Furthermore, no studies have investigated the impact of comprehensive multi-modal distillation, including intra-modal, cross-modal, and multi-modal fusion distillation, simultaneously. Our SimDistill makes progress by effectively performing multi-modal fusion distillation through its simulated multi-modal architecture. This complements intra-modal and cross-modal distillation (Sec. 4.3), resulting in improved performance.

Training Objective

Apart from the above distillation losses, the student model is also optimized by the common loss of 3D object detection task $\mathcal{L}_{det}$. The overall training objective $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{IMD} + \mathcal{L}_{CMD} + \mathcal{L}_{MMD-F} + \mathcal{L}_{MMD-P} + \mathcal{L}_{det}. \quad (9)$$

Experiment

Experiment Setting

Datasets and Evaluation Metrics We follow the common practice (Huang et al. 2021; Liu et al. 2023; Liang et al. 2022; Li et al. 2023b; Chen et al. 2023) to evaluate our method on the most challenging benchmark, *i.e.*, nuScenes (Caesar et al. 2020). It comprises 700 scenes for training, 150 scenes for validation, and 150 scenes for testing. Each scene includes panoramic LiDAR data and surrounding camera images, which are synchronized to provide convenience for multi-modal-based research. The dataset comprises a total of 23 object categories, and 10 popular classes are considered for computing the final metrics. To align with the official evaluation, we adopt mean Average Precision (**mAP**) and nuScenes detection score (**NDS**) as the main metrics with other 5 metrics for reference.

Implementation Details Our method is implemented with PyTorch using 8 NVIDIA A100 (40G Memory), based on the MMDetection3D codebase (Contributors 2020). We adopt BEVFusion (Liu et al. 2023) as the default teacher model, which takes images with a size of 256×704 and LiDAR point cloud with a voxel size of (0.075m, 0.075m, 0.2m) as input and uses VoxelNet (Zhou and Tuzel 2018) and Swin-T (Liu et al. 2021) as backbones for the two modalities, respectively. During distillation, we utilize the official BEVFusion checkpoint, freeze the teacher model, and train the student model for 20 epochs with batch size 24. The backbone and input resolution are kept the same as BEVFusion-C in both our SimDisitll and our competitor BEVDistill. More implementation details, ablation analysis, and visualizations can be found in Appendices.

Main Results

We compare our SimDistill with state-of-the-art methods on the nuScenes validation set and present the results in Table 1. We group the methods according to the input modality and present the knowledge distillation-based methods in the bottom part (except for baseline BEVFusion-C) for straightforward comparisons. From the table, we can see that fusion-based methods usually possess a stronger perception ability and achieve better performance. However, the high cost of LiDAR may restrict their practical usage. Compared with the baseline BEVFusion-C, SimDistill boosts the performance significantly by 4.8% mAP and 4.1% NDS, clearly validating the effectiveness of the proposed distillation method. Compared with the concurrent distillation methods BEVDistill and UniDistill, our SimDistill achieves much better performance under the same setting.

Ablation Studies

Why Choose Multi-modal Architectures?

To demonstrate the superiority of the proposed simulated

	Teacher	Student	Distillation	mAP↑	NDS↑
a	BEVFusion	BEVFusion-C	MMD-F	35.94	41.75
b	BEVFusion	SimDistill	MMD-F	38.34	44.15
c	BEVFusion-L	BEVFusion-C	CMD-v	35.88	42.87
d	BEVFusion-L	SimDistill	CMD-v	36.80	42.79

Table 2: Ablation study of the model architecture. CMD-v is the vanilla version of CMD without using GCM here.

multi-modal structure, we replace the multi-modal teacher BEVFusion and the simulated multi-modal student SimDistill with their single-modal counterpart BEVFusion-L (*i.e.*, the LiDAR branch of BEVFusion) and BEVFusion-C, respectively. The results are presented in Table 2. We first investigate the influence of using a simulated multi-modal student. In models (a) and (b), we adopt the multi-modal teacher (BEVFusion) but distill the fusion feature to different student architectures. The experiment results show that the simulated multi-modal student (b) outperforms the single-modal one (a) with a clear gain of 2.4 in both mAP and NDS. We then change the teacher to a single-modal one (BEVFusion-L) to verify the performance of the student. Although directly learning from a cross-modal teacher adversely affects performance due to the modality gap, the multi-modal student (d) still achieves better performance in mAP and comparable results in NDS compared with the single-modal student (c). The two groups of comparisons validate the superiority of using a multi-modal student. Besides, the experiments of (b) and (d) both directly distill the learned feature from the teacher model to the student, which validates the importance of using a multi-modal teacher, *i.e.*, with a gain of 1.54% mAP and 1.36% NDS. In summary, it is crucial to employ multi-modal architectures for both teacher and student models to enhance knowledge transfer and achieve better performance. In addition, employing the proposed simulated multi-modal student model maintains the advantage of cost-effective camera-only deployment.

