ON THE IMPORTANCE OF PRETRAINED KNOWLEDGE DISTILLATION FOR 3D OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-camera 3D object detection for autonomous driving is quite challenging and has drawn great attention from both academia and industry. The core issue of the vision-only methods is that it is difficult to mine accurate geometry-aware features from images. To improve the performance of vision-only approaches, one promising ingredient in the recipe lies in how to use visual features to simulate the geometry information of LiDAR, since point cloud data inherently carries 3D spatial information. In this paper, we resort to knowledge distillation to leverage useful representations from the LiADR-based expert to enhance feature learning in the camera-based pipeline. It is observed that the joint optimization of expert-apprentice distillation as well as the target task might be difficult to learn in the conventional distillation paradigm. Inspired by the great blossom and impressive results of foundation models in general vision, we propose a pretrained distillation paradigm, termed as **PreDistill**, to decouple the training procedure into two stages. The apprentice network first emphasizes the knowledge transfer from the expert; then it performs finetuning on the downstream target task. Such a strategy would facilitate the optimal representation learning with targeted goals and ease the joint feature learning as resided in conventional single-stage counterpart. PreDistill serves as a convenient plug-and-play that is flexible to extend to multiple state-of-the-art detectors. Without bells and whistles, building on top of the most recent approaches, *e.g.*, BEVFusion-C, BEVFormer, and BEVDepth, we could guarantee a unanimous gain of 7.6%, 1.0%, and 0.6% in terms of NDS metric on nuScenes benchmark. Code and model checkpoints would be available.

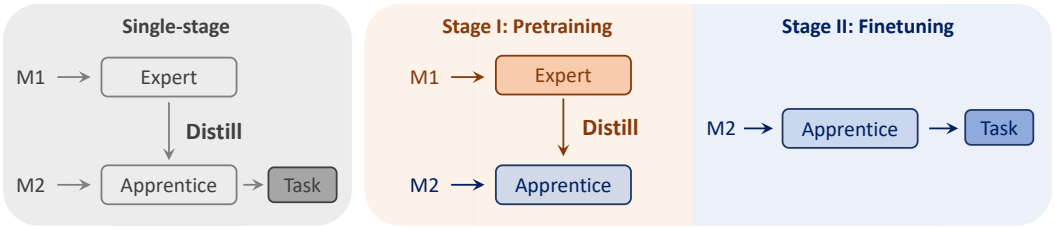
1 INTRODUCTION

Recognizing objects in 3D space is a fundamental and challenging task in autonomous driving. Camera-based approaches (Philion & Fidler, 2020; Li et al., 2022b;c; Liu et al., 2022) obsess the advantage of semantic and visual information of objects, while LiDAR-based methods (Yan et al., 2018; Lang et al., 2019; Shi et al., 2021; Yin et al., 2021) present better geometry-aware representation from point cloud data. As the sensor configurations get more diverse, representing features in bird's-eye-view (BEV) to indulge the goodness of both modalities in 2D and 3D space is trending and has drawn massive attention from both academia and industry.

Bear in mind that there is a distinct performance gap between camera-only and LiDAR-based approaches based on pubic 3D detection benchmarks (Sun et al., 2020; Caesar et al., 2020). The gap is due to the fact that the LiDAR modality has explicit geometry information but not the case in camera modality. The core and most challenging part of a superior vision-only 3D detector are to equip features with well-learned geometry representations in 3D space. However, relying on visual information only to reconstruct 3D scenes will inevitably cause ambiguity and inaccuracy.

In this paper, we cast the 3D object detection problem using camera input only, and try to mitigate the performance gap between camera and LiDAR. That is, with the aid of LiDAR information available during training, how to devise a general pipeline to extract geometry information and incorporate it into the feature learning of the camera pipeline is a critical problem. As such, among the many solutions, knowledge distillation could be an option.

Knowledge distillation (Hinton et al., 2015) is proposed to transfer knowledge from the expert model to guide the apprentice (student) network. The core idea is to learn or mimic the features, or behav-



(a) Single-stage Distillation (b) Pretrained (Two-stage) Distillation

Figure 1: Comparison of two different paradigms for knowledge distillation, where $M1$ and $M2$ represent different sources of input (modality) for the expert and apprentice respectively. For **(a) single-stage distillation**, the target task as well as the learning of apprentice network is optimized simultaneously within the distillation framework; for **(b) pretrained distillation** investigated in this paper, the optimization of apprentice network and downstream task is decoupled into two stages. This is inspired by the impressive practice from visual pretrained models, benefiting from learning more representative features with targeted goals.

iors in Robotics terminology, of a target network to a downstream, lightweight apprentice model. As depicted in Fig. 1, there are two typical distillation paradigms, based on whether the target task is optimized simultaneously within the distillation framework.

