MonoTAKD: Teaching Assistant Knowledge Distillation for Monocular 3D Object Detection

Hou-I Liu[1,2]*, Christine Wu[2], Jen-Hao Cheng[2], Wenhao Chai[2], Shian-Yun Wang[3], Gaowen Liu[4], Hugo Latapie[5], Jhih-Ciang Wu[6], Jenq-Neng Hwang[2]†, Hong-Han Shuai[1], Wen-Huang Cheng[7]

[1]National Yang Ming Chiao Tung University, [2]University of Washington, [3]University of Southern California,
[4]Cisco Systems, Inc., [5]Taijitu AI, Inc., [6]National Taiwan Normal University, [7]National Taiwan University

Abstract

Monocular 3D object detection (Mono3D) holds noteworthy promise for autonomous driving applications owing to the cost-effectiveness and rich visual context of monocular camera sensors. However, depth ambiguity poses a significant challenge, as it requires extracting precise 3D scene geometry from a single image, resulting in suboptimal performance when transferring knowledge from a LiDAR-based teacher model to a camera-based student model. To facilitate effective distillation, we introduce Monocular Teaching Assistant Knowledge Distillation (MonoTAKD), which proposes a camera-based teaching assistant (TA) model to transfer robust 3D visual knowledge to the student model, leveraging the smaller feature representation gap. Additionally, we define 3D spatial cues as residual features that capture the differences between the teacher and the TA models. We then leverage these cues to improve the student model's 3D perception capabilities. Experimental results show that our MonoTAKD achieves state-of-the-art performance on the KITTI3D dataset. Furthermore, we evaluate the performance on nuScenes and KITTI raw datasets to demonstrate the generalization of our model to multi-view 3D and unsupervised data settings. Our code is available at ***https://github.com/hoiliu-0801/MonoTAKD****.*

1. Introduction

Monocular 3D object detection (Mono3D) has garnered significant attention in autonomous driving, driven by the affordability and practicality of monocular camera sensors [9, 14, 24, 49, 51]. One widely used approach for improving 3D detection performance is to leverage depth maps as auxiliary supervision. For example, MonoDTR [14] and MonoPGC [41] apply an additional depth branch to predict depth maps and integrate them with the visual features using a transformer-based decoder. However, the accuracy of depth maps estimated from a single image remains constrained for the camera-based models, primarily due to the absence of stereoscopic data, which is critical to the understanding of 3D scene geometry [10, 47].

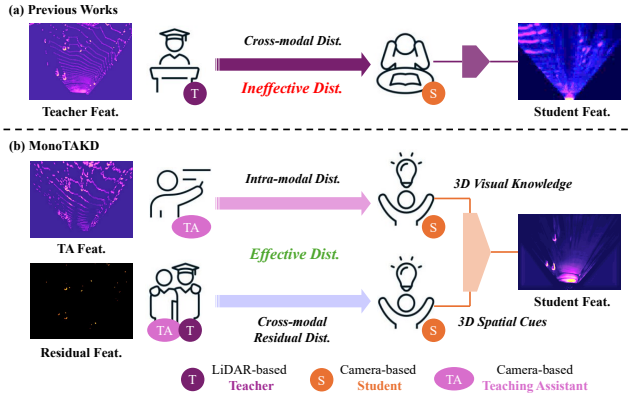

Figure 1. Comparison between previous methods and MonoTAKD (Ours). Previous works [7, 12] face a significant challenge in the distillation process due to a substantial gap in feature representation. Our MonoTAKD incorporates intra-modal and cross-modal residual distillation to enhance learning across this feature representation gap. We visualize the BEV features of each model, along with the residual features derived between the TA and T models, highlighted in orange. Best view with zoom-in and color.

To enhance the extraction and comprehension of 3D information, a promising alternative to depth-guided approaches is to utilize the cross-modal distillation [12, 50]. This approach enables models trained with images to acquire 3D information directly from LiDAR data. For instance, CMKD [12] and UniDistill [50] employ a LiDAR-based model (teacher) to extract critical 3D information and distill this knowledge into the camera-based model (student) through the bird's eye view (BEV) distillation [6]. Nevertheless, the feature representations from different modalities are inherently distinct, which causes ineffective distillation and restricts the performance of the camera-

*This work was done during Hou-I's VISIT program at UW
†Corresponding email: hwang@uw.edu

based student [15, 39], as depicted in Fig. 1 (a).

Although recent approaches involve using an adaptation module [12, 39] or applying regularization [16] to minimize the global distance between cross-modal features, effectively transferring knowledge between such a considerable modality gap continues to be an open challenge. Therefore, we introduce a novel Teaching Assistant Knowledge Distillation framework, MonoTAKD. First, we introduce a robust camera-based model as the teaching assistant (TA) model to guide the camera-based student model with 3D visual knowledge, as shown at the top of Fig. 1 (b). The TA model directly accesses the ground truth (GT) depth map and integrates visual features from the camera-based model to reconstruct optimal BEV features, referred to as 3D visual knowledge. The TA model has two primary functions: (1) By providing the optimal BEV features to the student model, it mitigates the distortion brought by inaccurate depth estimation in the camera-based model. (2) This approach facilitates effective distillation through intra-modal distillation (IMD), as the feature representation gap is smaller within the same modality (camera) compared to cross-modal scenarios (LiDAR and camera) [42].

Specifically, we attempt to make the BEV features of the student model close to those of the TA model[1]. As such, the 3D perception of the student model is enhanced without relying on LiDAR data. It is worth noting that there is no distillation relationship between the teacher and TA models, as visual and LiDAR features serve as independent learning targets. Additionally, we use end-to-end rather than multi-step [26] distillation during training.

