A. Additional Dataset Analysis

A.1. Detailed Dataset Analisis

Comparison in the object density against other datasets. We further compare the average number of 3D annotations per frame of different datasets. As is demonstrated in Table 5, we compute the density for A*3D Dataset and borrow the statistics from CityScapes 3D [11] for the following datasets: KITTI, ApolloScapes, Argoverse, nuScenes, Waymo, and CityScapes 3D. Compared with other datasets, we have a high object density across all classes.

	Car	Big Vehicle	Cyclist	Pedestrian	All
KITTI [12]	4.2	0.2	0.0	0.0	4.40
ApolloScapes [16]	11.6	0.0	0.0	0.0	11.60
Argoverse [7]	4.1	0.3	0.1	0.001	4.50
nuScenes [5]	3.0	0.6	0.07	0.07	3.74
Waymo [38]	3.2	0.0	0.04	0.0	3.24
CityScapes 3D [11]	6.4	0.2	1.2	0.2	8.0
A*3D [32]	3.9	**1.0**	0.3	0.6	5.8
Ours	**14.0**	0.6	**3.9**	**5.5**	**24.0**

Table 5. The average number of 3D annotations in each image at coarse-level. Compared with other datasets, we have a high object density across all classes.

Size and orientation. Only motor vehicles are taken into account for size analysis, *i.e*., cars, and big vehicles since non-motor categories usually have similar sizes. The size and orientation distributions are presented in Fig. 9. Due to various camera specifications and diverse scenes, the high-frequency orientations are not constrained to a single peak.

The Mean and Std Dev of fine-grained categories. We further compute the mean and standard deviation (Std Dev) of each fine-grained category, which is presented in Table 6. The mean and Std Dev values can be utilized for predefining the mean size and the disturbance range. For example, Monoflex [50] estimates the offset of length, width, and height w.r.t the mean values, instead of directly regressing the sizes, which improves the robustness and accuracy of size prediction.

A.2. More samples of the Rope3D Dataset.

We present more roadside data samples for visualization in Fig. 10, including different weather conditions, collecting time and object densities.

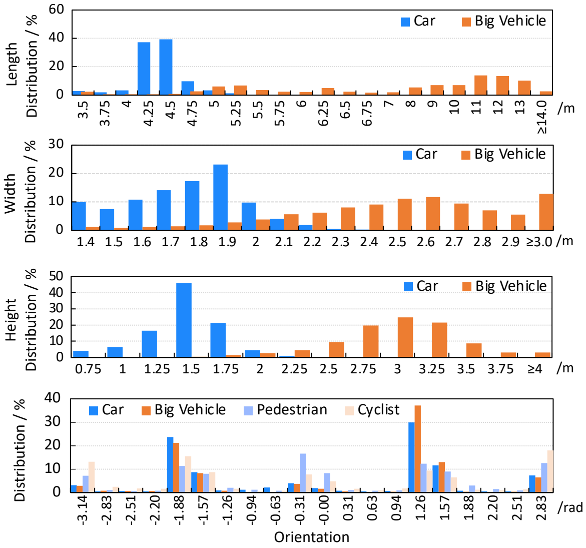

Figure 9. From top to bottom are the distribution of length, width, height and orientation over the motor vehicles, respectively. Only the 3D sizes of car and big vehicles are counted, since the sizes of pedestrian and cyclists only have little changes.

Category	Metric	Length / m	Height / m	Width / m
Car	mean	4.247	1.325	1.706
	Std Dev	0.315	0.258	0.234
Truck	mean	7.122	2.623	1.706
	Std Dev	2.067	0.628	0.492
Van	mean	4.651	1.750	1.757
	Std Dev	0.429	0.311	0.268
Bus	mean	10.575	3.009	2.533
	Std Dev	1.806	0.404	0.426
Pedestrian	mean	0.478	1.610	0.501
	Std Dev	0.178	0.160	0.143
Cyclist	mean	1.525	1.382	0.505
	Std Dev	0.264	0.280	0.217
Tricyclist	mean	2.631	1.539	1.077
	Std Dev	0.497	0.196	0.292
Motorcyclist	mean	1.692	1.418	0.613
	Std Dev	0.276	0.175	0.211

Table 6. The Mean and Std Dev size of each fine-grained category over the Rope3D Dataset.

