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

Supplementary Material

Due to the page constraint of the main paper, we provide more quantitative and qualitative results in this supplementary material, which is organized as follows:

- Dataset description of the KITTI raw set in Section A.
- The implementation and training details for the KITTI3D and nuScenes datasets are documented in Section B.
- Justification and analysis of the TA model in Section C.
- More quantitative results for MonoTAKD in Section D.
- More ablation studies for MonoTAKD in Section E.
- Qualitative results for MonoTAKD in Section F.

A. Datasets

KITTI Raw. The KITTI Raw dataset includes approximately 48K unlabeled data used for semi-supervised training. Following [12, 29], we train on the Eigen clean subset (22K) of the KITTI raw dataset and evaluate on the KITTI test set (3,769). The evaluation metric and the implementation of KITTI raw are the same as the KITT3D dataset.

B. Implementation Details

For the KITTI3D dataset, we use a pre-trained Second [45] as the LiDAR-based teacher. Both the camera-based TA and camera-based student are derived from CaDDN [32], using ResNet50 as their backbone. In addition, we use PointPillar [20] as the BEV detector. Initially, we trained the TA model using a pre-trained model for 5 epochs. Then, a pre-trained teacher model and a frozen TA model are used to train the student model for another 60 epochs. Training is performed with an NVIDIA Titan XP GPU in an end-to-end manner. We set the batch size to 2, and the learning rate is $2e^{-4}$ with the one-cycle learning rate strategy. The IoU thresholds for the Car, Pedestrian, and Cyclist categories are 0.7, 0.5, and 0.5, respectively. As for the discrete depth bins D, we set D to 120, and the minimum and maximum depths are set to 2.0 and 46.8 meters, respectively.

In the case of the nuScenes dataset, we adopt a pre-trained CenterPoint [48] as the LiDAR-based teacher and use BEVDepth [22] for both the TA and the student. Due to a higher resolution and a larger model size, we set the batch size to 8 and trained the models with eight NVIDIA V100 GPUs. We set the learning rate of $2e^{-4}$ with a multi-step learning rate decay schedule and a decay rate of 0.1 and train the model for 25 epochs.

C. Justification and Analysis of TA

Novelty of the TA. Unlike previous TAKD [26], relying on step-by-step distillation, we bypass this and simultaneously distill complementary knowledge to the student: 3D visual knowledge from the camera-based TA and precise LiDAR-exclusive 3D features from the LiDAR-based teacher. This approach presents a novel solution to the cross-modal distillation problem, which goes beyond addressing the differences in the model's architecture. Experimental results show that MonoTAKD outperforms TAKD by 4.5%, 3.8%, and 2.7% in AP_{3D} for easy, moderate, and hard scenarios, as step-by-step distillation cannot bridge the modality gap and also complicates the training procedure.

Quality and complexity of TA. To ensure high-quality features from the TA model, we fine-tune it starting from a pre-trained camera-based model for 5 epochs, achieving rapid convergence within 3 hours (simple training procedure) due to the incorporation of the GT depth, as shown in Table C1. Additionally, since the TA model is excluded during inference, it does not affect the student's inference time.

Table C1. Performance of our teacher model $\mathcal{T}$ and teaching assistant model $\mathcal{A}$. † represents the incorporation of the GT depth.

Model	Epochs	Training Time (hr)	AP_{3D}		
			Easy	Mod.	Hard
$\mathcal{T}$	N/A	pre-trained	87.68	76.32	73.28
$\mathcal{A}$	N/A	pre-trained	23.47	16.31	13.84
$\mathcal{A}^{\dagger}$	N/A	pre-trained	54.84	35.44	30.45
$\mathcal{A}^{\dagger}$	5	3	62.91	43.35	34.99
$\mathcal{A}^{\dagger}$	10	6	62.83	42.98	34.82

Applicability of TA. One concern is whether depth maps are always available for training the TA model. Most autonomous driving datasets, including KITTI3D, nuScenes, and Waymo, provide 3D detection labels derived from LiDAR point clouds, which can be readily converted into GT depth maps for TA training. However, when depth maps are not directly accessible (e.g., radar 3D object detection), distance information can be used as an alternative.

In summary, the overall performance, considering AP, training complexity, and model complexity, provides a superior solution compared to the existing Mono3D approach. Further discussion can be found in section 5.