There are some works to distillate the expert to the apprentice for autonomous driving tasks using single-stage distillation. Some resort to a stereo input setting (Guo et al., 2021) to generate depth estimation via disparity maps in the camera branch, and the LiDAR information is further incorporated via BEV feature distillation. Others attempt multi-level (feature and response) distillation in a monocular setting (Chong et al., 2022) or unify LiDAR and camera voxelized features for fusion task (Li et al., 2022a). Despite various settings and applications, these literature enjoy impressive performance gain under the single-stage distillation spirit. However, for cross-modal distillation, the network might be confused about which part to learn and optimize because of the huge difference in between, resulting in a marginal or even inferior improvement for challenging scenarios.

Inspired by the two-phase training procedure and great success in visual pretrained models (Bai et al., 2022b; Kuncoro et al., 2020; Wu et al., 2022), we hypothesize that decoupling the expert-apprentice distillation from the target task into two stages, might ease the difficulty of joint feature optimization. That is, instead of performing the task at hand within distillation, we would first perform pretraining to impose the apprentice network on feature distillation from the expert alone; the adaptation and finetuning of the target task would then be optimized based on the well-pretrained knowledge from the preceding stage. To the best of our knowledge, there are few attempts following this two-stage distillation spirit. Sautier et al. (2022) proposed a similar pipeline to leverage the benefits of pretraining in a contrastive self-supervised fashion. It utilizes the semantic features from camera as an expert to guide the representation learning in the LiDAR branch for segmentation task. Note that we do *not* intentionally advocate pretrained distillation over the single-stage option; rather in this paper, we try to investigate another perspective of distillation via a disentangled spirit.

To this end, we present a general pretrained distillation pipeline for knowledge transfer from LiDAR to a camera-only detection network, namely **PreDistill**. The overall pipeline is described in Fig. 2. It consists of three phases, where two novel refinements are proposed to facilitate feature representation learning. A selective focus module is devised to emphasize foreground instances and background information in BEV space; a duplication operation is performed during the finetuning period, allowing for faster convergence and better feature alignment between two modalities.

The contributions are summarized as follows:

- We propose PreDistill, a pretrained distillation paradigm for knowledge transfer. It decouples the expert-apprentice distillation process from the downstream finetuning task. Such a scheme would facilitate better representative learning for the apprentice model.
- We provide another perspective to utilize geometry-aware LiDAR information to enhance the performance of vision-centric object detection. This is achieved by two proposed strategies in the pipeline, namely selective focus and duplication.

- We demonstrate that PreDistill serves as a plug-and-play module extensible to various state-of-the-art detectors. Without bells and whistles, building on top of BEVFusion, BEVFormer, and BEVDepth approaches, we could guarantee an unanimous gain of 7.6%, 1.0%, and 0.6% in terms of NDS metric on nuScenes benchmark.

2 Related Work

2.1 Knowledge Distillation

Knowledge distillation (Hinton et al., 2015) is initially proposed to transfer the learned knowledge from a large network to a small one for model compression. This strategy has been proved effective in various computer vision tasks, such as 2D object detection (Chen et al., 2017; Wang et al., 2019; Dai et al., 2021; Yang et al., 2022b) as well as 3D domains (Cho et al., 2022; Yang et al., 2022a; Zhang et al., 2022). As stated in Sec. 1, there are two types of distillation based on whether the task at hand is optimized alongside the optimization of the apprentice network.

Single-stage Distillation. Most works fall into this category in various forms, and yet in pursuit of the same purpose to optimize the student network. Gupta et al. (2016) transferred knowledge between different modalities to facilitate the feature learning in the few labelled or even unlabelled student task. MonoDistill (Chong et al., 2022) designed a multi-level strategy to provide guidance from the LiDAR teacher to the camera student, considering from both feature and result space. LIGA-Stereo (Guo et al., 2021) learned a stereo-based model to imitate the high-level geometry-aware representation from a LiDAR-based model in 3D and BEV space. The view transformation process can benefit from the relatively accurate depth information. UVTR (Li et al., 2022a) transferred knowledge from geometry-rich teacher (LiDAR) to geometry-inferior student (camera) by minimizing distance without excluding background features. The idea of modality switch and unified voxelization allows flexible extension for sensor fusion task from a distillation perspective. These work train the knowledge distillation framework end-to-end alongside the target task in the student model, which may suffer from the difficulty of feature joint optimization.

