Distilling Focal Knowledge from Imperfect Expert for 3D Object Detection

Jia Zeng[1,2] Li Chen[1] Hanming Deng[3] Lewei Lu[3]
Junchi Yan[2,1] Yu Qiao[1] Hongyang Li[1,2*]
[1]OpenDriveLab, Shanghai AI Lab [2]Shanghai Jiao Tong University [3]SenseTime Research
{zengjia,lichen,qiaoyu,lihongyang}@pjlab.org.cn
{denghanming,luotto}@sensetime.com yanjunchi@sjtu.edu.cn

Abstract

Multi-camera 3D object detection blossoms in recent years and most of state-of-the-art methods are built up on the bird's-eye-view (BEV) representations. Albeit remarkable performance, these works suffer from low efficiency. Typically, knowledge distillation can be used for model compression. However, due to unclear 3D geometry reasoning, expert features usually contain some noisy and confusing areas. In this work, we investigate on how to distill the knowledge from an imperfect expert. We propose FD3D, a Focal Distiller for 3D object detection. Specifically, a set of queries are leveraged to locate the instance-level areas for masked feature generation, to intensify feature representation ability in these areas. Moreover, these queries search out the representative fine-grained positions for refined distillation. We verify the effectiveness of our method by applying it to two popular detection models, BEVFormer and DETR3D. The results demonstrate that our method achieves improvements of 4.07 and 3.17 points respectively in terms of NDS metric on nuScenes benchmark. Code is hosted at https://github.com/OpenPerceptionX/BEVPerception-Survey-Recipe.

1. Introduction

Accurate 3D object detection is a vital component in autonomous driving. To achieve this, most methods [14, 37] resort to LiDAR sensors and dominate the public benchmarks [1, 27]. Despite the performance gap, pure vision approaches are still worthy of in-depth inquiry, since cameras can provide rich semantic information and are low-cost and easy-to-deploy. Among these, bird's-eye-view (BEV) detection has drawn extensive attention from both industry and academia, and shown great potential to narrow down the performance gap [15, 21]. However, such models tend to be computationally consuming.

*Corresponding author at lihongyang@pjlab.org.cn

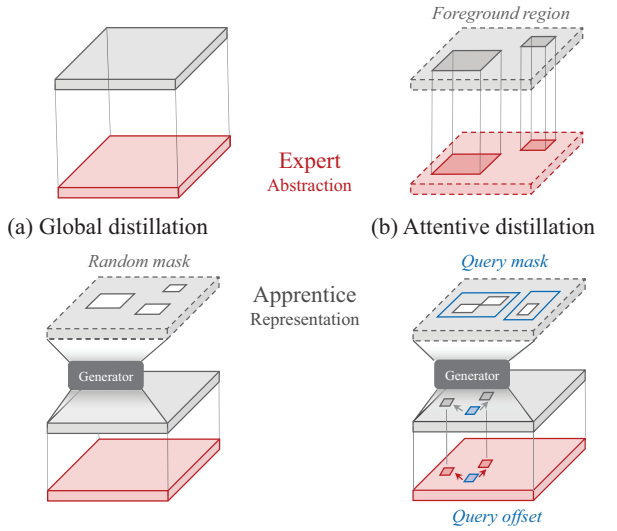

Figure 1. Illustration of the proposed generative and focal distillation method. Compared with others, the proposed manner in 1 (d) leverages queries to generate instance masks for masked generative distillation, rather than random masks in 1 (c). Moreover, the queries meanwhile search for the representative position to perform refined distillation, where the distillation region selection is more fine-grained and flexible than 1 (b).

In common practice, knowledge distillation can compress the model and is usually applied to alleviate computation overhead. One possible solution is to utilize the LiDAR-based model as the expert [4, 17], but this requires complex spatial-temporal calibrations and also needs to handle heterogeneous problems from different modalities. An intuitive question is, can we distill these models solely based on camera sensors? In this work, we intend to address this problem and focus on the camera-only distillation setting. To the best of our knowledge, our work is the first solution tailored for this setting.

Distillation methods in 2D object detection have derived

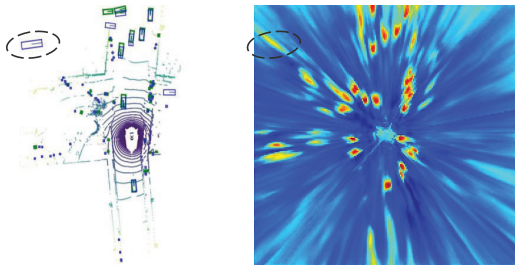

Figure 2. Visualization of the predicted bounding boxes and bird's eye view feature of BEVFormer [19]. At the center is the autonomous vehicle. In the left subfigure, green and blue boxes denote the ground truth and predictions respectively. In the right subfigure, the BEV feature is ray-shaped and contains a lot of noise. Areas with incorrect high activation appear behind objects due to occlusion, which easily introduces false positives, *e.g.*, the region circled by the ellipse.