Even when the student model perfectly replicates the BEV features of the TA model through IMD, a feature representation gap still exists between the replicated BEV features and those of the teacher model. This gap arises because the monocular image inherently lacks certain 3D spatial cues that are exclusive to the LiDAR modality. To this end, we formulate these 3D spatial cues by computing the feature difference in BEV features between the teacher and TA models. Consequently, we define this feature difference as the residual features that are distilled into the student model through the proposed cross-modal residual distillation (CMRD), as illustrated at the bottom of Fig. 1 (b). This approach empowers the student model to concentrate on learning the crucial 3D spatial cues instead of being compelled to replicate the complex entirety of BEV features from the LiDAR-based teacher. We believe that learning the key differences (residual features) between the LiDAR and camera model is pivotal for the camera-based student to enhance its 3D perception.

Camera-based models struggle to extract 3D spatial cues due to their limited understanding of 3D scene geometry. Thus, we propose a spatial alignment module (SAM) to refine the student's BEV features. This module enhances the 3D spatial cues by capturing global information and compensating for the spatial shifts, improving the 3D representation of the student's BEV features. Notably, SAM is an optional module with an acceptable FLOPs overhead. Even without SAM, our MonoTAKD still significantly surpasses state-of-the-art methods by a noticeable margin.

Our contributions are three-fold:

- Our MonoTAKD utilizes intra-modal distillation to transfer 3D visual knowledge and cross-modal residual distillation to convey essential 3D spatial cues, both directed to the student model.
- We develop a SAM to garner rich global information and to compensate for the spatial shift caused by feature distortion. Also, we design an FFM to expertly integrate features from different modalities, resulting in a more comprehensive 3D representation.
- Experimental results show that MonoTAKD achieves state-of-the-art performance on the KITTI3D dataset and validates its generalizability to multi-view and unsupervised settings on the nuScenes and KITTI raw datasets.

2. Related Work

2.1. LiDAR-based 3D Object Detection

LiDAR sensors have been widely used in 3D object detection since point clouds can represent precise 3D environmental information [34]. For example, point-based methods [31, 33] take raw points as input and process the point-wise features with a sizeable multi-layer perceptron. Voxel-based methods [8, 40, 45] convert point clouds into voxel grids and extract the voxelized features through 3D sparse convolution layers. Although LiDAR-based 3D object detection methods have proven to be high-performing techniques, LiDAR systems are expensive, making them impractical for autonomous driving.

2.2. Depth-guided Mono3D

Monocular 3D object detection methods offer a promising low-cost solution for autonomous driving applications. Depth-guided methods [9, 32, 37, 49] leverage depth information to help 3D perception. Early methods [9, 37] implemented an off-the-shelf depth estimator to predict depth maps, which are then integrated into visual features through a fusion module. CaDDN [32] predicts depth distribution bins with image features to reconstruct a 3D frustum feature. However, image features lack the understanding of 3D scene geometry, leading to unreliable depth predictions.

2.3. Semi-supervised Mono3D

Semi-supervised Mono3D methods typically incorporate unlabeled data into the training dataset, increasing the quan-

[1]We conduct knowledge distillation of student and TA models in a unified BEV space to facilitate 3D detection.

tity of valuable data and thereby enhancing the model's robustness. LPCG [29] conducts instant segmentation on images and uses a heuristic algorithm to create 3D pseudo labels. Mix-Teaching [46] exploits the pre-trained model to generate pseudo labels and pastes them into the image background regions to increase the number of instances in unlabeled data. Yet, increasing training data also brings significant training time, and the generated noisy pseudo-labels may hurt the performance.

2.4. Cross-modal Distillation for Mono3D

Several works [7, 12, 39, 42, 50] apply cross-modal distillation to Mono3D. For example, MonoDistill [7] uses a camera-based model as the teacher model to process the sparse depth map and transfer inferred 3D information to the student model through feature and logit distillation. To enhance 3D information, CMKD [12] and DistillBEV [39] employ a LiDAR-based model to extract enriched 3D features from the LiDAR point cloud. Then, they perform the BEV distillation to transfer this knowledge from the LiDAR modality to the camera modality.

Nevertheless, prior research in this area has predominantly focused on using a single modality teacher, hindering the comprehensive acquisition of 3D knowledge. A camera-based student often struggles to gain this knowledge due to substantial differences in feature representation. Cross-modal feature alignment typically relies on a complex adaptation module, which may be insufficient for effectively bridging this gap [12, 39]. In contrast, our MonoTAKD employs intra-modal distillation to effectively transfer 3D visual knowledge and introduces cross-modal residual distillation to convey essential 3D spatial cues.

3. Method

3.1. Overview

As shown in Fig. 2, MonoTAKD comprises a pre-trained LiDAR-based model as the teacher $\mathcal{T}$ and two camera-based models, serving as the teaching assistant $\mathcal{A}$ and student $\mathcal{S}$. These three models communicate by propagating BEV features, enabling knowledge distillation (KD) during training to enhance the performance of Mono3D tasks. The core innovation lies in integrating two distillation techniques: intra-modal distillation (IMD), which enriches 3D visual knowledge, and cross-modal residual distillation (CMRD), which provides 3D spatial cues. To further strengthen the BEV feature representation in $\mathcal{S}$, we introduce a spatial alignment module (SAM) to refine its BEV features. Finally, a feature fusion module (FFM) is applied to unify features from both distillation branches of $\mathcal{S}$, ensuring a more cohesive 3D representation.