Pretrained Distillation. As the success of visual pretrained models continues, some attempt to apply knowledge distillation in the pretraining before finetuning the downstream task. Kuncoro et al. (2020) distilled the approximate marginal distribution over words in context to inject syntactic biases in the language model BERT (Devlin et al., 2018). TinyViT (Wu et al., 2022) extracted knowledge from a large model to a small one in the large-scale data pretraining setting where the small student model can benefit superior feature learning from massive pretraining data. The tasks are performed in the general vision domains. Liu et al. (2021) introduced a learned 2D model to pretrain a 3D counterpart by contrastive learning between two modalities. The proposed method is verified in the indoor point cloud setting for detection and segmentation tasks. SLidR (Sautier et al., 2022) in contrast adopted a camera model as teacher to facilitate the feature learning of a LiDAR network in a contrastive self-supervised manner. Our work is inspired by the aforementioned pretrained distillation approaches from various applications, in the sense of decoupling the knowledge distillation and the downstream task finetuning to better optimize feature learning in the student model.

2.2 3D Object Detection

Camera-based 3D Object Detection. Recent years have witnessed a great blossom of vision-centric approaches to achieve impressive results on many benchmarks. Due to several benefits to perform perception task in BEV space (Li et al., 2022c), recent methods tend to conduct view transformation first to obtain BEV features, and then predict detection results based on the BEV features.

Lift-Splat-Shoot (LSS) (Philion & Fidler, 2020) proposed leveraging depth distribution to model uncertainty in depth estimation, and plenty of works follow this paradigm for perspective transformation (Huang et al., 2021; Li et al., 2022b). To further solve the issue of inaccurate depth estimation, BEVDepth (Li et al., 2022b) utilized explicit depth supervision. BEVFormer (Li et al., 2022c) performed view transformation by exploiting deformable attention and devised grid-shaped queries in the BEV space. In this work, we also aim for the camera-based detection task, and propose a novel distillation paradigm to improve the performance of existent state-of-the-art detectors.

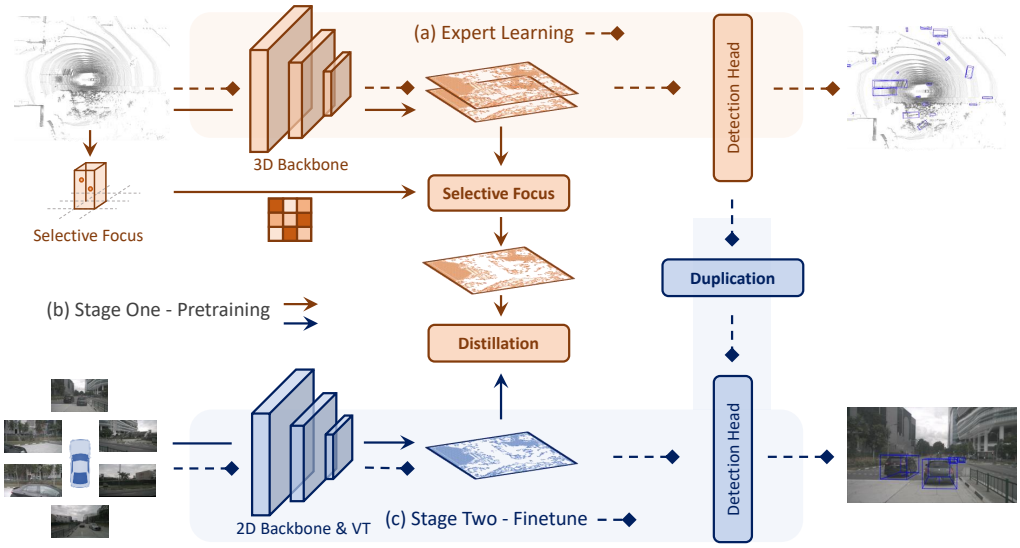


Figure 2: The proposed **PreDistill pipeline** is formulated in a pretrained distillation spirit and consists of three phases. View transformation is abbreviated as VT. In (a) expert learning, the LiDAR network is trained to obtain optimized features in the backbone and detection head respectively. The core part of PreDistill lies in the (b) pretraining phase, as indicated by solid arrows. The goal is to learn representative features for apprentice network from its LiDAR counterpart via knowledge distillation. A selective focus module is introduced to emphasize foreground features of the expert based on sparsity density in the point clouds. In the last phase (c), we perform the target task with pretrained apprentice network, and finetune the model with initial parameters in the head duplicated from those in the expert.

LiDAR- and Fusion-based 3D Object Detection. LiDAR-based approaches (Yan et al., 2018; Zhou & Tuzel, 2018; Yin et al., 2021; Lang et al., 2019) utilize accurate spatial information from point cloud data and thus yield superior performance compared to camera solutions. To further boost performance, fusion-based approaches take LiDAR and camera data as input, leveraging both the geometry and semantic advantages in two modalities. TransFusion (Bai et al., 2022a) adopted the bounding box prediction as a proposal to query the image feature, then fused the visual information to LiDAR features. BEVFusion (Liang et al., 2022) proposed a simple yet effective pipeline to combine BEV representations from different sensors and achieved impressive results.