various types as depicted in Fig. 1 (a)-(c), but their effectiveness is not verified in 3D object detection. The main challenge in camera-to-camera distillation for 3D object detection comes from the imperfect expert features. Due to the lack of accurate 3D information, expert features drawn from 2D images usually contain some noisy and confusing areas. To better illustrate this point, we visualize the BEV features from BEVFormer [19] in Fig. 2. The unclear occlusion reasoning makes the BEV features suffer from the ray-shape artifacts. The imperfect BEV features result in many false positives. Directly mimicking these features from experts such as Fig. 1 (a) may exacerbate the drawback. Some 2D detection distillation methods [29] propose to focus on foreground-orient regions, as shown in Fig. 1 (b). However, background regions are also important as proved in [7, 35, 13]. We categorize these methods as attentive distillation. Compared with 2D object detection, the imbalance between foreground and background in 3D areas is much more severe. Balancing these weights is not a general solution. The latest study MGD [36] demonstrates the effectiveness of masked image modeling (MIM) distillation as depicted in Fig. 1 (c). However, the global random mask cannot greatly enhance 3D object detection, which is validated in Tab. 4a.

To this end, we propose a Focal Distiller for 3D object detection, shortened as **FD3D**. The schematic diagram of FD3D is shown in Fig. 1 (d). Specifically, A set of queries are leveraged to locate the instance-level focal regions, masked generative distillation is performed within the regions. Moreover, these queries dynamically search fine-grained representative positions for focal distillation. Two complementary modules guide the apprentice network to generate enhanced feature representation on focal regions. In summary, our work makes three-fold contributions:

1. To the best of our knowledge, this is the first work to explore knowledge distillation on the camera-only 3D object detection. We reveal the challenge relies on how to distill focal knowledge from an imperfect 3D object detector expert.

2. We propose FD3D, which utilizes a set of queries to distill focal knowledge. With these queries, coarse-grained focal regions are selected for masked generation, and fine-grained focal regions are searched out for instance-oriented refinement distillation.

3. FD3D serves as a plug-and-play module. It can be easily extended to various detectors. The improvements with 4.07 and 3.17 NDS can be obtained with FD3D assembled in BEVFormer and DETR3D, respectively.

2. Related Work

Knowledge distillation on 2D object detection. Knowledge distillation is originally proposed by Hinton *et al.* [9]. It regards teachers' output logits as knowledge and is employed as a model compression approach in the classification task. FitNet [26] extends the knowledge to intermediate features. Recently, many works successfully apply knowledge distillation on object detection [3]. Knowledge distillation on object detection encounters extreme imbalance between positive and negative instances. Global feature imitation [29] such as Fig. 1 (a) brings marginal improvement, and even exacerbates the performance. It is almost a consensus that distillation should not treat all regions equally. Revisiting from the latest works, where to perform distillation is of great concern in distilling object detectors, and they invariably resort to attentive distillation as depicted in Fig. 1 (b). Mimic [16] performs distillation on ROI regions sampled by an RPN network, but this framework can only be applied to a two-stage detector. FIFG [29] claims the distillation should be performed on near-object regions. However, the near-object regions are defined by hand-crafted rules and require manual refinement. GID [6] focuses on regions where there is a significant difference between the student and the teacher outputs. DeFeat [7] separates foreground and background by ground truth and performs distillation independently. GID [6] and DeFeat [7] prove that the informative cues in the background also benefit the student network. FGD [35] designs a more elaborate structure to focus on critical pixels and channels. Albeit the effectiveness of the aforementioned studies, they require manual hyper-parameter tuning to adjust the concentration of distillation. ICD [13] encodes the instance annotations as queries, and attention maps are generated to guide the distillation region concentration. We deem such a query-based approach is more flexible. MGD [36] proposes a simple but