3.2. LiDAR-based Teacher

To exploit the 3D representation and to enhance the 3D perception capabilities of the student model $\mathcal{S}$, we use a pre-trained LiDAR-based model [45] as the teacher model $\mathcal{T}$ to directly encode 3D information for distillation. As shown in the top block of Fig. 2, $\mathcal{T}$ voxelizes the unordered point cloud data captured from LiDAR into voxel grids v, passing through the 3D sparse convolution layers ψ to obtain voxel features. We then employ height compression [12], denoted as ϕ, to embed these voxel features into the BEV space, forming the BEV feature $F^T \in \mathbb{R}^{H\times W\times C}$. The process to derive the BEV feature from $\mathcal{T}$ can be formulated as

$$F^T = \phi(\psi(v)), \tag{1}$$

3.3. Camera-based Teaching Assistant

Cross-modal distillation in Mono3D [7, 12] is challenging due to the knowledge heterogeneity between modalities, and relying on solely complex adaptation modules [12, 39] for feature alignment proves insufficient. To overcome the limitations of direct cross-modal distillation, we introduce a negotiator, *i.e.*, the Teaching Assistant (TA) model, to effectively bridge the teacher and student models, facilitating effective knowledge transfer via IMD. To alleviate knowledge heterogeneity, we select a camera-based model [32] as the TA model $\mathcal{A}$, which shares the same modality as the student model $\mathcal{S}$. We further ensure feature alignment is tractable by jointly projecting the latent representations into BEV space, allowing $\mathcal{S}$ to effectively obtain relevant 3D visual knowledge from $\mathcal{A}$.

As depicted in the middle block of Fig. 2, $\mathcal{A}$ processes a monocular image input by extracting visual features through a 2D backbone (ResNet50), including a 1×1 convolutional layer to adjust the channel dimensions. Unlike previous depth-guided methods [14, 49], $\mathcal{A}$ operates on the ground truth (GT) depth maps $d \in \mathbb{R}^{H\times W\times D}$ by computing the outer product with the resulting visual feature f, where D stands for the number of discrete depth bins. This approach effectively injects accurate depth information into the TA model, reducing feature distortion from inaccurate depth estimations and enabling the TA model to achieve performance closer to the optimal levels expected from a camera-based model. Formally, we formulate the BEV feature from $\mathcal{A}$ as

$$F^A = \phi(f \otimes d), \tag{2}$$

where d represents the GT depth maps, $\otimes$ denotes the outer product, ϕ is the projection function similar to Eq. 1 with minor modifications (*e.g.* interpolation [12]), and 1×1 convolution for arranging the desired dimension. Hereafter, we refer to F^A as 3D visual knowledge, which serves as one of the learning targets for the student model.

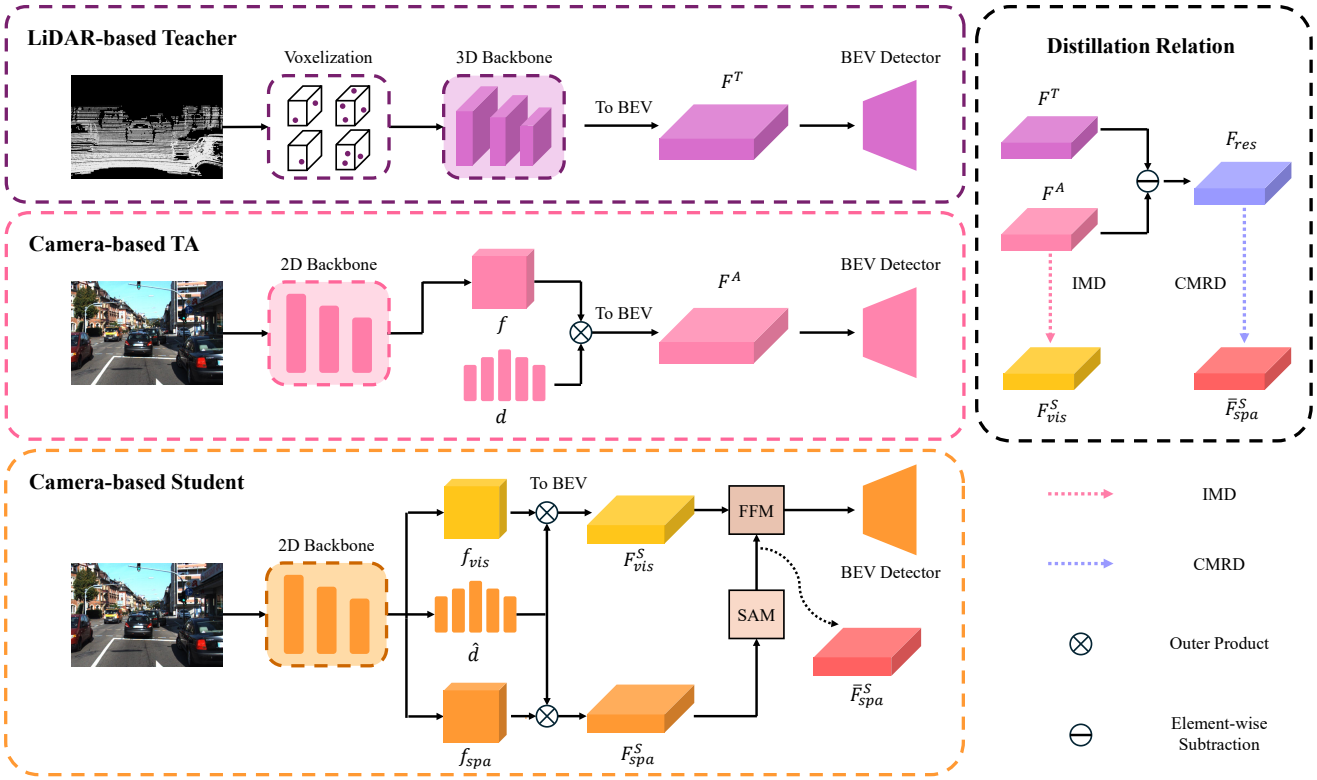

Figure 2. Overall architecture of the MonoTAKD . The top, middle, and bottom rows show the architecture of the LiDAR-based teacher, the camera-based teaching assistant (TA), and a camera-based student. We design the intra-modal distillation (IMD) and the cross-modal residual distillation (CMRD) processes to guide the camera-based student. In addition, a spatial alignment module (SAM) and a feature fusion module (FFM) are employed to improve the BEV feature representation.