3 Method

Fig. 2 illustrates the overall pipeline of our method. We first elaborate on the network structures of the apprentice and expert in Sec. 3.1. Then the training and inference procedures of PreDistill are presented in Sec. 3.2. Two important refinements on the pipeline are also described in Sec 3.3.

3.1 Network Architecture

Apprentice Model. State-of-the-art multi-camera methods usually transform image features into BEV space, and perform detection based on BEV representation. View transformation is the main component to construct BEV representations, and it can be categorized either as 2D-3D process or 3D-2D. Recent camera-based approaches (Li et al., 2022c; Huang & Huang, 2022) introduce the temporal design to further leverage motion and geometry information to boost model performance.

To prove the generalization ability of our proposed pipeline, we consider multiple models under different settings as our baselines. BEVFusion-C (Liang et al., 2022) serves as a baseline model of 2D-3D view transformation without temporal information. BEVDepth (Li et al., 2022b) task multiple frames as input and utilizes 2D-3D view transformation. With temporal information, BEVFormer (Li et al., 2022c) using deformable attention as a 3D-2D view transformation method is also

taken into account. As our method can be applied to camera models which have the explicit BEV representation, without loss of generality, we abstract the network modules before the BEV feature map as the backbone of the apprentice. The rest of the network is denoted as the detection head.

Expert Model. We adopt the LiDAR backbone from TransFusion-L (Bai et al., 2022a), which is a popular and efficient architectural design. Since we aim to align the BEV representation between LiDAR and camera data, the same detection head network was adopted in both expert and apprentice models, for better utilizing the consistency of BEV features. Note that the expert would first be trained to secure a decent performance on the 3D detection task, ensuring an distinct performance upper bound for the apprentice distillation pipeline.

3.2 PREDISTILL: TRAINING AND INFERENCE

Due to the lack of spatial and structural cues, 3D detection methods taking RGB images as input hardly provide satisfactory perception results. To bridge the performance gap between LiDAR- and camera-based methods, we propose a general training pipeline, which utilizes LiDAR information to guide the training process of camera models. As depicted in Fig. 2, the proposed pipeline can be divided into three steps: expert learning, pretraining, and finetuning.

Expert Learning. An expert network, which consists of an off-the-shelf LiDAR backbone b_L and a detection head h_L, is first trained on the 3D object detection task. Taking a LiDAR point cloud $P \in \mathbb{R}^{N \times K}$ as input, where N is the number of points and K is the number of dimensions of input data, the backbone produces a BEV representation $b_L(P) \in \mathbb{R}^{D \times X \times Y}$, where D is the number of channels, and the spatial shape is defined as X and Y. A common detection loss is utilized to supervise the expert model on the output $h_L(b_L(P))$. The learned backbone and detection head are denoted as $\tilde{b}_L$ and $\tilde{h}_L$ respectively.

Pretraining. We denote the input data of the apprentice as $I \in \mathbb{R}^{M \times 3 \times H \times W}$, which are multi-view RGB images in the resolution of H and W; they are supposed to cover the entire horizontal FOV (field of view) of the surrounding environment with M images. Taking images I as input, the backbone b_C of the apprentice outputs BEV feature $b_C(I) \in \mathbb{R}^{D \times X \times Y}$, owning the same shape as in the expert network. The selective focus module, which would be described in next section, provides a weighted mask $S \in \mathbb{R}^{X \times Y}$ to describe the density information of the point cloud. Given a learned LiDAR backbone, the weighted $\mathcal{L}_2$ pretraining loss is denoted as:

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{X \cdot Y} \sum_{i}^{X} \sum_{j}^{Y} S_{ij} \left|\left| \tilde{b}_L(P)_{ij} - b_C(I)_{ij} \right|\right|^2. \tag{1}$$

Finetuning. The apprentice network is trained for the 3D detection task, where the camera head h_C shares identical architecture with that of LiADR's h_L. Since the semantic and geometry information of BEV features between the expert and apprentice is already aligned, we can directly duplicate the detection head of expert model to the apprentice models towards a better startpoint for weight initialization. This allows for fast convergence to an optimal learning.

Without any extra cost, the inference process of the well trained apprentice network follows the same procedure as does in conventional camera-based approaches. Our proposed PreDistill framework is a general plug-and-play module that can be flexible to multiple detectors.

