BEV-LGKD: A Unified LiDAR-Guided Knowledge Distillation Framework for BEV 3D Object Detection

Anonymous CVPR submission

Paper ID 1100

Abstract

Recently, Bird's-Eye-View (BEV) representation has gained increasing attention in multi-view 3D object detection, which has demonstrated promising applications in autonomous driving. Although multi-view camera systems can be deployed at low cost, the lack of depth information makes current approaches adopt large models for good performance. Therefore, it is essential to improve the efficiency of BEV 3D object detection. Knowledge Distillation (KD) is one of the most practical techniques to train efficient yet accurate models. However, BEV KD is still under-explored to the best of our knowledge. Different from image classification tasks, BEV 3D object detection approaches are more complicated and consist of several components. In this paper, we propose a unified framework named BEV-LGKD to transfer the knowledge in the teacher-student manner. However, directly applying the teacher-student paradigm to BEV features fails to achieve satisfying results due to heavy background information in RGB cameras. To solve this problem, we propose to leverage the localization advantage of LiDAR points. Specifically, we transform the LiDAR points to BEV space and generate the foreground mask and view-dependent mask for the teacher-student paradigm. It is to be noted that our method only uses LiDAR points to guide the KD between RGB models. As the quality of depth estimation is crucial for BEV perception, we further introduce depth distillation to our framework. Our unified framework is simple yet effective and achieves a significant performance boost. Code will be released.

1. Introduction

3D object detection is an essential computer vision technique with wide application scenarios such as autonomous driving. Recently, multi-view 3D object detection has gained increasing attention thanks to significant improvements in the results of Bird's-Eye-View (BEV) perception. As a common representation of surrounding scene, BEV can clearly present the location and scale of objects. Compared with multi-modal systems, multi-view camera systems can be deployed at low cost, while the lack of depth information makes current approaches adopt large models for good performance. Therefore, it is crucial to improve the efficiency of current approaches for practical deployment on vehicles.

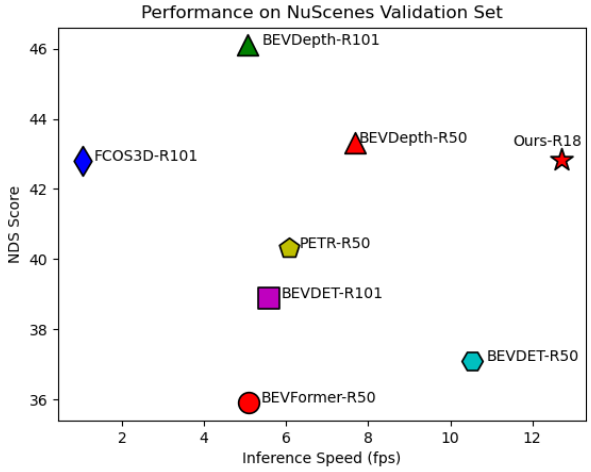


Figure 1. Inference speed of different methods on nuScenes val set. We test the inference speed on a single NVIDIA V100-32G GPU with the batch size of 1.

Knowledge Distillation (KD) is an effective method to train efficient yet accurate neural networks and has been extensively studied in fundamental tasks like image classification. [18] presents the well-known teacher-student paradigm, which forces the logits of a smaller network (student) to match the logits predicted by a large network (teacher). Many subsequent works follow this paradigm but match the hidden layer features as extra knowledge. However, different from image classification, BEV 3D object detection approaches are more complicated and consist of several components. Firstly, it is a multi-view system in

which each camera covers a certain field of view. Existing methods construct the BEV feature by aggregating the camera features, which contain heavy background information [20, 25, 28]. Secondly, as pointed out by [25], the quality of intermediate depth is the key to improving BEV 3D object detection.

To solve the problems, we propose a unified framework named BEV-LGKD based on the teacher-student paradigm. Compared with cameras, LiDAR points can capture precise 3D spatial information of foreground objects. Although many approaches have been proposed to explore the incorporation of LiDAR points for BEV 3D object detection, it should be noted that our method only uses LiDAR points to guide the KD between RGB models. The additional computational cost of our method is transforming the LiDAR points to BEV space, which is almost negligible compared with network training. We leverage the BEV LiDAR points and the camera parameters to obtain the foreground mask and view-dependent mask for teacher-student paradigm. The foreground mask can select the most informative regions for feature matching. The view-dependent mask exploits the characteristics of each view's feature. Therefore, the proposed BEV-LGKD can outperform other feature-based distillation methods and is especially suitable for BEV 3D object detection. We further observe that the quality of intermediate depth can be improved with a large depth estimation network. Therefore, we design a novel depth distillation loss to further improve the 3D object detection performance. Our unified framework is simple yet effective and achieve significant performance gains.

The contributions can be concluded as follows:

- We propose a unified BEV KD framework named BEV-LGKD that effectively leverages the LiDAR points to select informative foreground regions for BEV feature matching.
- We introduce a novel depth distillation loss to our framework, helping the student model obtain more accurate depth estimation results, further improving the whole KD performance.
- We conduct extensive experiments to evaluate the advantages of the proposed framework against other feature-based distillation methods.