3.4. Camera-based Student

The goal in designing the student model is to learn 3D scene geometry from more powerful models. Based on this perspective, we adopt a camera-based model that closely mirrors $\mathcal{A}$. Specifically, we use the same 2D backbone to extract the visual feature and incorporate an off-the-shelf depth estimator to predict depth map $\hat{d}$. After processing the image through the 2D backbone, $\mathcal{S}$ employs two separate branches to generate distinct visual features, denoted as f_{vis} and f_{spa}, with different weights. f_{vis} focuses on learning 3D visual knowledge from $\mathcal{A}$, while f_{spa} is responsible for learning 3D spatial cues from the relationship between $\mathcal{A}$ and $\mathcal{T}$, complementing the visual information. This dual-branch structure optimizes the student's capacity to understand both visual and spatial information crucial for robust 3D perception. Together with the predicted depth map $\hat{d}$, the paired representations go through the projection described in Eq. 2, resulting in BEV features F^S_{vis} and F^S_{spa}.

With the established BEV features (F^T, F^A, F^S_{vis}, F^S_{spa}) from the three models, the following sections detail the distillation, refinement, and fusion processes that drive comprehensive integration of 3D information across modalities.

Intra-modal Distillation. Since the camera-based $\mathcal{A}$ and $\mathcal{S}$ share the same modality, their feature representation gap is expected to be smaller than the gap observed in direct cross-modal distillation between $\mathcal{T}$ and $\mathcal{S}$. This similarity in representation makes intra-modal distillation (IMD) a more effective strategy for knowledge transfer. In our approach, we leverage the BEV features F^A as 3D visual knowledge and distill them into the student's BEV features F^S_{vis}.

The IMD process is optimized by minimizing the mean square error (MSE) between corresponding feature representations, which can be expressed as

$$\mathcal{L}_{IMD} = \text{MSE}(F^S_{vis}, F^A). \tag{3}$$

Spatial Alignment Module. To strengthen the student's ability to capture 3D information, we introduce a Spatial Alignment Module (SAM) to refine its BEV features. We denote $\bar{F}^S_{spa}$ as the enhanced BEV features obtained by processing F^S_{spa} through SAM. As illustrated in Fig. 3, SAM captures extensive global information and mitigates spatial feature misalignment through Atrous and Deformable convolutions. More precisely, Atrous Spatial Pyramid Pooling (ASPP) [4] expands the receptive fields using multiscale dilated convolutions, encouraging the model to capture richer

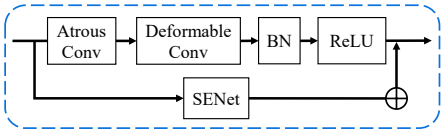

Figure 3. Spatial Alignment Module (SAM). SAM cascades the Atrous and Deformable convolutions to learn the alignment of BEV features. SENet is adopted for channel attention.

3D information. On the other hand, Deformable convolutions [52] adjust spatial offsets to address distortions in BEV features caused by inaccurate depth estimations, thus reducing spatial misalignment in feature representation. Additionally, we incorporate a SENet-like block [13] within SAM to recalibrate features adaptively along the channel dimension, enhancing the model's sensitivity to essential 3D spatial cues. This well-designed module allows SAM to enhance feature alignment and effectively integrate valuable spatial information into the student's BEV features.

Cross-modal Residual Distillation. While we designate the BEV features of the TA model as the learning target for the student model in Eq. 3, they still fall short of fully capturing the 3D spatial richness provided by LiDAR. Although incorporating GT depth maps into the TA model narrows such a gap, a discrepancy remains due to the lack of precise 3D spatial cues intrinsic to LiDAR data. Unlike previous cross-modal distillation methods, which require the student model to mimic the raw features from the teacher, our MonoTAKD identifies the missing 3D spatial cues as residual features, providing an additional learning objective. Particularly, the distillation of residual features covers vital spatial information and enhances the 3D perception of the student model. The loss function can be defined as

$$\mathcal{L}_{CMRD} = \text{MSE}(\bar{F}^S_{spa}, F_{res}), \tag{4}$$

where $F_{res} = F^T \ominus F^A$ represents the residual features. We note that F_{res} is calculated as the difference of BEV features between $\mathcal{T}$ and $\mathcal{A}$ by element-wise subtraction and is binarized by a predefined threshold to suppress background noise and emphasize essential spatial regions (see Fig. 4).

Feature Fusion Module. To aggregate two BEV features from the student model, *i.e.*, F^S_{vis} and $\bar{F}^S_{spa}$, we apply a Feature Fusion Module (FFM) to integrate their expertise in both visual and spatial 3D information. Specifically, we begin by fusing features F^S_{vis} and $\bar{F}^S_{spa}$ (the enhanced feature by SAM) using element-wise addition. This combined representation is then processed through two convolutional layers. The fusion process can be simplified as

$$F^S = \text{FFM}(F^S_{vis} \oplus \bar{F}^S_{spa}). \tag{5}$$

The resulting fused feature F^S is subsequently fed into a BEV detector [20] for 3D object detection, generating the prediction for object localization and recognition.