3.3 REFINEMENTS

Selective Focus in Pretraining. Given the outputs after view transformation, the BEV features $b_C(I)$ contain some noise due to the inaccurate depth information. In contrast, the sparse point cloud data provides more attentive feature representations $b_L(P)$ in the BEV space. The gap between these two BEV representations makes it difficult to transfer knowledge directly. As regions with a small number of points in LiDAR data are less likely to provide useful features with high confidence, distilling knowledge in these regions may deviate the network from the right optimization objective. Thus we leverage the statistical hints from the density of point cloud data to restrict the distillation areas. Specifically, we define a weighted mask $S \in \mathbb{R}^{X \times Y}$ with the same spatial shape as the BEV feature. S_{ij} represents the weight counted according to the number of points within a

pillar at location (i, j). With the newly introduced selective focus module, our proposed PreDistill significantly improves the performance on top of several state-of-the-art detectors.

Duplication in Finetuning. The learned detection head $\tilde{h}_L$ of the expert, which is compatible with the LiDAR BEV feature maps, remains representative after being trained on the detection task. As the objective of the pretraining phase is to align the BEV representations from two modalities, the learned camera BEV representation $\tilde{b}_C(I)$ is supposed to follow the same distribution as in $\tilde{b}_L(P)$. Thus, we regard the learned detection head of the expert model as an ideal initial weight for the head of the apprentice. In practice, duplication not only makes the optimization easier but also brings a faster convergence speed.

4 EXPERIMENTS

In this section, we first introduce our experimental setups, including dataset, metrics, and implementation details. Then experiments on various multi-camera methods are conducted to validate the effectiveness of our proposed pipeline.

4.1 DATASET AND METRICS

The nuScenes dataset (Caesar et al., 2020) is a large-scale autonomous driving dataset, which contains 1000 driving scenes, in which 700, 150, and 150 scenes are for training, validation, and testing respectively. Each scene has a duration of about 20 seconds and is sampled at 2Hz for annotation. Each frame consists of RGB images with a resolution of 900×1600 from 6 cameras covering the entire horizontal FOV.

Two main metrics are provided for the 3D detection task, namely mean average precision (mAP) and nuScenes detection score (NDS). The mAP is computed using the center distance on the ground plane to match the predicted results and ground truths. The nuScenes dataset defines five types of true positive (TP) metrics, namely mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE), to measure translation, scale, orientation, velocity, and attribute errors respectively. The NDS is defined as a weighted sum of mAP and five TP metrics.

4.2 IMPLEMENTATION DETAILS

In the phase of expert learning, we follow the common usage of data augmentations, including random rotation, scaling, translating, and flipping. The input point clouds of expert models are filtered by the range of $[-51.2m, 51.2m]$, except that the range of BEVFusion-C follows the open-source implementation as $[-54m, 54m]$. Since the BEV features of different camera models are in different shapes, the voxel sizes of LiDAR models are changed accordingly. In the phase of pretrain, augmentation in both modalities is turned off to spatially align BEV representations. Apprentices are pretrained by 12 epochs by the AdamW optimizer. To make a fair comparison to baseline models, settings in the phase of finetuning like data augmentations and training epochs are kept the same for each camera network, despite smaller learning rates of $5e^{-5}$, $2e^{-4}$, and $5e^{-5}$ for BEVFusion-C, BEVDepth, and BEVFormer respectively. For all phases of the training pipeline, we only use data from the nuScenes dataset without introducing any external data.

BEVFusion-C. BEVFusion-C (Liang et al., 2022) uses Dual-Swin-Tiny (Liu et al., 2020) as the image backbone. The BEV feature map is in the shape of $512 \times 180 \times 180$, where 512 is the number of channels and 180 is the spatial shape along two axes. Input images have a resolution of 448×800.

BEVDepth. BEVDepth (Li et al., 2022b) adopts Res-50 (He et al., 2016) as its image backbone. It has a BEV representation in 256 channels and the shape of 128×128, and takes images with a resolution of 256×704 as input.

BEVFormer. BEVFormer (Li et al., 2022c) is trained with the same settings as the submitted version to the nuScenes leaderboard that VoVNet-99 (Lee et al., 2019) is used as the image backbone and the BEV feature is set to the size of 200×200 with 256 channels. The image backbone consumes RGB images in the shape of 900×1600 as input.

Table 1: 3D detection improvement on nuScenes *val* set. DST denotes the Dual-Swin-Tiny backbone. VoVNet-99 is pretrained on the depth estimation task with extra data (Park et al., 2021). The results indicate that our pipeline benefit various multi-camera methods.