2. Related work

Knowledge distillation Knowledge Distillation (KD) is invented to transfer useful knowledge from a large teacher model to a small student model. Early works concentrate on matching the logits of image classification via adjusting a temperature coefficient during training [11, 18]. The following methods extend the teacher-student paradigm to dense prediction tasks, demonstrating its effectiveness [6, 27, 40]. As for the object detection task, [6] proves the importance of feature learning during KD process. [14, 52] find that features of foreground and background regions contribute differently to the KD process. [22] proposes an instance-conditional framework to improve the KD performance. KD has been demonstrated the effectiveness on many vision tasks such as optical flow prediction [41], depth estimation [27, 34, 45] and semantic segmentation [17, 21, 27]. Although there are some KD methods for object detection, BEV KD is still under-explored to the best of our knowledge. The most related method is one ICLR 2023 blind submission, which studies the problem of cross-modal distillation. They match the features of LiDAR model and RGB model, while our method only uses LiDAR points to guide the KD between RGB models. The additional computational cost of our method is transforming the LiDAR points to BEV space, which is almost negligible compared with network training.

Vision-centric 3D object detection Vision-centric 3D object detection is useful for applications like autonomous driving since camera systems can be deployed at a low cost. FCOS3D [43] first decouples 3D targets into 2D and 3D attributes, and then predicts 3D objects by projecting the 3D center from 2D feature planes. PGD [42] analyzes the advantage of applying depth distribution to 3D monocular object detection. DETR3D [44] follows the seminal work of DETR [3], which uses object queries to match the position and class information of the instances. Recently, Bird's-Eye-View (BEV), as a unified representation of surrounding views, has attracted increasing attention from the community. BEVDet [20] utilizes the LSS operation [33] to transform 2D image features to 3D BEV feature. PersDet [53] improves the BEV feature generation and proposes the perspective BEV detection framework. PETR [28] introduces 3D positional embedding to obtain the 3D position-aware features. BEVDet4D [19] and PETRv2 [29] both fuse the multi-frame features using a spatial-temporal alignment operation and achieve significant performance improvement. BEVFormer [26] exploits the spatial and temporal cross-attention mechanism to query a BEV feature according to its position in BEV space. BEVDepth [25] finds that accurate depth estimation is essential for accurate BEV 3D object detection. BEVStereo [24] proposes to improve the depth estimation of camera-based systems by leveraging the temporal multi-view stereo (MVS) technology. They further design an iterative algorithm to update more valuable candidates, making it adaptive to moving candidates. STS [46] also tries to improve the depth estimation. They propose a novel technique that leverages the geometry correspondence to facilitate accurate depth learning. Although there are plenty of methods for BEV 3D object detection, BEV KD is under-explored as far as we know.

Depth estimation Since depth estimation is essential

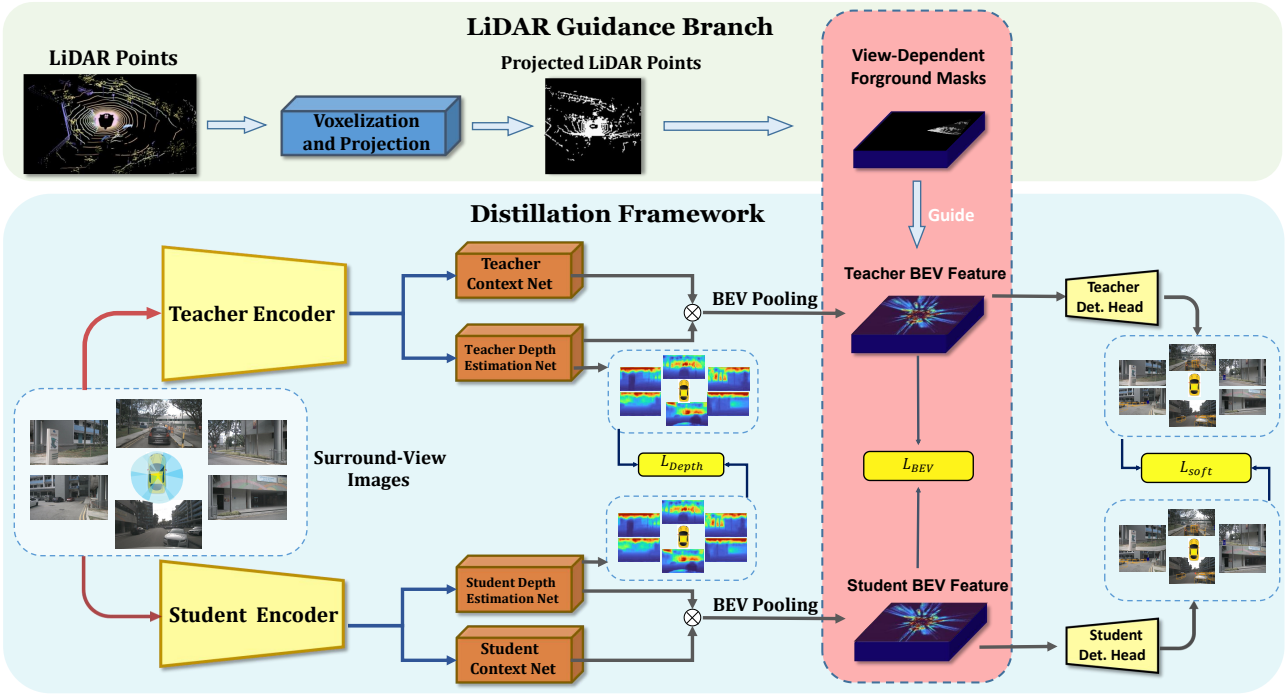


Figure 2. **The framework of the proposed LGKD method**. The framework of LGKD follows the teacher-student paradigm with RGB images as the inputs. It consists of three distillation components: LiDAR-Guided BEV Distillation, Depth Distillation and Soft-label Distillation.