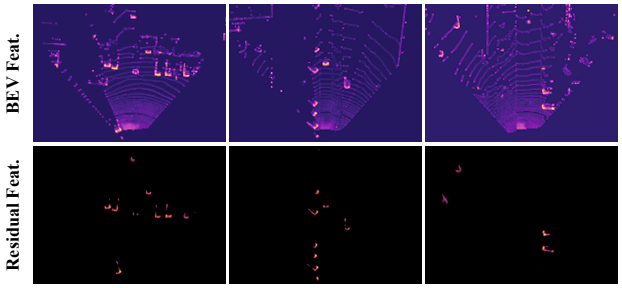

Figure 4. Comparing the BEV features (F^T) from the teacher model with the residual features (F_{res}) reveals that our residuals effectively capture essential 3D spatial cues, emphasizing critical information over less important elements, such as ripples and background noise present in F^T.

3.5. Loss Function

Our proposed MonoTAKD framework is end-to-end trained with an objective function $\mathcal{L}_{total}$ that combines multiple components to effectively guide KD. This total loss function can be expressed as

$$\mathcal{L}_{total} = \mathcal{L}_{IMD} + \mathcal{L}_{CMRD} + \mathcal{L}_{logit}, \tag{6}$$

where the first two terms are introduced in Eq. 3 and Eq. 4. The last term in Eq. 6 is logit distillation for the student model, formulated as

$$\mathcal{L}_{logit} = \mathcal{L}_{cls} + \mathcal{L}_{reg}, \tag{7}$$

where the former term represents the classification loss accomplished by quality focal loss [21]. The latter utilizes the Smooth L1 loss for the bounding box regression for precise localization. Both classification and regression losses are computed over the predictions from the teacher model.

4. Experiments

4.1. Datasets and Evaluation Metrics

KITTI3D. The KITTI 3D detection dataset [11] is a widely used benchmark in the field of autonomous driving. It consists of 7,481 training images and 7,518 testing images with synchronized LiDAR point clouds.Following the data split provided by [5], the training images are further split into two groups: 3,712 images and 3,769 images as the train and val sets, respectively. We report the 3D detection performance AP_{3D} and the BEV detection performance AP_{BEV} with 40 recall positions [35]. Our detection results include three difficulty levels: easy, moderate, and hard.

nuScenes. The nuScenes dataset [3] includes 1,000 multimodal video sequences, 700 for training, 150 for validation, and 150 for testing. It offers synchronized sensor data

Table 1. Experimental results on the KITTI *test* set for the Car category. We use **bold** and underline to indicate the best and the second-best results, respectively. † denotes the KD-based methods.

Method	Venue	Extra Data	AP_{3D}			AP_{BEV}		
			Easy	Mod.	Hard	Easy	Mod.	Hard
DDMP-3D [36]	CVPR 21	Pre-trained Depth	19.71	12.78	9.80	28.08	17.89	13.44
DD3D [27]	ICCV 21		23.22	16.34	14.20	30.98	22.56	20.03
Kinematic3D [1]	ECCV 20	Temporal	19.07	12.72	9.17	26.69	17.52	13.10
Dfm [1]	ECCV 22		22.94	16.82	14.65	31.71	22.89	19.97
CaDDN [32]	CVPR 21	LiDAR Auxiliary	19.17	13.41	11.46	27.94	18.91	17.19
MonoDTR [14]	CVPR 22		21.99	15.39	12.73	28.59	20.38	17.14
MonoNerd [43]	ICCV 23		22.75	17.13	15.63	31.13	23.46	20.97
MonoPGC [41]	ICRA 23		24.68	17.17	14.14	32.50	23.14	20.30
OccupancyM3D [30]	CVPR 24		25.55	17.02	14.79	35.38	24.18	21.37
MonoDistill† [7]	ICLR 22	None	22.97	16.03	13.60	31.87	22.59	19.72
Cube R-CNN [2]	CVPR 23		23.59	15.01	12.56	31.70	21.20	18.43
MonoUNI [17]	NeurIPS 23		24.75	16.73	13.49	33.28	23.05	19.39
MonoATT [51]	CVPR 23		24.72	17.37	15.00	36.87	24.42	21.88
MonoDETR [49]	ICCV 23		25.00	16.47	13.58	33.60	22.11	18.60
CMKD† [12]	ECCV 22		25.09	16.99	15.30	33.69	23.10	20.67
ADD† [42]	AAAI 23		25.61	16.81	13.79	35.20	23.58	20.08
MonoCD [44]	CVPR 24		25.53	16.59	14.53	33.41	22.81	19.57
MonoTAKD	-	None	**27.91**	**19.43**	**16.51**	**38.75**	**27.76**	**24.14**

Table 2. Experimental results on the nuScenes *val* set. We use **bold** to indicate the best results in each setting.

Method	Modality	Backbone	NDS↑	mAP↑
PETRv2 [25]	C	R50	0.456	0.349
P2D [18]	C	R50	0.474	0.360
PGD [38]	C	R101	0.428	0.369
MonoDETR [49]	C	R101	0.526	0.423
BEVFormer [23]	C	R50	0.423	0.352
+BEVDistill [6]	$L \rightarrow C$	R50	0.457	0.386
+DistillBEV [39]	$L \rightarrow C$	R50	0.476	0.367
+STXD [16]	$L \rightarrow C$	R50	0.481	0.374
+TAKD (Ours)	$L \rightarrow C$	R50	**0.490**	**0.392**
BEVFormer	C	R101	0.445	0.374
+BEVDistill	$L \rightarrow C$	R101	0.468	0.389
+DistillBEV	$L \rightarrow C$	R101	0.545	0.446
+STXD	$L \rightarrow C$	R101	0.543	0.440
+TAKD (Ours)	$L \rightarrow C$	R101	**0.558**	**0.451**
BEVDepth [22]	C	R50	0.440	0.317
+STXD	$L \rightarrow C$	R50	0.483	0.371
+DistillBEV	$L \rightarrow C$	R50	0.510	0.403
+LabelDistill [19]	$L \rightarrow C$	R50	0.528	0.419
+TAKD (Ours)	$L \rightarrow C$	R50	**0.537**	**0.430**
BEVDepth	C	R101	0.535	0.412
+DistillBEV	$L \rightarrow C$	R101	0.547	0.450
+LabelDistill	$L \rightarrow C$	R101	0.553	0.451
+TAKD (Ours)	$L \rightarrow C$	R101	**0.564**	**0.466**

streams collected with 6 cameras and a 32-beam LiDAR sampled at 20Hz, covering the 360-degree field of view. We utilize nuScenes Detection Score (NDS) and the mean Average Precision (mAP) as the primary evaluation metrics. Due to space limits, other dataset descriptions and implementation details are provided in the supplementary material.