Method	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFusion-C	DST	0.344	0.358	0.703	0.296	0.740	1.249	0.400
+ PreDistill	*gain*	+4.2%	+7.6%	+6.0%	+1.8%	+18.9%	+34.9%	+18.3%
BEVDepth	Res-50	0.361	0.484	0.650	0.276	0.497	0.340	0.199
+ PreDistill	*gain*	+0.6%	+0.6%	+2.4%	+0.1%	-1.3%	+0.2%	+1.1%
BEVFormer	VoVNet-99	0.435	0.534	0.667	0.276	0.339	0.360	0.195
+ PreDistill	*gain*	+1.1%	+1.0%	+6.3%	+1.3%	-5.7%	+1.4%	+1.2%

Table 2: Comparisons among different distillation approaches on nuScenes *val* set with BEVFusion-C. Following MonoDistill (Chong et al., 2022), Sce. denotes the scene-level distillation in feature space and Obj. denotes the object-level distillation in result space. The results show the merit to use pretrained distillation compared to single-stage distillation.

Distillation	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
-	0.344	0.358	0.703	0.296	0.740	1.249	0.400
Sce.	0.357	0.363	0.685	0.293	0.742	1.273	0.434
Sce. & Obj.	0.359	0.363	0.679	0.293	0.772	1.280	0.426
PreDistill	**0.386**	**0.434**	**0.643**	**0.278**	**0.551**	**0.900**	**0.217**

4.3 Building on Top of State-of-the-arts

As shown in Tab. 1, the proposed pipeline brings performance improvement to various state-of-the-art methods on the 3D detection task on nuScense *val* set, validating the generalization ability of our method. The most significant improvement is by the BEVFusion-C with a gain of 4.2% mAP and 7.6% NDS. Without the use of temporal information, BEVFusion-C has limited performance on various attributes, particularly on the mAVE. While the LiDAR network consumes 10 consecutive frames as input, the implicit temporal information credits to the model performance and thus serves as better guidance for the BEVFusion-C model. Note that only marginal improvement is observed in BEVDepth. This might result from the small image backbone of Res-50 and the smallest input resolution among all baseline models. The results also demonstrate the importance of a strong image backbone for feature extraction. By using our method, we improve the performance of BEVFormer by a clear margin of 1.1% mAP and 1.0% NDS, showing the effectiveness of our pipeline.

For all methods, mATEs improve significantly, which credits to the precise localization information of point cloud data. By contrast, mASEs are just slightly better than baselines. The reason is that camera models can already achieve competitive scale estimation performance to LiDAR methods. This can also apply to the improvement of mAVE and mAAE. However, the proposed pipeline exerts a negative effect on the orientation estimation. Since LiDAR point clouds are usually sparse for objects at a far distance, it leads to large orientation errors.

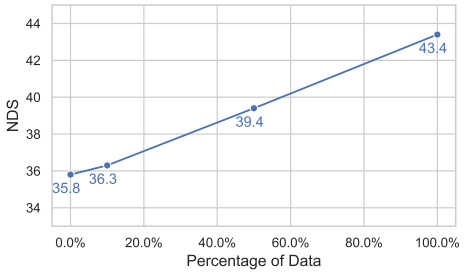
To further validate our results, based on the best-submitted setting of BEVFormer, we upload the results of the model trained by our pipeline. The reported scores on nuScenes *test* split are 49.1% mAP and 57.8% NDS, outperforming its original version by 1.0% mAP and 0.9% NDS.

4.4 Ablation Study

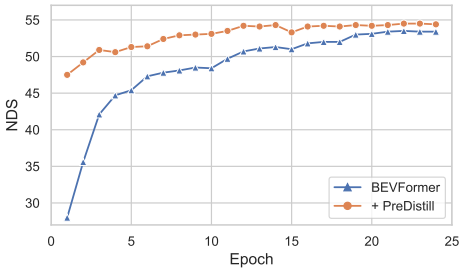
Distillation Paradigm. In Tab. 2, we follow the use of different distillation techniques at different stages of the network as in MonoDistill (Chong et al., 2022), using the typical single-stage distillation paradigm. MonoDistill projects point clouds into perspective space and utilizes 2D convolutional operations, shrinking the gap between two modalities. On the contrary, in our case, the backbone architectures of the expert and apprentice differ to a large extent. Therefore, directly utilizing the distillation modules from MonoDistill to our network in an end-to-end fashion (Row 2

Table 3: Comparisons among different pretraining strategies on nuScenes *val* set with BEVFormer, which uses VoVNet-99 as image backbone. Depth denotes that VoVNet-99 is pretrained on the depth estimation task with extra data (Park et al., 2021).

Depth	PreDistill	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
		0.324	0.433	0.784	0.287	0.524	0.482	0.212
	✓	0.403	0.504	0.651	0.275	0.490	0.369	0.191
✓		0.435	0.534	0.667	0.276	0.339	0.360	0.195
✓	✓	0.446	0.544	0.604	0.263	0.396	0.346	0.183



(a) Dependency of data scale in pretraining and model performance with BEVFusion-C.