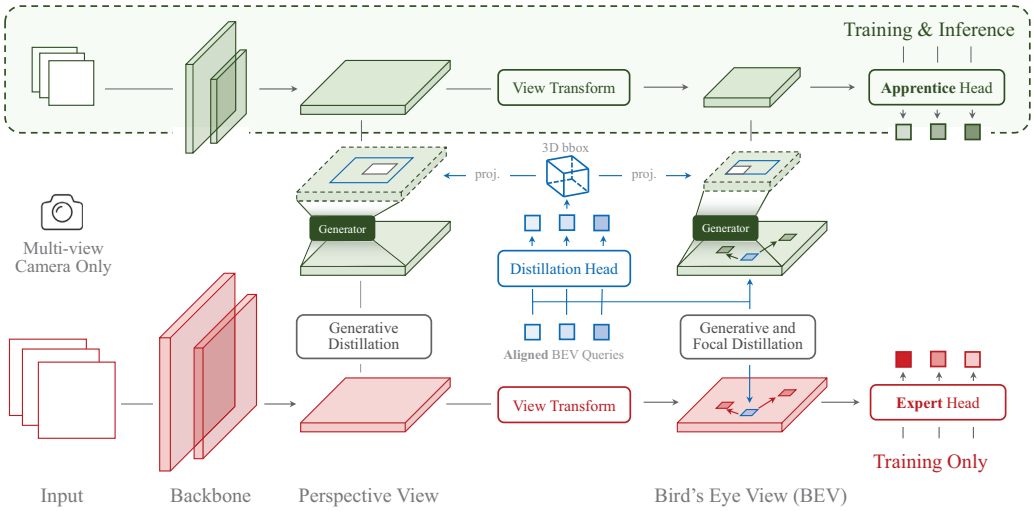

Figure 3. We present **FD3D**, a Focal Distiller for 3D object detection. The framework aims to transfer the dark knowledge of the focal region from the expert network to the compact apprentice network, where the feature distillation is performed on both perspective view (PV) and bird's-eye-view (BEV). We employ aligned BEV queries. With these queries, the instance masks in PV and BEV are generated, and masked generative distillation is performed within the mask region to enhance the feature representation. Moreover, these queries meanwhile leverage deformable attention mechanism to search out the representative features on focal BEV regions. The representative features are utilized for refined distillation.

effective distillation paradigm as shown in Fig. 1 (c): the student features are randomly masked and then recovered to mimic teacher features.

Multi-camera 3D Object Detection. Early studies [22, 30] independently process monocular images from multiple views, and aggregate the results across cameras for post-processing. FCOS3D [30] extends the popular anchor-free detector FCOS [28] by directly predicting 3D bounding boxes from 2D image features. Recent studies tend to predict object location in 3D space. One branch of researchers converts 2D image features to 3D space through depth estimation. The typical approach is to convert the estimated depth map into pseudo-LiDAR [31]. Lift-Splat-Shoot (LSS) [23] estimates the depth distribution of each image pixel and uses voxel pooling to generate BEV features. Its follow-up studies BEVDet [11], BEVDet4D [10], BEVDepth [18] further explore data augmentation, accelerating voxel pooling module and improving depth estimation accuracy. Another branch of studies indexes 2D image feature to 3D space through the camera extrinsic and intrinsic. Following DETR [2], DETR3D [32] defines a set of 3D object queries and samples 2D image features on the queries' projection points to predict 3D bounding boxes. BEVFormer [19] explicitly defines queries in BEV space, and performs 2D-to-3D transformation with deformable attention to generate the BEV feature for downstream tasks. PolarFormer [12] adopts polar coordinate system for more accurate BEV feature construction. These state-of-the-art models suffer from high computation overhead. It is natural to consider applying knowledge distillation to compress the model. However, regardless of the view transform approaches, the 2D-to-3D transformation is ill-posed, so the vision-based 3D object detector experts cannot provide a reliable feature for imitation.

LiDAR-based knowledge distillation. In 3D object detection, there are existing studies covering knowledge distillation under LiDAR [33, 34] and extending to LiDAR-camera cross-modality knowledge transfer. Monodistill [4] projects the LiDAR points to the image plane, aligning two modalities for knowledge transfer. UVTR [17] represents the image and LiDAR in a unified voxel space, and the feature distillation is performed on the query projection points. To exclude the heterogeneous problems from different modalities and focus on the exploration of distillation from an imperfect expect, the extent of this study is set at camera-only distillation.

3. Methodology

Fig. 3 illustrates the pipeline of our distillation approach. The overall structure of FD3D is presented in Sec. 3.1. The two proposed distillation modules are elaborated in Sec. 3.2 and Sec. 3.3.

3.1. Overall Structure

The state-of-the-art multi-camera 3D object detection networks suffer from high computation overhead. This shortcoming mainly originates from the heavy network ar-

Method	Backbone	Image Resolution	BEV Resolution	Encoder Layer
BEVFormer -Base* (E)	R101	900×1600	200×200	6
BEVFormer -Tiny (A)	R50	450×800	100×100	3
BEVFormer -Base (E)	R101-DCN†	900×1600	200×200	6
BEVFormer -Small (A)	R101-DCN†	450×800	100×100	3
DETR3D -R101 (E)	R101-DCN†	900×1600	-	-
DETR3D -R50 (A)	R50	900×1600	-	-

Table 1. The setting of expert-apprentice pairs. The symbol "E" and "A" denote "expert" and "apprentice" respectively. By settng different network depth, input resolution or BEV resolution, we obtain expert-apprentice pairs. * represents this version of BEVFormer-Base apply ResNet101 as backbone rather than ResNet101-DCN, which is different from the original version [19]. † indicates using nuScene pretrained weight with FCOS3D [30]. Unless otherwise specified, the backbone is pretrained on ImageNet only.

chitecture and high input resolution. By compressing the network depth or input resolution, we obtain more compact networks. However, the lightweight networks inevitably suffer from accuracy attenuation. Therefore, we employ the heavy state-of-the-art model as the expert, adopt the lightweight counterpart as the apprentice, and perform knowledge transfer between them. We freeze the expert network, and leverage the intermediate features of the expert network as auxiliary supervision for the apprentice network. The setting of adopted expert-apprentice pairs is depicted in Tab. 1. Their GLOPS and FPS are presented at Tab. 2.