4.2. Main Results

Results on KITTI3D. We demonstrate the effectiveness of our MonoTAKD on the KITTI test set for the Car category in Table 1. First, MonoTAKD outperforms the second-best result by +2.30, +2.06, and +1.21 in AP_{3D}, by +1.88, +3.34, +2.26 in AP_{BEV} under easy, moderate, and hard difficulty levels, respectively. When compared to CMKD, the top-performing KD-based method, MonoTAKD, consistently surpasses CMKD, achieving +2.82, +2.44, +1.21 improvements in AP_{3D} across all difficulty levels. Moreover, we observe a significant performance gain over all proceeding depth-guided methods. For instance, with respect to MonoDETR, our model achieves a notable increase of +2.93 (21.58%) in AP_{3D} under the hard level.

Our successful performance lies in incorporating intra-modal distillation (IMD) and cross-modal residual distillation (CMRD), which provide the student model with rich 3D visual knowledge and crucial 3D spatial cues. This combined approach enhances both the semantic information and understanding of 3D scene geometry for the student model. Note that the results for other classes (Pedestrian and Cyclist) are provided in the supplementary material.

Results on nuScenes. We further evaluate MonoTAKD on the nuScenes val set, as shown in Table 2. For consistent and effective distillation, we use CenterPoint as the LiDAR-based teacher and employ CaDDN for both the TA and student models, as validated in Table 3.

Our results show significant improvements in NDS and mAP for both BEVFormer and BEVDepth. This reveals the potential to scale our feature distillation losses from monocular to multi-view camera applications, which showcases the generalizability of our approach. Additionally, we ob-

Table 3. Compared the effect of distillation losses using various teacher, TA, and student models. * indicates the insertion of the ground truth depth. We use **bold** and underline to indicate the best and the second-best results, respectively.

Model Types			AP_{3D}			AP_{BEV}		
Teacher model	TA model	Student model	Easy	Mod.	Hard	Easy	Mod.	Hard
None	None	CaDDN	23.57	16.31	13.84	30.28	21.53	18.90
CenterPoint [48]	CaDDN*	CaDDN	27.06	19.38	17.50	35.66	26.03	23.09
PointPillar [20]	CaDDN*	CaDDN	31.28	20.80	17.58	42.21	28.48	25.73
Second [45]	CaDDN*	CaDDN	**34.36**	**22.61**	**19.88**	**42.86**	**29.41**	**26.47**
CenterPoint	MonoDETR*	CaDDN	26.56	18.84	16.46	34.62	25.85	22.29
PointPillar	MonoDETR*	CaDDN	28.11	20.00	17.24	40.21	27.10	24.65
Second	MonoDETR*	CaDDN	30.74	20.35	17.69	41.75	28.95	24.80
None	None	MonoDETR	28.84	20.61	16.38	37.86	26.95	22.80
CenterPoint	MonoDETR*	MonoDETR	30.78	21.17	18.41	39.73	27.22	24.69
PointPillar	MonoDETR*	MonoDETR	31.25	21.47	18.57	38.68	27.16	24.83
Second	MonoDETR*	MonoDETR	33.18	21.97	18.55	41.98	28.43	25.24

serve higher performance using BEVDepth as the student model due to its similarity to CenterPoint's dense prediction head, enhancing the effectiveness of logit distillation.

4.3. Ablation Study

For easier comparison with other methods, we conduct ablation studies on the KITTI val set for the Car category, using ResNet50 as the backbone for the TA and student models.

Generalization Study with Different Teacher and Teaching Assistant Models. We employ CaDDN and MonoDETR without distillation techniques as the baseline models in this experiment. The findings in Table 3 suggest utilizing the same model for the TA and student models. As such, the implementation is simpler and more generalizable than creating a custom TA model for each detector. In addition, we observe that distilling knowledge across heterogeneous architectures (transformer to CNN) may be ineffective due to substantial differences in feature encoding. Our study demonstrates that MonoTAKD consistently boosts performance under the guidance of various LiDAR-based teacher and camera-based TA models. Ultimately, Second is selected as our LiDAR-based teacher, with CaDDN serving as both the TA and student models.

Effectiveness of Different Feature Distillation. We present the effect of each feature distillation loss on MonoTAKD's performance in terms of AP_{3D}, as shown in Table 4. Setting 1 represents the baseline performance of the student without any distillation loss guidance. Cross-modal distillation (CMD) refers to the student learning directly from the LiDAR BEV features [12]. Settings 2 and 3 investigate the effectiveness of distillation between IMD and CMD. The student model performs better with IMD because the distillation is across a narrower feature representation gap than CMD. Settings 3 and 4 show that CMRD slightly outperforms CMD, although it does not provide the stable and robust 3D visual knowledge that IMD offers.