for vision-centric 3D object detection, we introduce related methods on depth estimation. Monocular and stereo are two most typical ways of depth estimation. Monocular methods [1, 8–10, 12, 36] generally build an encoder-decoder architecture to regress the depth map from contextual features. Previous methods tend to either use a regression head to predict dense depth map [8, 35, 36] or use a classification head to predict a distribution along the depth range [1, 9]. Compared with monocular methods, stereo methods usually construct a cost volume to regress disparities based on photometric consistency [4, 15, 23, 32, 39, 51]. We conduct analysis to the depth estimation component of BEV 3D object detection and find that KD between depth estimation models can further improve the performance of BEV 3D object detection.

3. Method

This section provides a detailed introduction to the proposed BEV LiDAR-Guided Knowledge Distillation (BEV-LGKD) framework. Our goal is to transfer the knowledge from a large RGB teacher model to a small RGB student model. The overall framework is illustrated by Fig. 2. We use BEVDepth [25] as the baseline since it is a simple yet effective method. Our framework consists of three components: LiDAR-Guided BEV Distillation, Depth Distillation and Soft-label Distillation. We will introduce each component in the following sections.

Given an input multi-view image $I_k \in R^{3\times H\times W}$, BEVDepth adopts a shared backbone model to extract the feature $F_k \in R^{C\times H_f\times W_f}$, where k is the index of the camera. They also predict the depth distribution map for each input image $D_k \in R^{D\times H_f\times W_f}$. Then they project the camera features to viewing frustum $V_k \in R^{C\times D\times H_f\times W_f}$ and sum up the frustum features falling into the same flattened BEV grid $B \in R^{C\times H_e\times W_e}$. Finally, the task-specific heads are applied to the BEV feature. We first introduce how to distill the knowledge between BEV features.

3.1. LiDAR-guided BEV distillation

BEV is a unified representation of surrounding views, thus BEV 3D object detection has become prevailing in multi-view 3D object detection recently. In this paper, we explore the BEV feature KD to improve the performance of multi-view 3D object detection. According to recent 2D and 3D object detection works [5, 31, 49, 50], features close to object centers typically contain more useful information while features in background regions are less useful for KD. Therefore, directly applying the teacher-student KD paradigm to BEV feature fails to achieve satisfying results due to heavy background in BEV features. This is because BEV feature is aggregated from the multi-view RGB features, which contain heavy background information. To

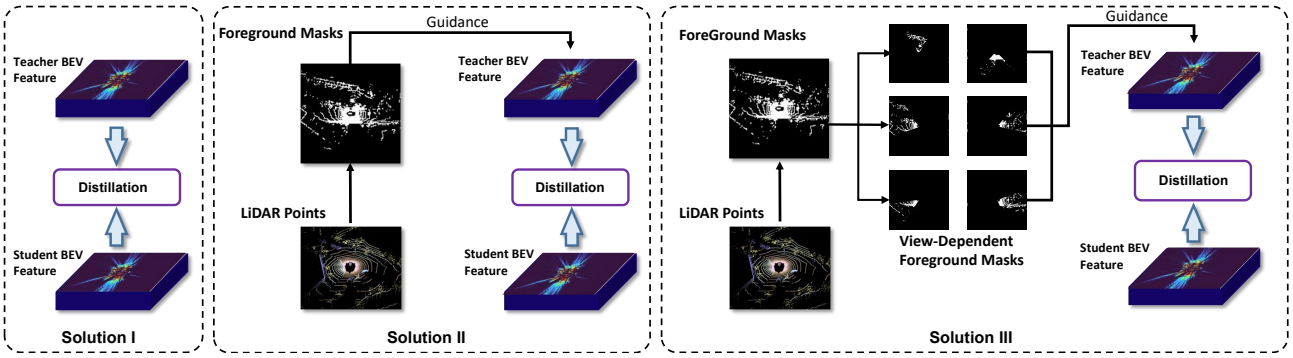


Figure 3. **Different solutions of BEV feature distillation**. Solution I is the direct distillation in the feature level. Solution II is the feature distillation guided by LiDAR foreground mask. Solution III is the feature distillation guided by both LiDAR foreground and view-dependent masks.

solve these limitations, we adopt the strong localization ability of the LiDAR sensor to help the KD between RGB models. LiDAR points can provide accurate location information of foreground objects. As illustrated by Fig. 3, we propose to adopt LiDAR points to distill useful information from teacher model to student model.