D. More Quantitative Results

Results for Pedestrian and Cyclist. We present a detailed comparison with other state-of-the-art methods for the non-car categories on the KITTI test set. Table D2 demonstrates that MonoTAKD outperforms other methods not only in the

Table D2. Experimental results for Pedestrian and Cyclist categories on the KITTI *test* set. We use **bold** and underline to indicate the best and the second-best results, respectively.

Method	Venue	Pedestrian AP_{3D}/AP_{BEV}			Cyclist AP_{3D}/AP_{BEV}		
		Easy	Mod.	Hard	Easy	Mod.	Hard
MonoATT [51]	CVPR 2023	10.55/11.63	6.66/7.40	5.43/6.56	5.74/6.73	3.68/4.44	2.94/3.75
Cube R-CNN [2]	CVPR 2023	11.17/11.67	6.95/7.65	5.87/6.60	3.65/5.01	2.67/3.35	2.28/3.32
CaDDN [32]	CVPR 2021	12.87/14.72	8.14/9.41	6.76/8.17	7.00/9.67	3.41/5.38	3.30/4.75
DD3D [27]	ICCV 2021	13.91/15.90	9.30/10.85	8.05/8.05	2.39/3.20	1.52/1.99	1.31/2.39
MonoNerd [43]	ICCV 2023	13.20/15.27	8.26/9.66	7.02/8.28	4.79/5.24	2.48/2.80	2.16/2.55
MonoUNI [17]	NeurIPS 2023	15.78/16.54	10.34/10.90	8.74/9.17	7.34/8.25	4.28/5.03	3.78/4.50
OccupancyM3D [30]	CVPR 2024	14.68/16.54	9.15/10.65	7.80/9.16	7.37/8.58	3.56/4.35	2.84/3.55
MonoTAKD	-	**16.15/19.79**	**10.41/13.62**	**9.68/11.92**	**13.54/16.90**	**7.23/9.42**	**6.86/8.29**

Table D3. Experimental results on the KITTI *test* set for the Car category, leveraging unlabeled data. We use **bold** and underline to indicate the best and the second-best results, respectively.

Method	Venue	Extra Data	AP_{3D}			AP_{BEV}		
			Easy	Mod.	Hard	Easy	Mod.	Hard
LPCG [29]	ECCV 22	Raw	25.56	17.80	15.38	35.96	24.81	21.86
Mix-Teaching [46]	CSVT 23		26.89	18.54	15.79	35.74	24.23	20.80
CMKD [12]	ECCV 22		28.55	18.69	16.77	38.98	25.82	22.80
MonoTAKD	-	Raw	**29.86**	**21.26**	**18.27**	**43.83**	**32.31**	**28.48**

Car category but also in the Pedestrian and Cyclist categories. This success indicates that the approach is well-suited for a broad range of autonomous driving applications, including tasks like trajectory prediction.

Results on KITTI raw. To improve the transferability and to generalize the application of MonoTAKD on real-world scenes, we explore the performance of MonoTAKD in a semi-supervised manner. As illustrated in Table D3, our MonoTAKD outperforms CMKD in AP_{3D}/AP_{BEV} across all three difficulty levels, respectively.

Owing to MonoTAKD's outstanding performance in semi-supervised settings, it is evident that our distillation method adeptly extracts valuable 3D features from unlabeled data. Thus, MonoTAKD can provide comprehensive guidance for the student model across all difficulty levels.

E. More Ablation Studies

Backbone choices on KITTI3D. We analyze the backbone choice of our MonoTAKD in Table E4. According to the table, Swin-T, a transformer-based backbone, exhibits higher FLOPs and underperforms in both speed and accuracy. We believe the performance drop is because of the heterogeneity of the architecture between teacher and student (CNN and Transformer). Conversely, MobileNetV3, a lightweight backbone, excels in speed and efficiency with lower FLOPs but has a trade-off with lower accuracy.

After comparing ResNet50 and ResNet101, we determined that ResNet50 is the optimal backbone for the student model, delivering enhanced performance with higher AP, improved FPS, and reduced FLOPs. This finding highlights that in Mono3D tasks, a larger or more complex backbone does not necessarily translate to better performance. Note that, we only compare the FLOPs of the backbone. The total FLOPs can be found in Table 6.