To further highlight the effect of residual features, we incorporate IMD to support the basic 3D perception and compare the results between settings 5 and 6. Our analysis reveals that focusing on learning 3D spatial cues (residual features) within the LiDAR BEV feature is more effective than exhaustively replicating all 3D information from the LiDAR modality. Specifically, the student model can prioritize accurately discerning the shape and position of foreground objects while minimizing attention to background noise, which could otherwise hinder its understanding of 3D scene geometry. As a result, setting 6 (MonoTAKD) outperforms baseline, setting 1, in AP_{3D} by +10.01, +6.45, and +6.44 for three difficulty levels, respectively. In our settings, the FFM is excluded from settings 1-4 since the BEV feature fusion is not required.

Table 4. Effectiveness of different feature distillation losses.

Settings	Loss			AP_{3D}		
	IMD	CMD	CMRD	Easy	Mod.	Hard
1				24.35	16.16	13.44
2	✓			31.11	20.24	16.91
3		✓		28.22	18.29	15.10
4			✓	29.68	19.57	16.22
5	✓	✓		30.73	19.88	16.43
6	✓		✓	**34.36**	**22.61**	**19.88**

Additionally, we report training curves for each feature distillation loss at the moderate level, as depicted in Fig. 5. When comparing the training curves between IMD and CMD, CMD consistently shows lower AP_{3D} values. This underscores the superior effectiveness of IMD over CMD, attributed to the narrower feature representation gap within the same modality.

Finally, our IMD+CMRD (MonoTAKD) remarkably outperforms IMD+CMD throughout the training process. This observation indicates that the student model greatly benefits from learning residual features rather than learning from the entirety of LiDAR BEV features, thereby enhancing the effectiveness of CMD. Notably, each feature distillation loss converges within approximately 60 epochs, demonstrating that the proposed feature distillation losses

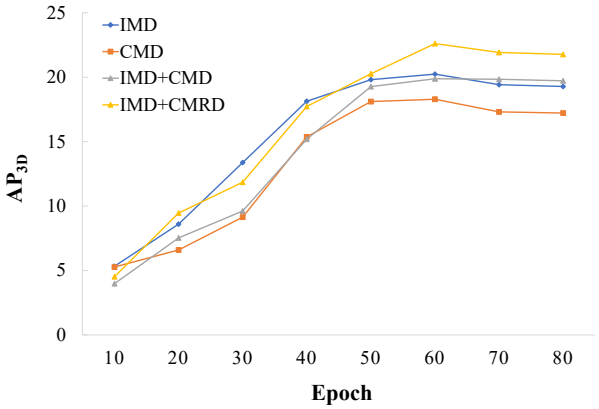

Figure 5. Convergence curves of different feature distillation on the KITTI *val* set. The x-axis shows the number of epochs, and the y-axis denotes AP_{3D} for the Car category at the moderate level.

Table 5. Ablation of the components of MonoTAKD. DConv. indicates the deformable convolution.

Components				AP_{3D}		
ASPP	DConv.	SENet	FFM	Easy	Mod.	Hard
				31.36	20.80	16.53
✓				32.28	21.45	17.18
	✓			32.46	21.51	17.99
✓	✓			33.19	21.81	18.23
✓	✓	✓		33.26	21.47	18.61
✓	✓	✓	✓	**34.36**	**22.61**	**19.88**

Table 6. Efficiency analysis of MonoTAKD.

Model	Parameters (M)	FLOPs (G)	AP_{3D}		
			Easy	Moderate	Hard
MonoNeRD [43]	83.0	356.57	20.64	15.44	13.99
DD3Dv2 [28]	80.3	163.00	26.23	21.21	18.83
MonoDETR [49]	47.4	57.51	28.84	20.61	16.38
CMKD [12]	**45.1**	**41.32**	23.53	16.33	14.44
MonoTAKD-Lite	**45.1**	**41.32**	31.36	20.80	16.53
MonoTAKD	47.8	44.90	**34.36**	**22.61**	**19.88**

and modules do not extend training time. Furthermore, all experiments can be conducted on a consumer-grade GPU with only 12GB VRAM, indicating that our method imposes minimal impact on training costs.

Effectiveness of the Proposed Modules. Table 5 illustrates the performance boost brought by each component of SAM and the FFM. The baseline is the MonoTAKD with IMD and CMRD but without SAM and FFM. In our experiment, FFM is designed to augment BEV feature fusion across different branches of the student model, whereas configurations without it default to element-wise addition.

With the combination of all components, MonoTAKD improves +3.00, +1.81, and +3.35 in AP_{3D} across three difficulty levels, highlighting the effectiveness of SAM in learning spatial shifts in both near and far regions, and FFM in integrating BEV features within the student model.

Efficiency Analysis. To further evaluate the practicality of the model at the application level, we tabulate the FLOPs and parameters of state-of-the-art Mono3D methods in Table 6. By leveraging the KD-based framework, CMKD reaches a substantial reduction in parameters and FLOPs compared to depth-guided methods [28, 49]. However, distilling knowledge across the large feature gap between teacher and student models results in a notable AP drop.

To ensure a fair comparison with CMKD, we remove both SAM and FFM modules from our student model, creating a lightweight version, MonoTAKD-lite, which matches the baseline in Table 5 and aligns with the CMKD (CaDDN) architecture. MonoTAKD-Lite effectively mitigates the performance drop while remaining a compact student model. Notably, even without the additional modules, our MonoTAKD-Lite demonstrates a clear performance gain over state-of-the-art approaches, offering higher AP_{3D} with reduced FLOPs and fewer parameters.