To be more specific, we first transform the LiDAR points to the BEV space and then voxelize [55] the projected LiDAR points to form a binary mask $M \in R^{1\times H_e\times W_e}$. The binary mask indicates the occupancy status of flattened BEV grid. We further perform a Gaussian smoothness to extend the localization information from isolated positions $M_g = g_\sigma(M)$. Since BEV feature is aggregated from multiple views, we split the foreground mask into multiple overlapped masks $\{M_g^k\}$ according to the view-dependent masks. The view-dependent masks are calculated by the camera field of view. Given the BEV features of teacher $B_t \in R^{C\times H_e\times W_e}$ and student $B_s \in R^{C\times H_e\times W_e}$. Our LiDAR-Guided BEV distillation loss is defined as follows.

$$L_{bev} = \sum_{k=1}^{K} \ell_2(M_g^k \odot B_t, M_g^k \odot B_s) \tag{1}$$

where ℓ_2 is the L2 distance between masked BEV features and K is the number of cameras.

3.2. Depth distillation

As pointed out by BEVDepth, accurate depth estimation is essential for the performance of BEV 3D object detection [25]. Therefore, we design two kinds of losses for depth distillation to further improve the KD performance. The coarse depth is defined as the distribution of predefined depths. The fine depth is defined as the regressed dense depth values.

Coarse depth loss Since the coarse depth is supervised by the Binary Cross-Entropy (BCE) loss as the classification tasks, we follow the well-know method [18] and define the coarse depth loss as follows.

$$L_{cd} = T^2 \sum_{k=1}^{K} \ell_{ce}(D_k^s/T, D_k^t/T) \tag{2}$$

where T is the temperature coefficient, and ℓ_{ce} is the Cross-Entropy loss.

Fine depth loss In order to enhance the depth distillation, we add a decoder ϕ to the context feature of BEVDepth [25] to regress the fine depth $\widetilde{D}_k = \phi(F_k) \in R^{H_f\times W_f}$. We use different structures for the decoders of teacher model and student model. For the fine depth, we use the L2 distance between the depth predictions of teacher and student:

$$L_{fd} = \sum_{k=1}^{K} \ell_2(\widetilde{D}_k^s, \widetilde{D}_k^t) \tag{3}$$

The final loss for depth distillation is defined as:

$$L_d = L_{cd} + \alpha L_{fd} \tag{4}$$

where α is the weight to balance coarse depth loss and fine depth loss. The depth distillation is illustrated in Fig. 2.

3.3. Soft-label distillation

For the detection task, we utilize a CenterNet [54] head to predict object locations and labels. Based on the BEV feature map $B \in R^{C\times H_e\times W_e}$, a heatmap would be regressed to represent the centers and categories in the BEV view. With a soft-label distillation loss $L_{soft} = \ell^d(O_s, O_t)$, the student model would have a faster convergence process under the guide of the teacher model.

3.4. Loss functions

In the phase of task learning, following BEVDepth [25], we use the multi-task loss of 3D object detection and depth estimation. Since this is not our main contribution, we refer the readers to BEVDepth [25]. In the distillation phase,

Table 1. 3D object detection results of different methods on nuScenes val set. We list the results of the state-of-the-art methods to make direct comparisons. We compare the results of DETR3D [44], FCOS3D [43], BEVDET [20], PETR [28], PGD [42], CenterNet [7], PersDet [53] and BEVDepth [25]. "C" in the modality column refers to the camera-only methods, and "L" refers to the Lidar-only methods. As can be seen, our method with a light-weight backbone (ResNet-18) can achieve competitive performance compared with BEVDepth (ResNet-50). Besides, BEV-LGKD achieves a significant performance boost compared with baseline.

Method	Image Size	Backbone	Modality	NDS ↑	mAP ↑	mATE ↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
PointPillars	-	-	L	0.597	0.487	0.315	0.260	0.368	0.323	0.203
CenterNet	-	DLA	C	0.328	0.306	0.716	0.264	0.609	1.426	0.658
FCOS3D	900 × 1600	ResNet-101	C	0.415	0.343	0.725	0.263	0.422	1.292	0.153
PGD	900 × 1600	ResNet-101	C	0.428	0.369	0.683	0.260	0.439	1.268	0.185
BEVDet	384 × 1056	ResNet-101	C	0.389	0.317	0.704	0.273	0.531	0.940	0.250
BEVDet	512 × 1408	Swin-T	C	0.417	0.349	0.637	0.269	0.490	0.914	0.268
PersDet	512 × 1408	ResNet-50	C	0.408	0.346	0.660	0.279	0.540	0.964	0.207
BEVDepth	256 × 708	ResNet-50	C	0.435	0.330	0.702	0.280	0.535	0.553	0.227
DETR3D	900 × 1600	ResNet-101	C	0.374	0.303	0.860	0.278	0.437	0.967	0.235
PETR	512 × 1408	ResNet-101	C	0.421	0.351	0.710	0.270	0.490	0.885	0.224
CrossDTR	512 × 1408	ResNet-101	C	0.426	0.370	0.773	0.269	0.482	0.866	0.203
BEVFormer-T	900 × 1600	ResNet-50	C	0.359	0.357	0.899	0.294	0.655	0.657	0.216
Ours (baseline)	256 × 708	**ResNet-18**	C	0.372	0.275	0.740	0.289	0.708	0.689	0.229
Ours (LGKD)	256 × 708	**ResNet-18**	C	0.425	0.305	0.701	0.273	0.560	0.500	0.215

the distillation loss is the weighted sum of our three components:

$$L = L_{soft} + \beta L_{bev} + \gamma L_d \tag{5}$$

the hyper-parameters are fixed during all our experiments.