(b) The number of epochs needed to converge during finetuning with BEVFormer.

Figure 3: During pretraining, our method requires a small amount of data to outperform the baseline, while during finetuning, it takes fewer iterations to converge.

& 3) brings a marginal improvement. By contrast, our method (Row 4) outperforms the baseline significantly. This confirms the superiority of our method compared to single-stage methods.

Backbone Pretraining. Park et al. (2021) demonstrates that depth pretraining using the large-scale DDAD15M dataset significantly improves the performance on 3D detection task. Experimentally in Tab. 3, though BEVFormer only pretrained by our method (Row 2) does not outperform the depth-pretrained version (Row 3), it is significantly better than the baseline (Row 1). Nevertheless, DDAD15M contains approximately 15M image frames, and only part of the data is open-sourced. Our method can be served as an alternative pretraining strategy for powerful image backbones.

Data Scale. We study the model performance using 10%, 50%, and 100% of training data in the pretraining phase. 0% denotes that the model is trained without our method. In Fig. 3a, with the growth of data used in the pretraining stage, the model performance improves roughly in a linear manner, indicating the increase of data might further impose a positive influence. It is also observed that increasing the number of pretraining epochs does not mask an obvious impact. Thus, the amount of data plays an important role in our proposed pipeline.

Convergence Speed. Given the pretrained distillation, the camera network requires fewer training epoches to reach a satisfying performance. As shown in Fig. 3b, approximately taking only half of the number of epochs, BEVFormer achieves higher performance than its original version.

4.5 Design Choices

The selective focus module generates weighted masks for the pretraining process. Intuitively, since the task of object detection concentrates on foreground objects, distilling object features refines the object-level representation. We design an object-level weight (Obj.), which is a heatmap indicating the locations of predicted objects. Another weight, the Den. weight, denotes the confidence based on point cloud density. As shown in Tab. 4a, both Obj. and Den. improve model performance compared to the baseline. However, Obj. only gains 1.6% of NDS while Den. improves mAP and NDS by 2.2% and 3.4%. The results demonstrate that the task of object detection not only focuses on foreground objects but also requires proper background information. Moreover, as the point cloud density varies in different locations, the maximal value of the weighted mask reaches about

Table 4: Comparisons among different design choices on nuScenes *val* set with BEVFusion-C.

(a) **Distillation Weight:** The Den. weight has a better improvement.

Obj.	Den.	mAP↑	NDS↑
		0.364	0.400
✓		0.364	0.416
	✓	0.386	**0.434**
✓	✓	**0.387**	0.421

(b) **Weight Scaling:** Values of weight that defer too much or too little bring negative effect.

Scaling	mAP↑	NDS↑
-	0.355	0.409
$sigmoid$	0.381	0.425
log	**0.386**	**0.434**

(c) **Pretraining Loss:** The choice of loss is important for distillation.

Loss	mAP↑	NDS↑
$\mathcal{L}_1$	0.334	0.348
$\mathcal{L}_2$	**0.386**	**0.434**
KL	0.376	0.410

(d) **Normalization:** Normalizing pretraining target causes significant difference.

Normalization	mAP↑	NDS↑
	0.361	0.391
✓	**0.386**	**0.434**

(e) **Duplication:** Duplicating detection head improves performance with free lunch.

Duplication	mAP↑	NDS↑
	0.380	0.411
✓	**0.386**	**0.434**

3,000. The extremely divergent distribution hinters the distillation process. As shown in Tab. 4b, predictions of the raw weighted mask (Row 1) are just slightly better than the baseline. We utilize the log and $sigmoid$ operations to scale the mask and find that log achieves a better performance.

Besides, as shown in Tab. 4c, we explore different distillation losses in the pretraining stage. Different from the common distillation setting where KL is used most frequently, in our case, the $\mathcal{L}_2$ loss achieves the highest performance. Instead of directly mimicking the LiDAR BEV feature, a channel-wise normalization is conducted to regularize feature maps. In Tab. 4d, the model performance benefits from the normalized BEV target by a large margin.

In the finetuning stage, Tab. 4e demonstrates the importance of duplicating the detection head from the expert to the apprentice. Credit to a better initialization of the detection head, an improvement of 0.6% mAP and 2.3% NDS can be observed.

4.6 Visualization

We present BEV representations of BEVFormer with and without our method in Fig. 4 in the appendix. As shown in the results, the proposed method refines the locations of predicted results and removes some false positive predictions. Visually, the salient heatmaps for each foreground instance, which might be caused by the inaccurate view transformation, are greatly refined, leading to more accurate detection results.