5. Discussion

To garner 3D perception from a 2D image, recent Mono3D methods can be categorized into three major strategies: (1) Extra training data. This method leverages LiDAR-generated auxiliary labels, such as depth maps [14, 41, 43], occupancy labels [30], temporal data [1, 16], to improve the training process. (2) Depth-guided. Using additional depth estimator [9, 37, 49, 51] or pre-training on additional depth datasets [27, 28] to improve depth estimation. (3) Cross-modal. Leveraging LiDAR-based teacher to camera-based student [12, 19, 39, 50]. Fair comparisons are challenged by heterogeneous experimental settings and varying model architectures. The lack of standardized efficiency metrics (FLOPs, parameter counts) makes it difficult to assess the true cost-benefit and trade-offs. We advocate for future work to report such metrics to enable fair comparisons across architectures and backbones.

6. Conclusion

In this paper, we propose a novel teaching assistant knowledge distillation for Mono3D (MonoTAKD) to improve the cross-modal distillation effectiveness. Specifically, we observe that the camera-based student model struggles to learn 3D information due to a significant feature representation gap between LiDAR and camera modalities. We alleviate this impediment by employing a camera-based teaching assistant model to deliver robust 3D visual knowledge through intra-modal distillation. To enhance the 3D perception of the student model, we formulate the LiDAR-exclusive 3D spatial cues as residual features and then distill them to the student model through cross-modal residual distillation. As a result, MonoTAKD establishes a new state-of-the-art benchmark on the KITTI 3D detection dataset. With its accurate yet cost-effective attributes, MonoTAKD poses a promising solution for autonomous driving applications.

7. Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan under Grants NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

References

[1] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *ECCV*, pages 135–152, 2020.

[2] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *CVPR*, pages 13154–13164, 2023.

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628, 2020.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4), 2018.

[5] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *NeurIPS*, 2015.

[6] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. In *ICLR*, 2023.

[7] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. In *ICLR*, 2022.

[8] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, pages 1201–1209, 2021.

[9] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *CVPRW*, 2020.

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018.

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32 (11):1231–1237, 2013.

[12] Yu Hong and et al. Cross-modality knowledge distillation network for monocular 3d object detection. In *ECCV*, pages 87–104, 2022.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[14] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, pages 4012–4021, 2022.

[15] Peixiang Huang, Li Liu, Renrui Zhang, Song Zhang, Xinli Xu, Baichao Wang, and Guoyi Liu. Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning. *arXiv preprint arXiv:2212.13979*, 2022.

[16] Sujin Jang, Dae Ung Jo, Sung Ju Hwang, Dongwook Lee, and Daehyun Ji. Stxd: Structural and temporal cross-modal distillation for multi-view 3d object detection. *NeurIPS*, 36, 2024.

[17] Jia Jinrang, Zhenjia Li, and Yifeng Shi. Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues. In *NeurIPS*, pages 11703–11715, 2023.

[18] Sanmin Kim, Youngseok Kim, In-Jae Lee, and Dongsuk Kum. Predict to detect: Prediction-guided 3d object detection using sequential images. In *ICCV*, pages 18057–18066, 2023.

[19] Sanmin Kim, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum. Labeldistill: Label-guided cross-modal knowledge distillation for camera-based 3d object detection. In *ECCV*, pages 19–37, 2024.

[20] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019.

[21] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *NeurIPS*, 33:21002–21012, 2020.

[22] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485, 2023.

[23] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18, 2022.

[24] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *AAAI*, pages 1810–1818, 2022.

[25] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *ICCV*, pages 3262–3272, 2023.

[26] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198, 2020.

[27] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, pages 3142–3152, 2021.

[28] Dennis Park, Jie Li, Dian Chen, Vitor Guizilini, and Adrien Gaidon. Depth is all you need for monocular 3d detection. In *ICRA*, pages 7024–7031, 2023.

[29] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. In *ECCV*, pages 123–139, 2022.

[30] Liang Peng, Junkai Xu, Haoran Cheng, Zheng Yang, Xiaopei Wu, Wei Qian, Wenxiao Wang, Boxi Wu, and Deng Cai. Learning occupancy for monocular 3d object detection. In *CVPR*, pages 10281–10292, 2024.

[31] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017.

[32] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, pages 8555–8564, 2021.

[33] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019.

[34] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 1–14, 2020.

[35] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel Lopez-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, pages 1991–1999, 2019.

[36] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, pages 454–463, 2021.

[37] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, pages 454–463, 2021.

[38] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *CoRL*, pages 1475–1485, 2021.

[39] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *ICCV*, pages 8637–8646, 2023.

[40] Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *CVPR*, pages 21653–21662, 2023.

[41] Zizhang Wu, Yuanzhu Gan, Lei Wang, Guilian Chen, and Jian Pu. Monopgc: Monocular 3d object detection with pixel geometry contexts. In *ICRA*, pages 4842–4849, 2023.

[42] Zizhang Wu, Yunzhe Wu, Jian Pu, Xianzhi Li, and Xiaoquan Wang. Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection. In *AAAI*, pages 2892–2900, 2023.

[43] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *ICCV*, pages 6814–6824, 2023.

[44] Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. Monocd: Monocular 3d object detection with complementary depths. In *CVPR*, pages 10248–10257, 2024.

[45] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[46] Lei Yang, Xinyu Zhang, Jun Li, Li Wang, Minghan Zhu, Chuang Zhang, and Huaping Liu. Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection. *TCSVT*, 33(11):6832–6844, 2023.

[47] Yiran Yang, Dongshuo Yin, Xuee Rong, Xian Sun, Wenhui Diao, and Xinming Li. Beyond the limitation of monocular 3d detector via knowledge distillation. In *ICCV*, pages 9077–9086, 2023.

[48] Tianwei Yin et al. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021.

[49] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *CVPR*, pages 9155–9166, 2023.

[50] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *CVPR*, pages 5116–5125, 2023.

[51] Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, and Minyi Guo. Monoatt: Online monocular 3d object detection with adaptive token transformer. In *CVPR*, pages 17493–17503, 2023.

[52] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019.