B. Additional Experiments

In this section, we show more experimental results. As is stated, we offer two kinds of validation sets, the homologous (I) in which the training and validation set have common scenes, and the heterologous (II) with the validation set has never seen the scenes in the training set and the camera specifications are possibly different with the training set.

Performance of pedestrian and cyclist. In addition to the results of motor vehicles in the main paper, we further present the results of pedestrians and cyclists under the homologous and heterologous settings in Table 7 for further evaluation. Monoflex [50] obtains superior performance especially on cyclist and pedestrian categories, which shows consistent behavior with the original work.

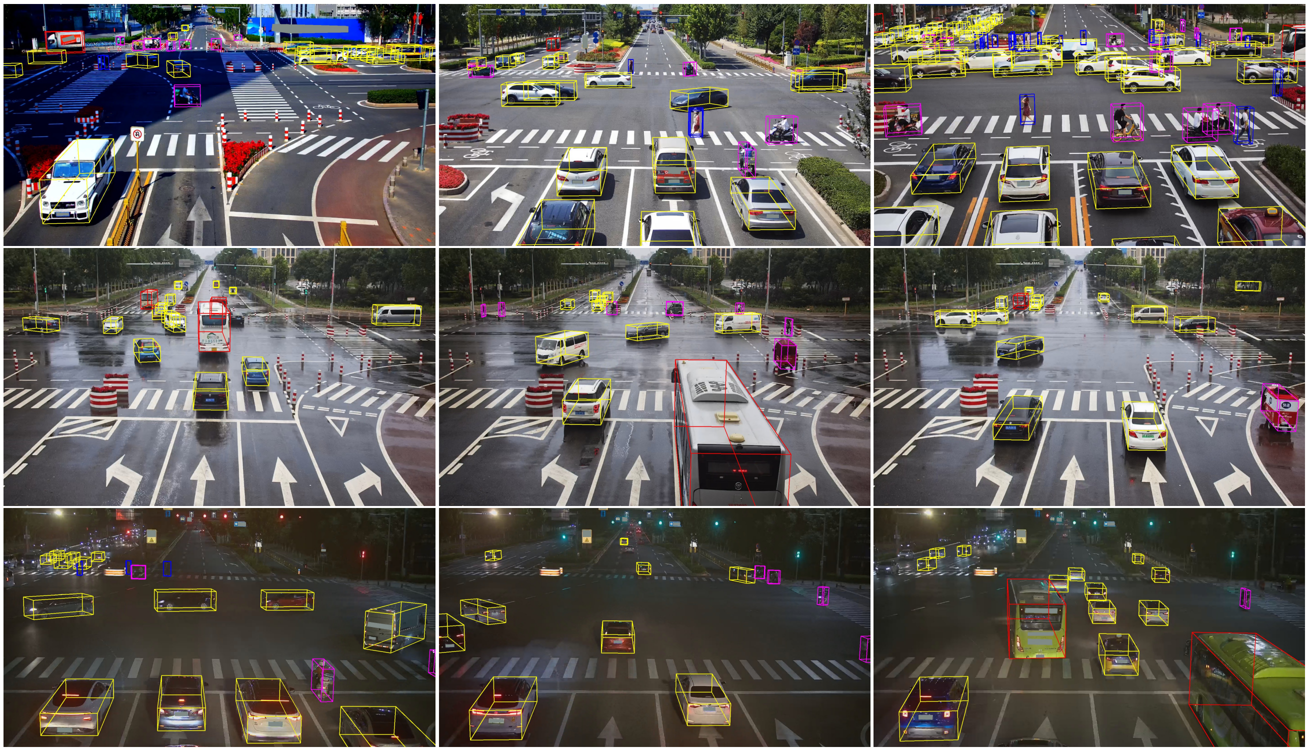

Figure 10. More collected examples. From top to bottom, each row corresponds to clear/sunny/cloudy, rainy and dawn/dusk.