4. Experiments

4.1. Settings

Datasets We use the nuScenes [2] dataset to evaluate the performance of our distillation framework. NuScenes contains 1k sequences, each of which is composed of six groups of surround-view camera images, one group of Lidar data and their sensor information. The camera images are collected with the resolution of 1600×900 at 12Hz and the LiDAR frequency for scanning is 20Hz. The dataset provides object annotations every 0.5 seconds, and the annotations include 3D bounding boxes for 10 classes {Car, Truck, Bus, Trailer, Construction vehicle, Pedestrian, Motorcycle, Bicycle, Barrier, Traffic cone }. We follow the official split that uses 750, 150, 150 sequences as training, validation and testing set respectively. So totally we get 28130 batches of data for training, 6019 batches for validation, and 6008 batches for testing.

Evaluation metric For the 3D object detection task, we use mean Average Precision(mAP) and Nuscenes Detection Score(NDS) as our main evaluation metrics. We also adopt other officially released metrics concluding Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). Note that NDS is a weighted sum of mAP and other metric scores. For the depth estimation task, we use Abs Relative difference (Abs_Rel) and prediction accuracy threshold ($\delta < 1.25$) as our main evaluation metrics. The other adopted metrics are Squared Relative difference (Sq_Rel), Relative Mean Square Error (RMSE) and RMSE_log.

Implementation details For both teacher and student models, we follow the pipeline of BEVDepth [25] while add the FCN decoders for depth distillation. The teacher model adopts ResNet-101 [16] or ResNet-50 as the heavy backbone, and the student model adopts ResNet-18 or MobileNetv2 as the light-weight backbone. The teacher model uses FPN with channel outputs {160, 160, 160, 160}, while the student model uses FPN with channel outputs {128, 128, 128, 128}. For BEV, we set the detection range to [-51.2m, 51.2m] along X and Y axis. The image scale for student model is 704×256. We adopt multi-frame fusion strategy that is proposed in BEVDepth [25]. We **do not** adopt CBGS [56] strategy in the experiments to gain extra improvements. In the optimization phase, we adopt Pytorch-lightning to compile the whole framework and use AdamW {weight_decay=1e-7} as the optimizer. We train the teacher model for 25 epochs with batch-size of 5 on 8 Nvidia Tesla V100 GPUs, and the student model with batch-size of 8 on 8 Nvidia Tesla V100 GPUs. The distillation process takes 35 epochs to finish.

Table 2. 3D object detection results of different methods on the nuScenes test set. We list the results of the state-of-the-art methods to make direct comparisons. We compare with the results of DETR3D [44], FCOS3D [43], BEVDET [20], PETR [28], PGD [42], CenterNet [7], "C" in the modality column refers to the camera-only methods, and "L" refers to the Lidar-only methods. As can be seen, our method with a light-weight backbone is still very competitive and BEV-LGKD also achieve a significant performance boost, demonstrating the generalization ability of BEV-LGKD.

Method	Image Size	Backbone	Modality	NDS ↑	mAP ↑	mATE ↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterNet	-	ResNet-101	C	0.400	0.338	0.658	0.255	0.629	1.629	0.142
FCOS3D	900 × 1600	ResNet-101	C	0.428	0.358	0.690	0.249	0.452	1.434	0.124
PGD	900 × 1600	ResNet-101	C	0.448	0.386	0.626	0.245	0.451	1.509	0.127
BEVDet	512 × 1408	Swin-S	C	0.463	0.398	0.556	0.239	0.414	0.101	0.153
DETR3D	900 × 1600	V2-99	C	0.479	0.412	0.641	0.255	0.394	0.845	0.133
PETR	512 × 1408	ResNet-101	C	0.455	0.391	0.647	0.251	0.433	0.933	0.143
Ours	256 × 708	**ResNet-18**	C	0.453	0.327	0.632	0.265	0.524	0.520	0.163