As depicted in Fig. 3, a distillation framework between the expert network and the apprentice network is constructed. For the BEV perception model, the distillation is performed on both the perspective view (PV) and BEV features. We set up an extra distillation head and employ a set of aligned BEV queries Q for distillation. The structure of the distillation head is similar to the head of Deformable DETR [39]. Based on the deformable attention mechanism, N queries sample key features from BEV features and generate 3D bounding boxes. The generated 3D bounding boxes are projected into PV and BEV. The projection areas are used for masked generative distillation, and the sampled BEV features in the deformable attention process are adopted for refined distillation. In order to urge the projection areas locate at the foreground area and the sampled BEV features come from the object region, the distillation head is also optimized with auxiliary detection losses.

The distillation head is built upon the expert BEV features. The distillation queries go through a self-attention layer, and then sample expert BEV feature F^E to generate y^E through a deformable attention (DEFORMATTN) layer and a feed-forward network ($\mathcal{FFN}$). The latter process is formulated as:

$$y^E = \mathcal{FFN}(\text{DEFORMATTN}(Q, F^E)), \tag{1}$$

where $y^E = \{[c_1^E, b_1^E], [c_2^E, b_2^E], \dots, [c_N^E, b_N^E]\}$, and c_i^E, b_i^E represent classification score and regression bounding box corresponding to i-th distillation query. Based on bipartite matching approach in [2], we search for the optimal permutations $\hat{\sigma}$ and optimize the distillation head by $\mathcal{L}_{dh}$ as follows:

$$\begin{cases} \hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_i^N \alpha \mathcal{L}_{cls}(c^E_{\sigma(i)}, c_i^{gt}) + \beta \mathcal{L}_{bbox}(b^E_{\sigma(i)}, b_i^{gt}), \\ \mathcal{L}_{dh} = \sum_i^N \alpha \mathcal{L}_{cls}(c^E_{\hat{\sigma}(i)}, c_i^{gt}) + \beta \mathcal{L}_{bbox}(b^E_{\hat{\sigma}(i)}, b_i^{gt}), \end{cases} \tag{2}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{bbox}$ take the same form as FocalLoss [20] and L1 loss respectively, b^{gt} and c^{gt} represent ground truth boxes coordinates and classes, and α as well as β balance the loss weight.

During the optimization process of the distillation head, the queries with a confidence higher than ρ_c and GIOU greater than ρ_b are filtered out as focal queries, termed as Q^f. The number of Q^f is denoted as N_f. Finally, the generated 3D boxes and sampled BEV features corresponding to the focal queries Q^f are selected to perform masked generative distillation and refined distillation.

3.2. Selective Masked Generative Distillation

The generated 3D boxes by focal queries Q_f are projected to PV and BEV. The 3D boxes are projected to PV according to the camera extrinsics and intrinsics, and simply flattened to BEV from the top-down view. The masking process described below is the same for both PV and BEV, so we describe the process without distinction between PV and BEV.

An instance mask I_i^{fg} is generated with the projection area corresponding to one focal query Q_i^f. The dimension of the instance mask is identical to that of the apprentice feature. If a pixel is inside the bounding box b_i^E, the value is set to be 0, otherwise its value is set to be 1. The formulation of I_i^{fg} is described as follows:

$$I^{fg}_{i,uv} = \begin{cases} 0, & \text{if } (u,v) \text{ in } b_i^E \\ 1, & \text{Otherwise} \end{cases}. \tag{3}$$

All instance masks are merged together to form the final mask I^{fg}. The process of mask stacking is formulated as follows:

$$I^{fg} = I_0^{fg} \wedge I_1^{fg} \wedge \cdots \wedge I_{N-1}^{fg}. \tag{4}$$

The masked region indicates where the objects reside in, and the masked generation is performed within the masked region.

The selective mask indicates that the pixels of apprentice feature within I^{fg} are masked by ratio r. A random mask I^r with the same size of I^{fg} is generated:

$$I^r_{uv} = \begin{cases} 0, & \text{if } R_{u,v} < r \\ 1, & \text{Otherwise} \end{cases}, \tag{5}$$

where $R_{u,v}$ is a uniformly random number between 0 and 1. The process of the selective mask is expressed as an OR operation between I^{fg} and I^r:

$$I^m = I^{fg} \vee I^r. \tag{6}$$

Then the apprentice feature F^A is masked with I^m. The masked apprentice feature is forced to recover with a generator $\mathcal{G}$:

