A More Related Works

Depth Estimation. Depth estimation is a classical problem in computer vision. These method can be divided into single-view depth estimation and multi-view depth estimation. Single-view depth estimation is either regarded as a regression problem of a dense depth map or a classification problem of the depth distribution. (Bhat et al., 2021; Fu et al., 2018; Eigen et al., 2014; Poggi et al., 2020; Ranftl et al., 2021) generally build an encoder-decoder architecture to regress the depth map from contextual features. Multi-view depth estimation methods usually construct a cost volume to regress disparities based on photometric consistency (Wei et al., 2022; Guizilini et al., 2022; Zhang et al., 2019; Shen et al., 2021; Peng et al., 2022; Zhang et al., 2022d). For 3D object detection, previous methods (Park et al., 2021; Reading et al., 2021; Hong et al., 2022) also introduce additional networks for depth estimation to improve the localization accuracy in 3D space. Notably, MonoDETR (Zhang et al., 2022a;b) proposes to only predict the foreground depth maps instead of the dense depth values, but cannot leverage the advanced geometries provided by LiDAR modality. Different from them, our TiG-BEV conducts inner-depth supervision that captures local sptial structures of different foreground targets.

Knowledge Distillation has shown very promising ability in transferring learned representation from the larger model (teacher) to the smaller one (student). Prior works (Zagoruyko & Komodakis, 2016; Huang & Wang, 2017; Liu et al., 2021a; Tung & Mori, 2019a) are proposed to help the student network learn the structural representation for better generalization ability. Such teacher-student paradigms have also been extended to other vision tasks, including action recognition (Cui et al., 2020), video caption (Pan et al., 2020), 3D representation learning (Fu et al., 2022; Zhang et al., 2022c; Liu et al., 2021b; Sautier et al., 2022), object detection (Dai et al., 2021; Chen et al., 2017; Zhou et al., 2023) and semantic segmentation (Hou et al., 2022; Wang et al., 2020). However, only a few of works consider the multi-modal setting between different sensor sources. For 3D representation learning, there are some interesting approaches. I2P-MAE (Zhang et al., 2022c) leverages Masked Autoencoders to distill 2D pre-trained knowledge into 3D transformers. BEV-LGKD Li et al. (2022a) generates the foreground mask and view-dependent mask for better localization. BEVDistill (Chen et al., 2022b) transfer knowledge from LiDAR feature to the cam feature by dense feature distillation and sparse instance distillation. UniDistill (Zhou et al., 2023) focuses on transferring knowledge from multi-modality detectors to single-modality detectors in a universal manner. X^{3}KD (Klingner et al., 2023) is a knowledge distillation framework for multi-camera 3D object detection, leveraging cross-modal and cross-task information by distilling knowledge from LiDAR-based detectors and instance segmentation teachers. DistillBEV (Wang et al., 2023) involves feature imitation and attention imitation losses across multiple scales, enhancing feature alignment between a LiDAR-based teacher and a multi-camera BEV-based student detector. BEVSimDet (Zhao et al., 2023) proposes a simulated multi-modal student to simulate multi-modal features with image-only input. VCD (Huang et al., 2023) presents some useful designs for temporal fusion and introduce a fine-grained trajectory-based distillation module. FD3D (Zeng et al., 2023) uses queries for masked feature generation and then intensify feature representation for refined distillation.

Our TiG-BEV also follows such teacher-student paradigm and effectively distills knowledge from the LiDAR modality into the camera modality. By modeling the relative relationships inside foreground objects in a novel way, the performances of camera-only detectors are further enhanced.

Relationship Supervision. Some related works have investigated channel-wise and pixel-wise relationship supervision in various domains. (Gatys et al., 2016) studied pixel-wise relationships in image style transfer and found that matching higher layer style representations preserves local image structures at a larger scale, resulting in smoother visuals. (Tung & Mori, 2019b) proposed similarity-preserving knowledge distillation, guiding the student network towards teacher network's activation correlations. If two inputs produce similar activations in the teacher network, the student network should be guided towards a similar configuration. (Hou & Zheng, 2021) introduced a channel-wise relationship preserving loss for visualizing adapted knowledge in domain transfer. They claimed that channel-wise relationships remain effective after global pooling, unlike pixel-wise relationships, which can be overshadowed pre-classifier. Our TiG-BEV has also been inspired by these works and explored the designs of learning the internal relationships of foreground objects.