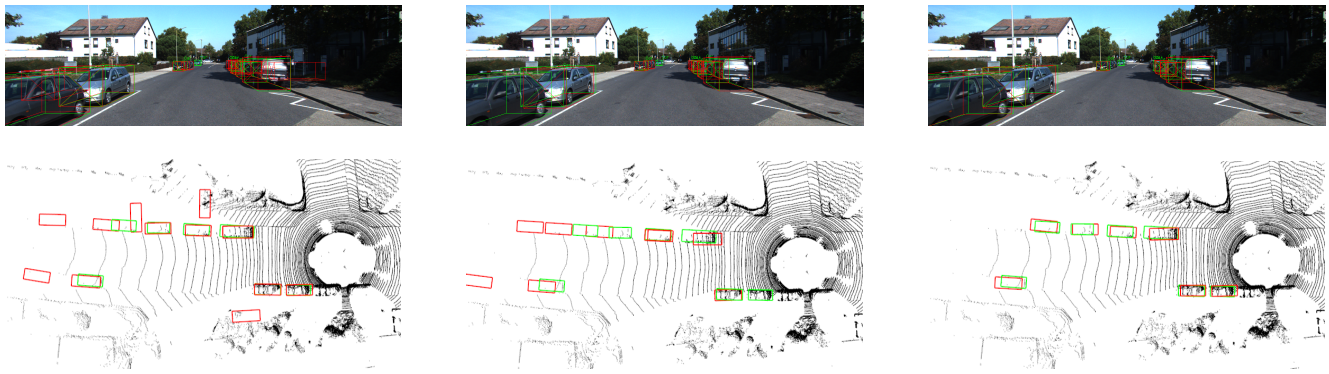
Table E4. Comparison of MonoTAKD with different backbones.

Backbone	Speed (FPS)	FLOPs (G)	AP_{3D}		
			Easy	Mod.	Hard
Swin-T	5.8	16.7	31.57	19.33	17.65
MobileNetV3	13.8	3.4	26.11	16.87	13.92
ResNet101	9.2	4.3	33.07	21.54	19.16
ResNet50	**11.9**	**4.1**	**34.36**	**22.61**	**19.88**

F. Qualitative Results

We compare our visualization results with state-of-the-art methods, CMKD [12] and MonoDETR [49], for both 3D object and BEV detection in Fig. F1. MonoTAKD comparatively has the best-fitted bounding box size estimation and the most accurate 3D localization among the three methods.

Lastly, Fig. F2 presents the BEV features of the teacher, TA, and the student. Notably, the student's BEV image exhibits distortion and blurriness. However, with the help of SAM and FFM modules, the student's BEV features successfully align more closely to resemble the BEV LiDAR features. This visual comparison illustrates how the proposed approaches collectively contribute to improving the student's 3D perception.



(a) CMKD (b) MonoDETR (c) MonoTAKD (ours)

Figure F1. Qualitative results on KITTI *val* set for the Car category. We compare the qualitative results among CMKD [12], MonoDETR [49], and our proposed MonoTAKD. The first and second rows represent detection results from a camera frontal view and a BEV, respectively. We use green and red boxes to indicate the ground truth and prediction bounding boxes.

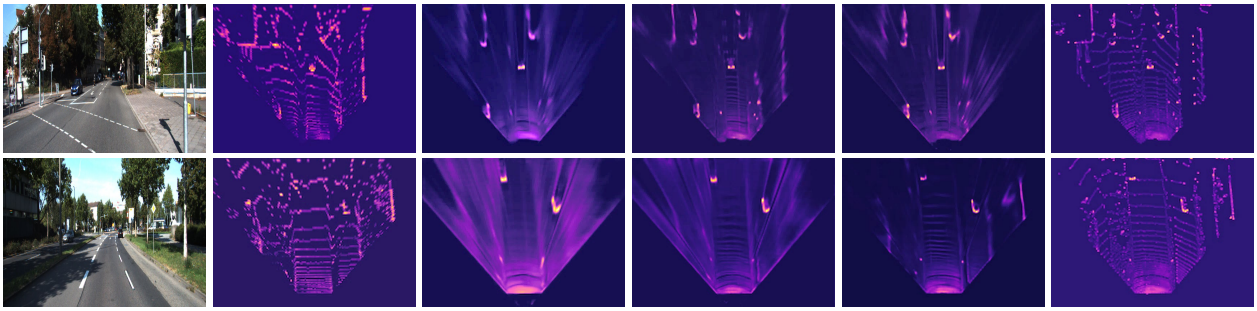


Figure F2. Visualization of the BEV features from the teacher, TA, and the two distillation branches of the student model on KITTI *val* set.