$$\hat{F}^A = \mathcal{G}(F^A \cdot I^m). \tag{7}$$

The generator $\mathcal{G}$ here includes two convolutional layers, the first one of which switches to a deconvolutional layer if the dimensions of the apprentice feature and expert feature are not identical. After generating $\hat{F}^A$, the channel-wise distribution [25] is calculated as follows:

$$\mathcal{L}_{SMGD} = \frac{\tau^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{H \times W} \varphi(F^E_{c,i}) \log \frac{\varphi(F^E_{c,i})}{\varphi(\hat{F}^A_{c,i})}, \tag{8}$$

$$\varphi(F_{c,i}) = \frac{exp(\frac{F_{c,i}}{\tau})}{\sum_{i=1}^{H \times W} exp(\frac{F_{c,i}}{\tau})}, \tag{9}$$

where τ is the temperature factor of the SoftMax layer, C means the feature channels, and H, W represent feature height and width respectively.

3.3. Query-based Focal Distillation

Focal distillation denotes the refined distillation focusing on the local feature pattern of focal positions. We aim to leverage a set of queries to automatically search out the representative positions of objects, and perform distillation on the identical positions between expert features and apprentice features. We resort to the deformable attention module to achieve the goal. The deformable attention module generates a set of offsets around a reference point. Based on the offsets to the reference point, the features on the crutial sampling points are sampled for object detection. We exploit this mechanism to conduct a refined distillation on such representative features as illustrated in Fig. 4.

A set of focal queries Q^f are filtered out as described in Sec. 3.1. For each focal query $q \in Q^f$, the reference point p_q is generated by the query feature z_q with a linear projection $\mathcal{MLP}$:

$$p_q = \mathcal{MLP}(z_q). \tag{10}$$

The reference point p_q represents the projection center of the estimated bounding box on the sampling feature plane.

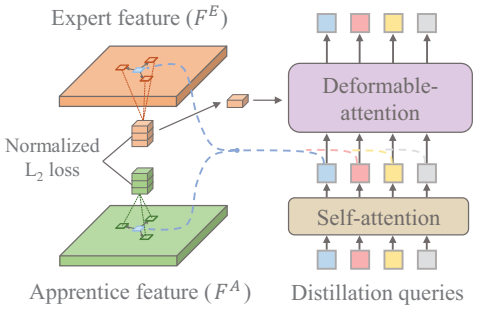

Figure 4. Process of query-based offset generation, feature sampling and query parameter update. Leveraging the deformable attention mechanism, the local expert feature and apprentice feature are sampled at the same position, based on the offset generated by the query. In addition, the query parameter is updated with the collected expert feature.

For each focal query q, M attention heads are attached on z_q. Each attention head generates K sampling offsets by another linear projection $\mathcal{MLP}'$. The offset is computed by the following:

$$\Delta p_{mqk} = \mathcal{MLP}'(z_q, m, k), \tag{11}$$

where m indexes the attention head, and k represents the sampling point.

The positions searched by the generated offsets are viewed as key representative positions. The query-based focal distillation is conducted on such key representative positions. Let F^E_{mqk}, F^A_{mqk} represent the sampled features from the identical key positions for expert and apprentice,respectively. They are calculated as follows:

$$F^E_{mqk} = F^E(p_q + \Delta p_{mqk}), \tag{12}$$

$$F^A_{mqk} = F^A(p_q + \Delta p_{mqk}). \tag{13}$$

The sampled expert feature and apprentice feature are normalized to $\widetilde{F}^T_{mqk}$ and $\widetilde{F}^S_{mqk}$ according to their maximum and minimum. Finally, the L2 loss is computed between the normalized key sampled features:

$$\mathcal{L}_{QFD} = \frac{1}{N_f} \sum_{q \in Q_f} \sum_{m=1}^{M} \sum_{k=1}^{K} L_2(\widetilde{F}^A_{mqk}, \widetilde{F}^E_{mqk}). \tag{14}$$

Through the loss for the fine-grained local region, the refined distillation on these representative positions is conducted. Query-based focal distillation aims to optimize the local patterns on the instances' representative positions.

3.4. Overall Distillation Loss

Besides being supervised by ground truth, the intermediate feature of the apprentice detector is also supervised by

the masked generative distillation loss $\mathcal{L}_{SMGD}$ and query-based focal distillation loss $\mathcal{L}_{QFD}$. The total knowledge distillation loss is defined as follows:

$$\mathcal{L}_{KD} = \lambda \mathcal{L}_{SMGD} + \eta \mathcal{L}_{QFD}. \tag{15}$$

$\mathcal{L}_{SMGD}$ is used to optimize the global feature distribution, indicating where should be with high activation. $\mathcal{L}_{QFD}$ is leveraged for optimizing the local feature pattern on searched focal positions. The hyper-parameters λ and η control the loss weight.

Note that only the gradients w.r.t. the vanilla detection loss and the distillation loss $\mathcal{L}_{KD}$ back-propagate to the apprentice network. $\mathcal{L}_{dh}$ only optimizes the parameters of the distillation head.