Table 7: **Comparison with BEVDistill (Chen et al., 2022b).** † and * denote the implementation of BEVDistill and ours, respectively. We present the performance improvement of the learning methods correspondingly to their implemented baselines.

Method	mAP↑	NDS↑
BEVDepth†	0.311	0.432
+ BEVDistill	0.332 (+2.1%)	0.454 (+2.2%)
BEVDepth*	0.329	0.431
+ Naive Distill	0.338 (+0.9%)	0.434 (+0.3%)
+ Inner-feature Distill	**0.359 (+3.0%)**	**0.454 (+2.3%)**
+ TiG-BEV	**0.366 (+3.7%)**	**0.461 (+3.0%)**

Table 8: **Comparison with Concurrent Works.** We present the performance improvement of some concurrent works, UniDistill (Zhou et al., 2023), X^{3}KD (Klingner et al., 2023), DistillBEV (Wang et al., 2023), correspondingly to their implemented baselines which uniformly use ResNet-50 as backbone with the image resolution of 256 × 704.

Student	mAP↑	NDS↑	Method	mAP↑	NDS↑	Venue
BEVDet	0.203	0.331	UniDistill	0.260(+5.7%)	0.373(+4.2%)	CVPR 2023
BEVDet	0.305	0.378	DistillBEV	0.327(+2.2%)	0.407(+2.9%)	ICCV 2023
BEVDet	0.298	0.379	TiG-BEV	0.331(+3.3%)	0.411(+3.2%)	Ours
BEVDet4D	0.328	0.459	DistillBEV	0.363(+3.5%)	0.484(+2.5%)	ICCV 2023
BEVDet4D	0.322	0.451	TiG-BEV	0.356(+3.4%)	0.477(+2.6%)	Ours
BEVDepth	0.359	0.472	X^{3}KD$_{modal}$	0.368(+0.9%)	0.494(+2.2%)	CVPR 2023
BEVDepth	0.364	0.484	DistillBEV	0.389(+2.5%)	0.498(+1.4%)	ICCV 2023
BEVDepth	0.357	0.481	TiG-BEV	0.383(+2.6%)	0.498(+1.7%)	Ours

B More Experiments

In this section, we further conduct a series of experiments to show the effectiveness of our approach.

B.1 Dataset

NuScenes dataset (Caesar et al., 2020) provides synced data captured from a 32-beam LiDAR at 20Hz and six cameras covering 360-degree horizontally at 12Hz. We adopt the official evaluation toolbox provided by nuScenes, which reports the nuScenes Detection Score (NDS) and mean Average Precision (mAP), along with 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).

B.2 Main Results

Comparison with BEVDistill (Chen et al., 2022b). In Table 7, we compare our TiG-BEV with another LiDAR-to-camera learning method BEVDistill in the same setting. As shown, on top of a better baseline model, our approach can achieve higher performance boost for both mAP and NDS. Besides, we compare with the naive distillation that directly applies MSE loss to the entire BEV features between camera and LiDAR, where our inner-feature distillation performs better. These well demonstrate the superiority of target inner-geometry learning to foreground-guided dense distillation.

Comparison with Concurrent Works. As shown in Table 8, we further compare our TiG-BEV with other recent works in the same setting which also belong to LiDAR-to-camera learning meth-

Table 9: **Performance Comparison without CBGS (Zhu et al., 2019).** For all methods, we adopt ResNet-101 as the 2D backbone and 512 × 1408 as the image resolution. * denotes our implementation.