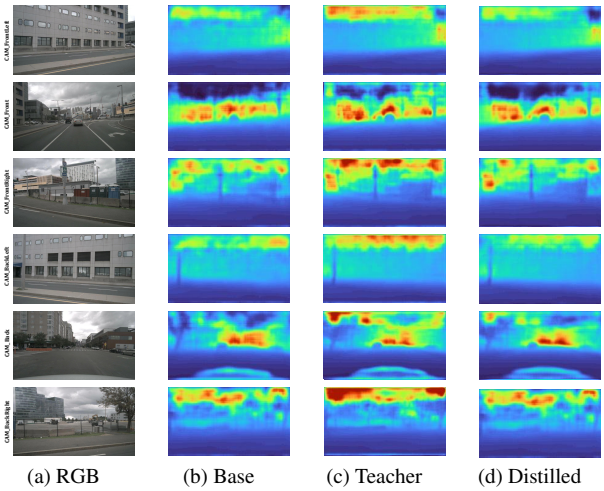
4.2. Comparison with the SOTA baselines

3D Object detection results on nuScenes validation set As shown in Tab. 1, we compare with the state-of-the-art methods on nuScenes val set. We only compare with methods based on RGB cameras since our method does not require RGB during inference. Since the camera-based methods lack depth information, a heavy backbone is usually required to achieve satisfying performance, leading to huge computational costs. In contrast, our method relies on a lightweight backbone and effectively transfers knowledge from a heavy backbone. Tab. 1 shows that our method with ResNet-18 can outperform typical monocular 3D object detection methods PGD [42], FOCOS3D [43] and CenterNet [7] with ResNet-101. For a fair comparison, we show the results of BEVDepth [25], which is also our baseline model. As can be seen, our method with ResNet-18 can achieve competitive performance with BEVDepth with ResNet-50. The gain of BEV-LGKD reaches **5.6%** on NDS compared with baseline, demonstrating the effectiveness of our method.

3D Object detection results on nuScenes test set We also report the results on nuScenes test set in Tab. 2. As can be seen, we also obtain remarkable results compared with the state-of-the-art methods. The proposed BEV-LGKD framework also achieves a significant performance boost, demonstrating the generalization ability of BEV-LGKD.

Depth estimation results on nuScenes val set We also compare the depth estimation on nuScenes val set. We compare our method with four state-of-the-art methods including Monodepth2 [10], FSM [13] and SurroundDepth [47]. Monodepth2 is self-supervised monocular depth estimation method while FSM and SurroundDepth are self-supervised multi-view depth estimation methods. The sparse ground truth is projected from LiDAR points by homographic warping. For the evaluation metrics, we get an average of results from 6 surround cameras. As shown in Tab. 3, our depth estimation results also have advantages on many metrics. Compared with monocular estimation methods like MonoDepth2, we obtain 11.3% decrease in Abs_Rel metric and 1.36 in Sq_Rel metric. We also outperform the multi-view methods like FSM. To make a more fair comparison, we calculate the depth estimation performance for individual cameras and compare it with Monodepth2 in Tab. 4. As can be seen, our method also outperforms the monocular methods and the gain of BEV-LGKD is significant.

Qualitative results We show the qualitative results for both 3D object detection and depth estimation tasks. As can be seen from Fig. 5 and Fig. 4, the proposed BEV-LGKD improves the performance of both 3D object detection and depth estimation.



(a) RGB (b) Base (c) Teacher (d) Distilled

Figure 4. Visualization of depth estimation results on nuScenes validation set. As shown in the figure, the depth distillation module improves the depth estimation performance with sharper edges and clearer shapes.

Table 3. Depth estimation results of different methods on nuScenes validation set. We compare with the results of FSM [13], SurroundDepth [47], Monodepth2 [10] . The depth estimation shares the backbone with 3D object detection. As can be seen, our method with ResNet-18 can outperform the results of Monodepth2 and SurroundDepth with ResNet-34 on the depth estimation task. The performance boost of BEV-LGKD is also significant. "S" refers to self-supervised learning methods and "F" refers to full-supervised learning methods.

Method	Backbone	Type	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta > 1.25$ ↑	$\delta > 1.25^2$ ↑	$\delta > 1.25^3$ ↑
Monodepth2	ResNet-34	S	0.303	2.675	9.210	0.437	0.506	0.764	0.881
FSM	ResNet-34	S	0.298	2.586	8.996	0.420	0.541	0.767	0.888
SurroudDepth	ResNet-34	S	0.245	3.067	6.835	0.321	0.719	0.878	0.935
Ours(teacher)	ResNet-101	F	0.147	1.055	4.595	0.219	0.829	0.931	0.969
Ours(student)	ResNet-18	F	0.190	1.411	5.447	0.263	0.761	0.896	0.957
Ours(base)	ResNet-18	F	0.204	1.603	5.892	0.274	0.737	0.880	0.921

4.3. Ablation study

Ablation study of each component We conduct experiments to verify the effectiveness of the distillation modules for 3D object detection on nuScenes val set. We design three distillation components to transfer reliable information from teacher model including Lidar-guided BEV Distillation, Depth Distillation and Soft-label Distillation. To evaluate the effectiveness of each component, we design experiments to evaluate the performance of each component. As shown in Tab. 5, each component contributes to the distillation process in terms of different metrics. For the NDS score, the soft-label distillation contributes 1.9% increase from 37.2% to 39.1%. The LiDAR-guided BEV distillation contributes most among the three components, improving the NDS score from **40.7%** to **42.5%** with a **1.8%** gain. For the mAP score, the soft-label distillation brings 1.5% improvements from 27.5% to 29.0% and the LiDAR-guided BEV distillation brings 1.1% improvements from 29.4% to 30.5%. We also provide the performance of the teacher model and our method can narrow the performance gap between the teacher model and student model on the NDS score from 9.9% to 4.6%.