4. Experiments

4.1. Dataset and metrics

We conduct experiments on nuScenes dataset [1], a large-scale autonomous driving dataset. This dataset contains 700, 150, 150 scenes for training, validation and testing, respectively. Each scene has roughly 20 seconds of duration. The key frame is annotated at 2 Hz.

The two dominant metrics of the nuScenes detection task are nuScenes Detection Score (NDS) and mean Average Precision (mAP). The mAP of nuScenes is computed by the center distance between predictions and annotations on the ground plane. In addition, the nuScenes dataset defines five true positive metrics mATE, mASE, mAOE, mAVE, mAAE for measuring translation, scale, orientation, velocity and attribute, respectively. NDS of nuScenes is a weighted sum of mAP and the five true positive metrics, defined as $NDS = \frac{1}{10}[5mAP + \sum_{mTP}(1 - min(1, mTP))]$.

4.2. Implementation Details

We conduct experiments on BEVFormer [19] and DETR3D [32]. Settings of expert-apprentice pairs are illustrated in Tab. 1. We obtain three groups of expert-apprentice combinations by changing the backbone, input image resolution and BEV grip resolution. ResNet [8] with or without deformable convolution [38] are used for all backbones. When applied to BEV perception model BEVFormer, the selective masked generation is performed on both the PV feature and BEV feature, and query-based focal distillation is performed on the BEV feature, while both selective masked generative distillation and query-based focal distillation are performed on PV feature when it comes to the BEV-free perception model DETR3D. We also conduct the experiment on BEVDepth in supplementary materials.

The codebase is developed upon MMDetection3D [5]. All models are trained on 8 NVIDIA A100 GPUs. For BEVFormer and DETR3D, the models are trained for 24 epochs with an initial learning rate of 2e-4 and per-GPU batch size of 1. No data augmentation is introduced when training BEVFormer and DETR3D. Feature pyramid position shift [24] is adopted for PV feature alignment between low-resolution apprentice and high-resolution expert. The hyperparameters r, τ, α, β, λ, η are set to be 0.5, 2.0, 2.0, 0.25, 0.25 and 0.001, respectively.

4.3. Comparison to State-of-the-arts

As depicted in Tab. 2, the proposed distillation method brings significant improvement on BEVFormer-Tiny, BEVFormer-Small and DETR3D by 4.07, 2.47 and 3.17 NDS, respectively. The gain comes mainly from the improved mAP and velocity estimation accuracy. The devised method brings 4.13 points to BEVFormer-Tiny and 3.08 points to BEVFormer-Small in terms of mAP, which demonstrates the enhancement of localization and classification. The proposed approach also achieves improvements of 4.97 points to BEVFormer-Tiny and 8.21 points to DETR3D-R50 in terms of mAVE. In regard to the classic distillation methods on the 2D field, regardless of BEVFormer-Tiny or DETR3D-R50, FitNet obtains marginal improvement, yet CWD achieves larger gains. MGD with the global random mask cannot outperform CWD, indicating that random masked generative distillation in 3D detection scenes is not as powerful as that in 2D scenes. The proposed FD3D outperforms all classic distillation methods on the 2D field by an obvious margin. The comparison reveals that the devised fine-grained distillation on focal regions truly improves the distillation quality. In the supplementary materials, we conduct further experiments to validate the efficacy of FD3D on another BEV perception model BEVDepth. It can be observed that FD3D leads to improved performance on BEVDepth. These findings demonstrate the generalizability of the proposed distillation method across various 3D object detectors. Moreover, we also present the gains achieved by FD3D on the nuScenes *test* set, which are consistent with the effects observed on the nuScenes *val* set.

4.4. Ablative Study

Contribution of each module. For a better understanding of the contribution of each proposed module to the distillation performance, we test each component independently on group BEVFormer Base-Tiny and report the corresponding accuracy in Tab. 3. The channel-wise distribution (CWD) is adopted as a baseline distillation method and obtains 41.80 NDS. When compared to this baseline, the selective masked generation can achieve an additional 1.07 NDS improvement, revealing that such technique can enhance feature learning on focal regions. Query-based focal distillation yields an extra 0.95 NDS gain over the baseline, thereby highlighting the significance of the refined distillation on the fine-grained position. With the two proposed