Method	mAP↑	NDS↑
BEVDet*	0.272	0.297
+ TiG-BEV	**0.375** (**+10.3%**)	**0.388** (**+9.1%**)
BEVDet4D*	0.336	0.435
+ TiG-BEV	**0.409** (**+7.3%**)	**0.489** (**+5.4%**)
BEVDepth*	0.393	0.487
+ TiG-BEV	**0.430** (**+3.7%**)	**0.514** (**+2.7%**)

Table 10: **Performance Comparison on KITTI Val Set (Geiger et al., 2012).**

Method	3D AP		
	Easy	Moderate	Hard
CMKD (Hong et al., 2022)	23.53	16.33	14.44
+ TiG-BEV	26.63	16.61	14.31

ods. As there is a lack of consistency in the baselines reproduced by everyone, we have also included the baseline in the table for comparison. It should be noted that, the baseline performance of UniDistill is unexpectedly low, making it difficult to find an appropriate setting to compare with. Besides it, our proposed method has been demonstrated to be simple yet efficient and highly competitive, achieving comparable performance to the latest state-of-the-art methods.

Without CBGS (Zhu et al., 2019) Strategy. In Table 9, we present the results of TiG-BEV without the CBGS training strategy. Without the resampling of training data, the performance improvement of learning target inner-geometry becomes more notable, **+10.3%, +7.3%,** and **+3.7%** mAP for the three baselines, which indicates the superior LiDAR-to-camera knowledge transfer of our TiG-BEV.

Performance on KITTI Val Set (Geiger et al., 2012). In addition, we also conduct an experiment on KITTI validation set which is one of the most popular datasets for monocular 3D object detection. The main baseline is CMKD (Hong et al., 2022), which is a state-of-the-art method on KITTI dataset and use a classical distillation method to transfer knowledge of teacher model. We add constraints on knowledge distillation of target objects with our TiG-BEV for further optimisation and achieve a better performance than the baseline.

Visualization. More visualization results of BEVDepth before and after our TiG-BEV are shown in Fig. 9. With the help of our inner-geometry learning, the detection of false positives and ghost objects can be reduced, and more missed objects have been detected. The most obvious and common improvement is that locations and orientations of the bounding boxes are further refined and they are more consistent with the ground-truth boxes within orange marks. What's more, we visualize our depth prediction with and without the inner-depth supervision in Figure 8, which effectively refines the contours and edges of foreground objects. And the conclusion implied in the visualization of predicted depth maps is consistent with Table 6.

B.3 Ablation Study

2D Backbones and Temporal Information. We further explore the influence of 2D backbones and temporal information to our TiG-BEV in Table 11. We observe that our TiG-BEV brings significant performance improvement consistent over different 2D backbones. Also, our target inner-

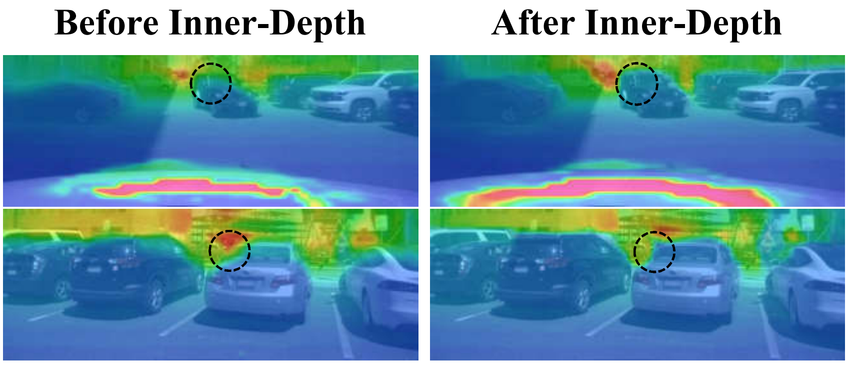

Figure 8: **Visualization of Predicted Depth Maps,** which are before and after the inner-depth supervision, respectively.

Table 11: **Ablation Study of 2D Backbones and Temporal Information.** CenterPoint (Yin et al., 2021) and BEVDepth (Li et al., 2022b) are adopted as the teacher and student models, respectively.