Ablation study of LiDAR guidance Our method adopts the localization ability of LiDAR points to guide the knowledge distillation between RGB models. To validate the effectiveness of the proposed foreground and view-dependent masks, we further conduct an ablation study of LiDAR guidance. As shown in Tab. 6, both the foreground mask and view-dependent mask contribute to the distillation process. Foreground masks contribute most to the performance gain since they can filter out the background information in BEV features.

Ablation study of other student backbones In order

Table 4. Depth estimation results of individual camera on nuScenes validation set.

Method	Abs-Rel.↓ F_Left	Front	F_Right	B_Left	Back	B_Right	Std.↓
Monodepth2	0.304	0.214	0.388	0.314	0.304	0.438	0.078
FSM	0.287	0.186	0.375	0.296	0.221	0.418	0.088
Ours(teacher)	0.150	0.094	0.165	0.159	0.127	0.186	0.032
Ours(student)	0.195	0.123	0.214	0.205	0.164	0.233	0.039

Table 5. Ablation study of each component. As can be seen, the proposed three components contribute to the performance of 3D object detection. $L^{\dagger}_{Task}$ denotes all the task losses concluding detection and depth estimation, $L^{dete.}_{soft}$ is the soft label distillation loss, L^{depth}_{dis} is the depth distillation loss, and L_{bev} is the BEV feature distillation loss.

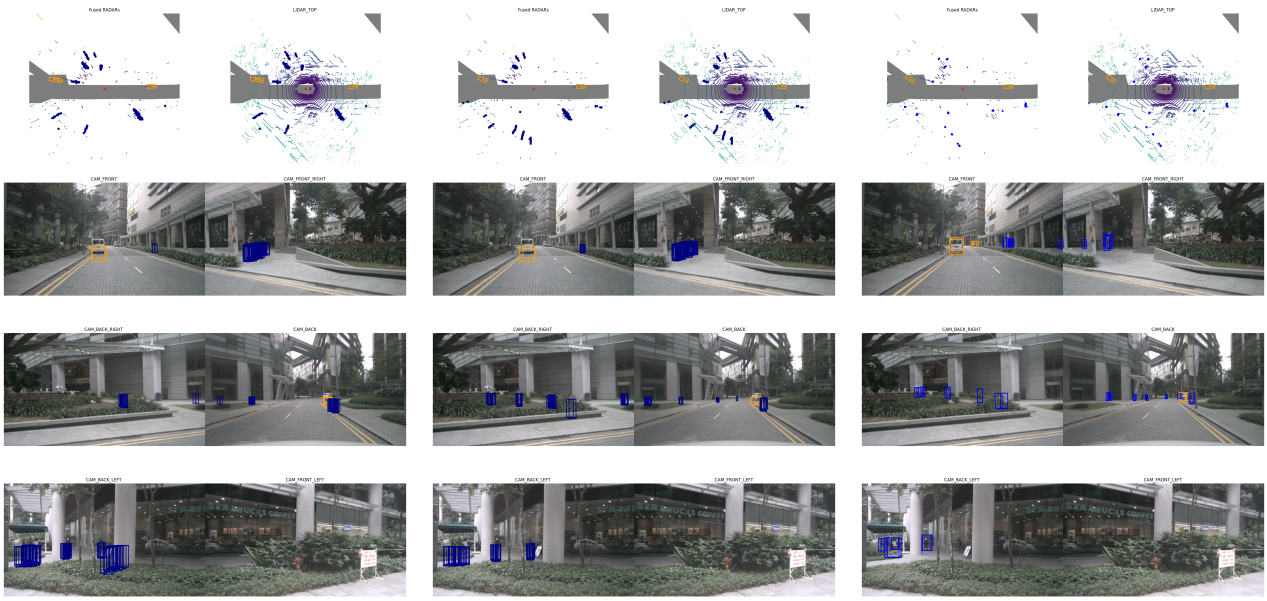
Phase	$L^{\dagger}_{Task}$	L_{soft}	L_{depth}	L_{bev}	NDS ↑	mAP ↑
Base (R18)	✓	-	-	-	0.372	0.275
Exp_1	✓	✓	-	-	0.391	0.290
Exp_2	✓	✓	✓	-	0.407	0.294
Exp_3	✓	✓	✓	✓	0.425	0.305
Exp_4	✓	-	✓	✓	0.416	0.301
Teacher (R101)	✓	-	-	-	0.471	0.350

Table 6. Ablation Study of LiDAR Guidance

Method	NDS↑	mAP↑	mATE↓	mAAE↓
Base-Direct	0.407(+ 3.5%)	0.294 (+ 1.5%)	0.720	0.225
Foreground	0.419(+ 4.7%)	0.304(+ 1.9%)	0.709	0.224
LGKD (Ours)	0.425(+ 5.3%)	0.305(+ 2.0%)	0.701	0.215

Table 7. Ablation Study of Other Distillation Methods

Method	NDS↑	mAP↑	mATE↓	mAAE↓
FitNet	0.389 (+ 1.7%)	0.283 (+ 0.8%)	0.742	0.228
LEKD	0.391(+ 1.9%)	0.288(+ 1.3%)	0.734	0.232
LGKD(Ours)	0.428(+ 5.3%)	0.305(+ 3.0%)	0.701	0.215
LEKD+LGKD	0.434(+ 6.2%)	0.310(+ 3.5%)	0.705	0.223



(a) Base (b) Distilled (c) Ground-Truth

Figure 5. Visualization results of 3D object detection on the nuScenes validation set. The student model (middle column) performs more accurate results than the un-distilled baseline (left column). We provide visualization for both surround camera views (bottom six images) and top-down LiDAR and RADAR views (top right and top left images).