Method	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	GFLOPs↓	FPS↑
BEVFormer-Base* (E)	47.37	36.44	73.51	28.14	44.98	42.38	19.50	1845.36	2.0
BEVFormer-Tiny (A)	39.02	26.87	83.43	29.48	59.73	50.03	21.49	381.95	7.3
+ FitNet [26]	40.57	28.26	81.52	28.96	52.96	51.32	20.84		
+ CWD [25]	41.80	29.79	80.81	28.55	**50.63**	49.87	21.21		
+ MGD [36]	41.33	29.08	80.19	28.83	52.83	50.00	**20.19**		
+ FD3D (Ours)	**43.09 (↑4.07)**	**31.00 (↑4.13)**	**79.31**	**28.00**	50.77	**45.06**	20.97		
BEVFormer-Base (E)	51.74	41.64	67.26	27.34	37.04	39.41	19.74	1323.41	1.8
BEVFormer-Small (A)	46.26	34.56	74.27	28.00	44.53	43.98	**19.41**	416.46	5.9
+ FD3D (Ours)	**48.73 (↑2.47)**	**37.64 (↑3.08)**	**71.94**	**27.50**	**40.89**	**39.41**	21.16		
DETR3D-R101 (E)	42.5	34.6	77.3	26.8	38.3	84.2	21.6	1016.83	2.5
DETR3D-R50 (A)	35.78	28.85	85.39	27.90	54.09	95.15	23.95	876.94	4.0
+ FitNet [26]	36.19	29.76	83.90	28.00	56.18	94.53	24.33		
+ CWD [25]	37.22	30.13	82.74	**27.65**	49.40	94.65	23.98		
+ MGD [36]	37.03	29.89	83.35	27.79	53.28	92.74	21.98		
+ FD3D (Ours)	**38.95 (↑3.17)**	**31.33 (↑2.48)**	**82.32**	**27.65**	**48.40**	**86.94**	**21.87**		

Table 2. 3D object detection improvement on nuScenes *val* set. The results demonstrate that the proposed distillation method benefits various 3D object detectors and surpasses the classic 2D distillation methods. We also show their efficiency metrics with GFLOPs and FPS. The FPS is measured on RTX 2080TI.

	Component			NDS	mAP
	CWD	SMGD	QFD		
BEVFormer Base-Tiny				39.02	26.87
	✓			41.80	29.79
		✓		42.87	30.59
			✓	42.75	30.47
		✓	✓	**43.09**	**31.00**

Table 3. Contribution by each component. SMGD and QFD are the proposed two novel modules. SMGD represents selective masked generative distillation, and QFD indicates query-based focal distillation. Channel-wise distribution (CWD) is applied as the baseline distillation method for comparison.

modules applied in conjunction, there is an observed increase in performance of 1.29 NDS, providing evidence of their efficacy and compatibility.

Mask region selection. As shown in Tab. 4a, variation in the impact of masked generative distillation across different regions is substantial. Global random mask leads to worsened results. Mask on foreground regions defined by the ground truth yields significant gains. Furthermore, applying masks to the predicted focal regions, as defined by focal queries, results in further improvement. This is probably attributed to the selection of focal queries, which effectively filter out regions with high confidence while excluding challenging-to-discriminate regions.

Distillation region. The terms "global distribution" and "local pattern" refer to two distillation mode focusing on the global overall activation and the local feature representation, respectively. And their performances are evaluated in Tab. 4b. The results demonstrate that distilling with global distribution alone already yields improvement, and when focusing on the local pattern of focal regions, a slightly higher gain is observed. The best performance is achieved through the combined optimization of them.

Distillation stage. The view transform module takes PV features as input and generates BEV features. PV features and BEV features are from different stages of the network. We explore the difference when global distillation (Fig. 1 (a)) is performed at different stages. The channel-wise distribution (CWD) is adopted in this ablation study. As shown in Tab. 4c, separate distillation on PV features obtains an improvement while separate distillation on BEV features leads to a deterioration in performance. This is due to the BEV feature containing a lot of shadow-like artifacts due to the unclear occlusion. This result demonstrates the necessity of selective distillation under BEV feature.

Feature alignment approach. When the BEV resolution differs between expert and apprentice, the BEV feature alignment is needed. There are two choices for alignment: upsampling the apprentice feature through a deconvolutional layer or directly downsampling the expert feature. We compare the two alignment approaches in Tab. 4d. Downsampling expert feature provokes inferior results than upsampling apprentice feature. This is potentially caused by the fatal information loss of the expert feature.

4.5. Visualization

The feature representation and prediction of the apprentice network BEVFormer-tiny with or without our proposed method are presented in Fig. 5. For PV, the feature activations from the first FPN layer in front left view, front right view and back left are displayed. The activation of many background regions is high without distillation, making the foreground and background indistinguishable. With

Group	Distillation Region	NDS	mAP
BEVFormer Base-Tiny	directly imitation	41.80	29.79
	global random mask	41.14	28.49
	mask on GT	42.41	29.85
	selective focal mask	**42.87**	**30.59**
DETR3D R101-R50	directly imitation	37.22	**30.13**
	global random mask	37.19	30.03
	mask on GT	37.58	30.11
	selective focal mask	**37.91**	29.87

(a) **Mask region selection**. Global random masked generation exacerbates the performance, while the selective masked generation on focal regions brings a remarkable improvement.