Backbone	Resolution	Multi-frame	Method	mAP	NDS
VoxelNet	-	✓	Teacher	0.564	0.646
ResNet-18	256 × 704	✓	Student + TiG-BEV	0.285 **0.323** (**+3.8%**)	0.405 **0.430** (**+2.5%**)
			Student + TiG-BEV	0.260 **0.294** (**+3.4%**)	0.295 **0.335** (**+4.0%**)
ResNet-50	256 × 704	✓	Student + TiG-BEV	0.329 **0.366** (**+3.7%**)	0.431 **0.461** (**+3.0%**)
			Student + TiG-BEV	0.298 **0.338** (**+4.0%**)	0.328 **0.375** (**+4.7%**)
ResNet-101	512 × 1408	✓	Student + TiG-BEV	0.393 **0.430** (**+3.7%**)	0.487 **0.514** (**+2.7%**)
			Student + TiG-BEV	0.345 **0.403** (**+5.8%**)	0.366 **0.416** (**+5.0%**)

geometry learning schemes can provide positive effect for both single-frame and multi-frame settings. The improvement of mAP ranges from **+3.4%** to **+5.8%** and the improvement of NDS ranges from **+2.5%** to **+5.0%**.

Performance of Small Object Detection. In Table 12, we explored the impact of each component of our method on small object detection. We consider pedestrian, motorcycle, bicycle, traffic cone and barrier, these five classes in NuScenes to be relatively small objects, and we calculated their detection results in terms of mAP on the validation set. This quantitative analysis further confirms that our method is also effective in detecting small objects and provides gains in performance with each component.

Performace of Different Distance Ranges. The detection of distant objects remains a long-standing challenge due to the sparsity of LiDAR points and the inaccuracy of depth estimation. Here we define 0 to 30 meters as close detection, and 30 to 60 meters as long-distance detection. As can be seen from the Table 13, our method can greatly improves the performance of close detection, and it also works even in the face of long-distance object detection with each designed module. In the future work, we will continue to pay attention to and strive to solve the problem of long-distance object detection.