Table 8. Ablation Study of Other Student Backbones.

Type	Backbone	NDS	mAP
Teacher	Res-101	0.471	0.350
Base$_M$	Mob.-v2	0.237	0.142
Student$_M$	Mob.-v2	0.275(+4.2%)	0.165 (+2.3%)

to evaluate the generalization of our method, we choose MobileNet-v2 [38] as another light-weight backbone. We maintain the overall architecture and hyper-parameters for a fair evaluation. As shown in Tab. 8, our method consistently improve the performance of 3D object detection using MobileNet-v2 as the light-weight backbone. The detailed record is listed in the supplementary file.

Ablation study of other distillation methods We also compare with other feature distillation methods to demonstrate the advantage of the proposed framework. We choose the feature distillation method FitNet [37] and LEKD [6] as the alternatives of our LiDAR-guided BEV distillation component. For a fair comparison, we keep the depth distillation and soft-label distillation. As can be seen from Tab. 7, the proposed LGKD method outperforms the feature distillation methods, demonstrating the advantage of BEV representation for multi-view perception tasks. We also conduct an experiment that combines feature distillation with BEV distillation and the additional margin is very small.

5. Discussions and limitations

In our work, we propose a novel distillation method to allow a lightweight model achieve remarkable performance for 3D object detection and depth estimation. However, our method still has some limitations. For example, we follow the pipeline of BEVDepth, which is sensitive to the accuracy of depth estimation. Besides, the performance on nuScenes test set is still not good enough compared with multi-modal methods [30, 48]. We will make attempts to further improve the performance.

6. Conclusion

To summarize, we propose a novel and unified framework named BEV-LGKD for BEV 3D object detection. Our framework consists of three components including LiDAR-Guided BEV Distillation, Depth Distillation and Soft-label Distillation. We leverage the localization ability of LiDAR points to generate the foreground and view-dependent masks, which effectively filter out the background information in BEV features. Our method only uses LiDAR data to guide the KD training and does not require LiDAR sensors during inference. Since depth estimation is essential for camera-based systems, we further introduce the depth distillation component to the framework, significantly improving the qualities of 3D object detection and depth estimation. We demonstrate the effectiveness of our method through extensive experiments.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 3

[2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 3

[5] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2703–2712, 2021. 3

[6] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 2, 8

[7] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 5, 6

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 3

[9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 3

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 3, 6, 7

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 2

[12] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 3

[13] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, 7(2):5397–5404, 2022. 6, 7

[14] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021. 2

[15] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[17] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019. 2

[18] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2, 4

[19] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2

[20] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 5, 6

[21] Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16876–16885, 2022. 2

[22] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, 34:16468–16480, 2021. 2

[23] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–590, 2018. 3

[24] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*, 2022. 2

[25] 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. *arXiv preprint arXiv:2206.10092*, 2022. 2, 3, 4, 5, 6

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera

images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 2

[27] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[28] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022. 2, 5, 6

[29] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 2

[30] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 8

[31] Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. On model calibration for long-tailed object detection and instance segmentation. *Advances in Neural Information Processing Systems*, 34:2529–2542, 2021. 3

[32] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022. 3

[33] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 2

[34] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019. 2

[35] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020. 3

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 3

[37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 8

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 8

[39] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 3

[40] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 2

[41] Hengli Wang, Peide Cai, Rui Fan, Yuxiang Sun, and Ming Liu. End-to-end interactive prediction and planning with optical flow distillation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2021. 2

[42] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2, 5, 6

[43] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 2, 5, 6

[44] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2, 5, 6

[45] Yiran Wang, Xingyi Li, Min Shi, Ke Xian, and Zhiguo Cao. Knowledge distillation for fast and accurate monocular depth estimation on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2457–2465, 2021. 2

[46] Zengran Wang, Chen Min, Zheng Ge, Yinhao Li, Zeming Li, Hongyu Yang, and Di Huang. Sts: Surround-view temporal stereo for multi-view 3d detection. *arXiv preprint arXiv:2208.10145*, 2022. 2

[47] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. *arXiv preprint arXiv:2204.03636*, 2022. 6, 7

[48] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *arXiv preprint arXiv:2208.11112*, 2022. 8

[49] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 3

[50] Yufei Yin, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Fi-wsod: Foreground information guided weakly supervised object detection. *IEEE Transactions on Multimedia*, 2022. 3

[51] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 185–194, 2019. 3

[52] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 2

[53] Hongyu Zhou, Zheng Ge, Weixin Mao, and Zeming Li. Persdet: Monocular 3d detection in perspective bird's-eye-view. *arXiv preprint arXiv:2208.09394*, 2022. 2, 5

[54] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 4

[55] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 4

[56] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 5