Group	Distillation Region	NDS	mAP
BEVFormer Base-Tiny	global distribution	41.80	29.79
	local pattern	41.91	29.89
	global distribution + local pattern	**42.68**	**30.30**
DETR3D R101-R50	global distribution	37.43	30.41
	local pattern	38.00	30.21
	global distribution + local pattern	**38.72**	**30.57**

(b) **Distillation region**. Global distribution means the channel-wise distribution of feature activation, and local pattern represents the feature values on the fine-grained focal regions. Distillation on both global distribution and local patterns achieves the best performance.

Group	Distillation Stage		NDS	mAP
	PV	BEV		
BEVFormer Base-Small			46.26	34.56
	✓		47.06	35.52
		✓	45.24	33.56
	✓	✓	**47.60**	**35.97**

(c) **Distillation stage**. Effect of global channel-wise distribution distillation (Fig. 1 (a)) on PV or BEV. The result indicates that combined distillation on PV and BEV achieves the highest improvement. Note that separate global distillation on BEV feature leads to a deterioration in performance.

Group	Feature Alignment	NDS	mAP
BEVFormer Base-Small	downsample	47.47	36.66
	deconvolution	**48.73**	**37.50**

(d) **Feature alignment.** Downsampling expert feature gets significantly worse results than upsampling apprentice feature through a deconvolutional layer.

Table 4. **FD3D ablations** on nuScenes *val* set. We show nuScenes detection score (NDS) and mean average precision (mAP) methics (%). Default settings are marked in gray.

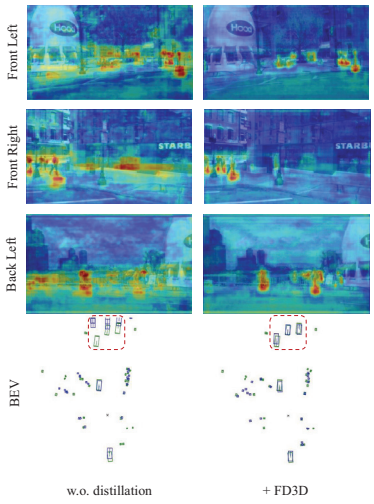

Figure 5. Visualization of feature activation and the predicted results before and after knowledge distillation. In comparison with the feature activation on PV, the foreground and background become clearly distinguishable with distillation. The predictions are illustrated on BEV, where green and blue boxes denote the ground truth and predictions respectively. With the devised method, the prediction accuracy of detection boxes is significantly improved.

the proposed approach, they become clearly distinguishable. The PV feature visualization confirms that the distillation makes the backbone of the apprentice network well-optimized. High-quality 2D image feature extraction is the basis for BEV feature production and subsequent detection. In addition, we give a comparison of predicted bounding boxes in BEV. Through the comparison of the area within the red dashed box, our approach significantly reduces the translation error. Though the accuracy of the bounding boxes is obviously improved, the false positive issues are still serious after distillation due to the inherent detects in the 3D object detection model.

5. Conclusion and Future Work

In this work, we apply knowledge distillation to camera-only 3D object detection, and we reveal that the challenge lies in how to distill focal knowledge when confronted with an imperfect expert. We devise FD3D, a flexible query-based approach that automatically searches the representative regions and highlights these regions for distillation. It serves as a plug-and-play module and successfully applies to various 3D object detectors.

Limitation and future work. The challenge of distilling knowledge from imperfect experts exists in not only the 3D object detection domain, but also in other fields. The proposed approach is needed to be extended to address more general scenarios.

Acknowledgements. This work is partially supported by NSFC (62206172, 62222607), and the Shanghai Committee of Science and Technology (21DZ1100100).

References

[1] 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. 1, 6

[2] 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. 3, 4

[3] 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

[4] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv preprint arXiv:2201.10830*, 2022. 1, 3

[5] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 6

[6] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021. 2

[7] 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

[8] 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. 6

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

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

[11] 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. 3

[12] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398*, 2022. 3

[13] 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

[14] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1

[15] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Enze Xie, Zhiqi Li, Hanming Deng, Hao Tian, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *arXiv preprint arXiv:2209.05324*, 2022. 1

[16] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017. 2

[17] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *arXiv preprint arXiv:2206.00630*, 2022. 1, 3

[18] 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. 3

[19] 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, 3, 4, 6

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[21] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022. 1

[22] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021. 3

[23] 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. 3

[24] Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Jiaya Jia. Multi-scale aligned distillation for low-resolution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14443–14453, 2021. 6

[25] 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. 5, 7

[26] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. *Advances in neural information processing systems*, 28, 2015. 2, 7

[27] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou,

Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1

[28] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 3

[29] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 2

[30] 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. 3, 4

[31] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 3

[32] 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. 3, 6

[33] Yue Wang and Justin M Solomon. Object dgcnn: 3d object detection using dynamic graphs. *Advances in Neural Information Processing Systems*, 34:20745–20758, 2021. 3

[34] Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3d object detection with knowledge distillation. *arXiv preprint arXiv:2205.15156*, 2022. 3

[35] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 2

[36] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022. 2, 7

[37] 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. 1

[38] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 6

[39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4