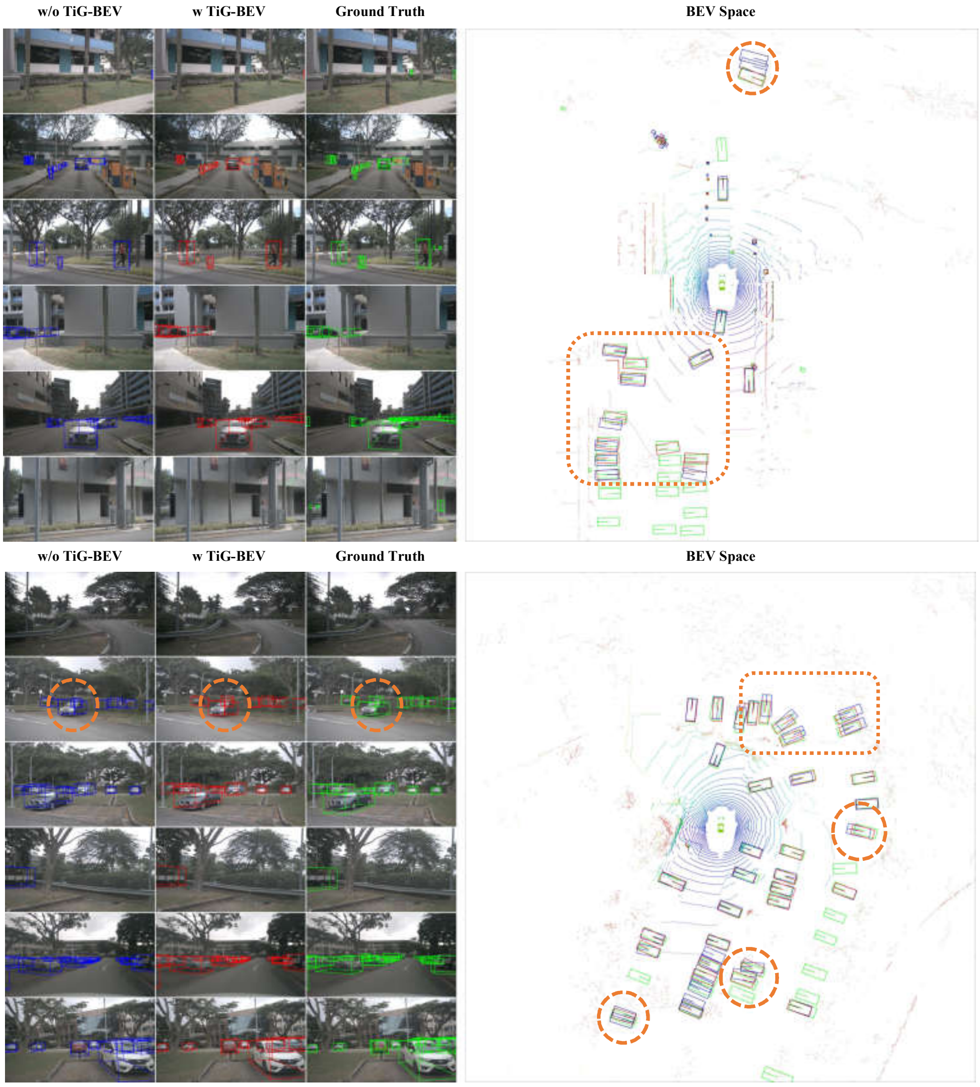

Figure 9: **Visualization of Detection Results**. From left to right, we show the 3D object detection before and after the TiG-BEV learning schemes, ground-truth annotations, along with the overall BEV-space results.

Table 12: **Ablation Study of Small Objects Detection Performance.** We use BEVDet4D as baseline and ResNet-101 as backbone with the image resolution of 512 × 1408.

$\mathcal{L}_{\text{depth}}^{A}$	$\mathcal{L}_{\text{depth}}^{R}$	$\mathcal{L}_{\text{bev}}$	mAP↑ (%)				
			pedestrian	motorcycle	bicycle	traffic_cone	barrier
			43.3	34.5	32.4	54.8	55.2
✓			44.6	35.6	32.8	56.3	55.9
✓	✓		45.4	37.3	33.7	57.8	58.8
✓	✓	✓	**45.7**	**38.8**	**37.5**	**58.9**	**57.2**

Table 13: **Ablation Study for Objects Detection Performance of Different Distance Ranges.** We use BEVDet4D as baseline and ResNet-101 as backbone with the image resolution of 512 × 1408.

Distance (m)	$\mathcal{L}^{A}_{\mathrm{depth}}$	$\mathcal{L}^{R}_{\mathrm{depth}}$	$\mathcal{L}_{\mathrm{bev}}$	mAP↑	NDS↑
[0,30)				44.2	53.7
[0,30)	✓			46.6	55.6
[0,30)	✓	✓		47.4	55.8
[0,30)	✓	✓	✓	**48.6**	**56.8**
[30,60)				10.2	28.7
[30,60)	✓			10.3	28.5
[30,60)	✓	✓		11.7	30.1
[30,60)	✓	✓	✓	**12.6**	**30.2**

B.4 Discussion

Improvements of Inner-depth Supervision vs. Inner-feature Distillation. Our inner-depth provides fine-grained depth cues for object-level geometry understanding, which is low-level information and implicitly benefits the mAP accuracy. The inner-feature directly supervises high-level semantics of BEV features, explicitly affecting the detection and recognition performance. They are complementary and can collaborate for better results.

Limitations and Future Works of Inner-depth Supervision. One limitation that will inevitably be involved in depth-related explicit supervision is the ground truth from LiDAR. Due to the sparsity of LiDAR, both absolute depth supervision and relative depth supervision will be affected to some extent, though the performance of object detection will be better with inner-depth supervision under the same conditions. For sparse point clouds, a good way to alleviate it is through depth completion.

Another limitation is the occlusion problem of the target. Due to visual occlusion, the ground truth of the occluded object that we can obtain is limited, which will affect the learning of depth prediction and inner-depth learning. A good way to alleviate it is to consider multi-frame images for inner-depth supervision of the same object's interior. But a potential risk here is that if the intrinsic and extrinsic parameters are inaccurate, it will directly affect the coordinate system transformation between multiple frames, thereby affecting the ground truth of depth values.