How Simulated Multi-modal Distillation Works?

To investigate the impact of distillation options, we perform ablation studies and summarize the results in Table 3. Model (a) denotes the baseline model with the proposed simulated multi-modal architecture without any knowledge distillation. We present the gains over Model (a) in the column mAP and NDS. As shown in (b), (c), (e), and (f), employing IMD, vanilla CMD, MMD-F, and MMD-P on the baseline model leads to 1.09%, 1.43%, 2.63%, and 1.02% absolute gains in mAP, respectively, where MMD-F brings the largest gain owing to the rich multi-modality knowledge contained in the fusion features. Interestingly, while the simulated LiDAR branch should possess more accurate 3D geometry than the camera branch, IMD (b) produces a slightly larger gain than vanilla CMD (c). We attribute it to the modality gap between the real LiDAR features of the teacher and the simulated ones of the student.

After using the proposed GCM (*i.e.*, Model (d)), we can see that it helps CMD achieve a gain of 4.12% mAP and 2.82% NDS over the baseline in (a), validating the effectiveness of GCM in overcoming the side effect of the modality gap during distillation. After incorporating all the components including simulated multi-modal architecture and all the distillation techniques, we get our SimDistill model in (g), which delivers the best performance of 40.40 mAP and 45.31 NDS, meanwhile achieving an improvement of 4.8% mAP and 4.1% NDS over the baseline model BEVFusion-C.

	IMD	CMD		MMD		mAP↑	NDS↑
		vanilla	GCM	-F	-P		
a						35.71 (-)	41.97 (-)
b	✓					37.14 (+1.43)	42.67 (+0.70)
c		✓				36.80 (+1.09)	42.79 (+0.82)
d		✓	✓			39.83 (+4.12)	44.79 (+2.82)
e				✓		38.34 (+2.63)	44.15 (+2.18)
f					✓	36.73 (+1.02)	42.52 (+0.55)
g	✓	✓	✓	✓	✓	40.40 (+4.69)	45.31 (+3.34)

Table 3: Ablation study of different distillation options.

Methods	FPS	GFlops	mAP	NDS
BEVDet (Huang et al. 2021)	15.6	215.3	31.2	39.2
BEVFormer (Li et al. 2022c)	2.4	1303.5	37.5	44.8
BEVDistill(Chen et al. 2023)	3.7	608.8	36.3	43.6
BEVFusion-C (Liu et al. 2023)	13.4	165.1	35.6	41.2
SimDistill	11.1	219.1	40.4	45.3

Table 4: Comparison of model efficiency.

Model Efficiency

We compare the model efficiency with other representative methods in Table 4. Our method achieves an inference speed of 11.1 FPS on a single GPU, running much faster than BEVDistill and BEVFormer. It is comparable to BEVFusion-C but a bit slower than BEVDet, mainly due to the additional simulated LiDAR branch in the architecture. Nevertheless, SimDistill significantly outperforms other methods in terms of mAP and NDS.

Conclusion

In this paper, we propose a novel simulated multi-modal distillation method named SimDistill for multi-view BEV 3D object detection by carefully investigating the architecture design and effective distillation techniques. We identify the importance of the multi-modal architecture for multi-modal knowledge distillation and devise a simulated multi-modal student model accordingly. Built upon it, we develop a novel simulated multi-modal distillation scheme that supports intra-modal, cross-modal, and multi-modal fusion knowledge distillation simultaneously. Experiments on the challenging nuScenes benchmark have validated the above findings and the superiority of the proposed distillation methods over state-of-the-art approaches. We believe SimDistill is compatible with other multi-modal teacher and diverse student models, which could lead to enhanced performance and remains a subject for future investigation.

Acknowledgments

This work was supported by Australian Research Council Projects IH-180100002.

References

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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*.

Chen, C.; Chen, Z.; Zhang, J.; and Tao, D. 2022. Sasa: Semantics-augmented set abstraction for point-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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 *The Eleventh International Conference on Learning Representations*.