5 Discussion and Conclusion

In this work, we propose the PreDistill, which is an effective training pipeline to improve the performance of multi-camera methods on the task of 3D object detection. With its simplicity, our method serves as a plug-and-play module for various models. For its limitation, as currently we only consider annotated data for all training stages, unlabeled data in the nuScenes dataset remains unused. Moreover, our pretraining stage shows the ability to leverage large-scale LiDAR-Camera data pairs without object annotations for better performance, which is critical for autonomous driving.

References

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust lidar-camera fusion for 3d object detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022a. 4, 5

Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. *arXiv preprint arXiv:2208.12256*, 2022b. 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 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. 3

Hyeon Cho, Junyong Choi, Geonwoo Baek, and Wonjun Hwang. itKD: Interchange transfer-based knowledge distillation for 3d object detection. *arXiv preprint arXiv:2205.15531*, 2022. 3

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. 2, 3, 7

Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. LIGA-Stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *IEEE International Conference on Computer Vision*, 2021. 2, 3

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network (2015). *arXiv preprint arXiv:1503.02531*, 2, 2015. 1, 3

Junjie Huang and Guan Huang. BEVDet4D: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 4

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

Adhiguna Kuncoro, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, and Phil Blunsom. Syntactic structure distillation pretraining for bidirectional encoders. *Transactions of the Association for Computational Linguistics*, 8, 2020. 2, 3

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 4

Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019. 6

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*, 2022a. 2, 3

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*, 2022b. 1, 3, 4, 6

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*, 2022c. 1, 3, 4, 6

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. BEVFusion: A simple and robust lidar-camera fusion framework. *arXiv preprint arXiv:2205.13790*, 2022. 4, 6

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. 1

Yudong Liu, Yongtao Wang, Siwei Wang, TingTing Liang, Qijie Zhao, Zhi Tang, and Haibin Ling. CBNet: A novel composite backbone network architecture for object detection. In *AAAI Conference on Artificial Intelligence*, 2020. 6

Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H Hsu. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.04687*, 2021. 3

Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *IEEE International Conference on Computer Vision*, 2021. 7, 8

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*. Springer, 2020. 1, 3

Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *arXiv preprint arXiv:2102.00463*, 2021. 1

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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast pretraining distillation for small vision transformers. *arXiv preprint arXiv:2207.10666*, 2022. 2, 3

Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 1, 4

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

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022b. 3

Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 4

Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. PointDistiller: Structured knowledge distillation towards efficient and compact 3d detection. *arXiv preprint arXiv:2205.11098*, 2022. 3

Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

A VISUAL BEV REPRESENTATIONS

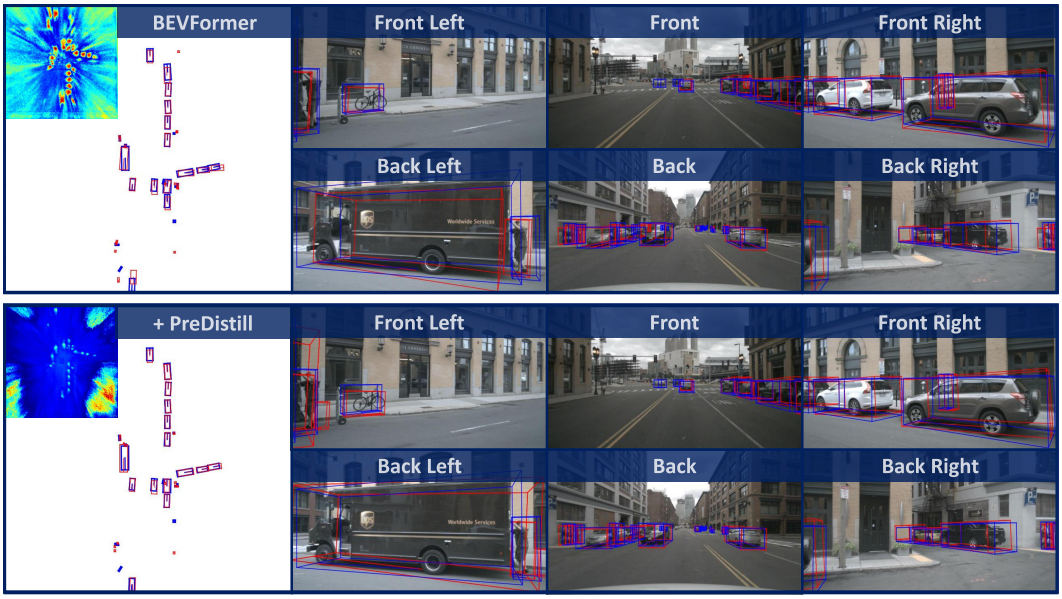


Figure 4: Qualitative results of BEV representations in BEVFormer. Blue and red boxes denote the ground truth and predictions respectively. The predictions of our method are more accurate.