Cho, H.; Choi, J.; Baek, G.; and Hwang, W. 2023. itkd: Interchange transfer-based knowledge distillation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chong, Z.; Ma, X.; Zhang, H.; Yue, Y.; Li, H.; Wang, Z.; and Ouyang, W. 2021. MonoDistill: Learning Spatial Features for Monocular 3D Object Detection. In *International Conference on Learning Representations*.

Chu, X.; Deng, J.; Zhao, Y.; Ji, J.; Zhang, Y.; Li, H.; and Zhang, Y. 2023. OA-BEV: Bringing object awareness to bird's-eye-view representation for multi-camera 3D object detection. *arXiv preprint arXiv:2301.05711*.

Contributors, M. 2020. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hong, Y.; Dai, H.; and Ding, Y. 2022. Cross-modality knowledge distillation network for monocular 3D object detection. In *Proceedings of the European Conference on Computer Vision*.

Huang, J.; and Huang, G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*.

Huang, J.; Huang, G.; Zhu, Z.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Huang, K.-C.; Wu, T.-H.; Su, H.-T.; and Hsu, W. H. 2022a. Monodtr: Monocular 3d object detection with depth-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Huang, P.; Liu, L.; Zhang, R.; Zhang, S.; Xu, X.; Wang, B.; and Liu, G. 2022b. TiG-BEV: Multi-view BEV 3D object detection via target inner-geometry learning. *arXiv preprint arXiv:2212.13979*.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, J.; Lu, M.; Liu, J.; Guo, Y.; Du, L.; and Zhang, S. 2022a. BEV-LGKD: A unified LiDAR-guided knowledge distillation framework for BEV 3D object detection. *arXiv preprint arXiv:2212.00623*.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*.

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*.

Li, Y.; Chen, Y.; Qi, X.; Li, Z.; Sun, J.; and Jia, J. 2022b. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022c. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision*.

Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; and Tang, Z. 2022. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*.

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

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.; and Han, S. 2023. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*.

Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; Yan, J.; and Ouyang, W. 2021. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ma, X.; Zhang, Y.; Xu, D.; Zhou, D.; Yi, S.; Li, H.; and Ouyang, W. 2021. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Park, J.; Xu, C.; Yang, S.; Keutzer, K.; Kitani, K. M.; Tomizuka, M.; and Zhan, W. 2022. Time Will Tell: New Outlooks and A Baseline for Temporal Multi-View 3D Object Detection. In *The Eleventh International Conference on Learning Representations*.

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

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*.

Reading, C.; Harakeh, A.; Chae, J.; and Waslander, S. L. 2021. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Simonelli, A.; Bulo, S. R.; Porzi, L.; López-Antequera, M.; and Kontschieder, P. 2019. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Vora, S.; Lang, A. H.; Helou, B.; and Beijbom, O. 2020. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, C.; Ma, C.; Zhu, M.; and Yang, X. 2021a. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021b. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wang, Y.; Fathi, A.; Kundu, A.; Ross, D. A.; Pantofaru, C.; Funkhouser, T.; and Solomon, J. 2020. Pillar-based object detection for autonomous driving. In *Proceedings of the European Conference on Computer Vision*.

Wang, Z.; Min, C.; Ge, Z.; Li, Y.; Li, Z.; Yang, H.; and Huang, D. 2022. Sts: Surround-view temporal stereo for multi-view 3d detection. *arXiv preprint arXiv:2208.10145*.

Xu, S.; Zhou, D.; Fang, J.; Yin, J.; Bin, Z.; and Zhang, L. 2021. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *IEEE International Intelligent Transportation Systems Conference*. IEEE.

Yan, J.; Liu, Y.; Sun, J.; Jia, F.; Li, S.; Wang, T.; and Zhang, X. 2023. Cross modal transformer via coordinates encoding for 3D object dectection. *arXiv preprint arXiv:2301.01283*.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*.

Yang, J.; Shi, S.; Ding, R.; Wang, Z.; and Qi, X. 2022. Towards efficient 3d object detection with knowledge distillation. *Advances in Neural Information Processing Systems*.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, J.; and Tao, D. 2020. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*.

Zhang, L.; Dong, R.; Tai, H.-S.; and Ma, K. 2023. Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhang, L.; Shi, Y.; Tai, H.-S.; Zhang, Z.; He, Y.; Wang, K.; and Ma, K. 2022. Structured knowledge distillation towards efficient and compact multi-view 3D fetection. *arXiv preprint arXiv:2211.08398*.

Zhao, H.; Zhang, J.; Zhang, S.; and Tao, D. 2022. Jperceiver: Joint perception network for depth, pose and layout estimation in driving scenes. In *European Conference on Computer Vision*.

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*.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.