Setting	Method	Backbone	Branch	IoU = 0.25				IoU = 0.5			
				Cyclist		Pedestrian		Cyclist		Pedestrian	
				$\text{AP}_{3\text{D}\|\text{R40}}$	$\text{Rope}_{\text{score}}$	$\text{AP}_{3\text{D}\|\text{R40}}$	$\text{Rope}_{\text{score}}$	$\text{AP}_{3\text{D}\|\text{R40}}$	$\text{Rope}_{\text{score}}$	$\text{AP}_{3\text{D}\|\text{R40}}$	$\text{Rope}_{\text{score}}$
$\mathbb{I}$	M3D-RPN-($\boldsymbol{G}$) [3]	ResNet34	A	12.45	28.64	2.29	20.07	2.61	20.79	0.34	18.63
	M3D-RPN-($\boldsymbol{D}$) [3]	ResNet34	A	22.26	36.61	6.98	24.00	5.64	23.35	1.16	19.47
	Kinematic3D-($\boldsymbol{G}$) [4]	DenseNet121	A	14.78	29.72	3.59	21.19	2.97	2.34	0.52	18.92
	MonoDLE-($\boldsymbol{G}$) [25]	DLA-34	K	24.26	37.35	4.14	21.85	4.68	21.70	0.44	18.91
	MonoFlex-($\boldsymbol{G}$) [50]	DLA-34	K	65.63	70.78	36.83	48.10	24.25	37.70	7.58	24.70
$\mathbb{III}$	M3D-RPN-($\boldsymbol{G}$) [3]	ResNet34	A	5.07	22.42	1.40	19.40	0.75	19.02	0.25	18.54
	M3D-RPN-($\boldsymbol{D}$) [3]	ResNet34	A	11.22	27.54	3.93	21.54	2.09	20.25	0.67	19.08
	Kinematic3D-($\boldsymbol{G}$) [4]	DenseNet121	A	4.84	21.15	2.98	20.52	0.72	17.94	0.73	19.02
	MonoDLE-($\boldsymbol{G}$) [25]	DLA-34	K	10.93	26.44	3.72	21.42	2.02	19.32	0.47	18.86
	MonoFlex-($\boldsymbol{G}$) [50]	DLA-34	K	44.27	53.58	25.48	39.04	12.30	28.00	4.29	22.09

Table 7. Performance of the Pedestrian and Cyclist on the Rope3D Dataset with IoU = 0.25 and 0.5 under two train-val splitting settings: the homologous ($\mathbb{I}$) and the heterologous ($\mathbb{III}$). -($\boldsymbol{G}$) denotes adapting the ground plane, -($\boldsymbol{D}$) means using the depth map of ground. The abbr. in the branch column denotes: A: anchor-based, K: keypoint-based.

Performance of fine-grained categories. We conduct two levels of categorization and the corresponding experiments. For the coarse-grained level, the monocular 3D object detection task mainly focuses on the most common traffic elements: Car, Big Vehicle, Pedestrian, and Cyclist. For fine-grained level, Car includes car and van, Big Vehicle can be further divided into truck and bus, and meanwhile, Cyclist can be subdivided into cyclist, motorcyclist, and tricyclist as they are driving non-motor vehicles. The performances of fine-grained-level-8 are compared in Table 8.

Performance of vanilla and improved approaches. Leveraging the depth map of ground plane, we try to alleviate the ambiguity caused by different camera specifications. For this purpose, we evaluate two approaches to incorporate depth information with the RGB appearance feature. The first one is directly concatenating the depth map with the original RGB channels as input, and the second is adopting another siamese network for depth feature extraction and further weighted fusion of the two depth predictions. The performances of these two methods are similar and we

Setting	Method	Backbone	$\text{AP}_{3D\|R40}$[Mod] / $\text{Rope}_{\text{score}}$							
			car	van	bus	truck	cyclist	motorcyclist	tricyclist	pedestrian
𝕀	M3D-RPN-$(\boldsymbol{G})$ [3]	ResNet34	41.15 / 52.38	31.19 / 44.31	32.60 / 44.58	26.54 / 39.89	6.48 / 23.80	10.23 / 26.84	20.81 / 35.39	2.01 / 19.87
	M3D-RPN-$(\boldsymbol{D})$ [3]	ResNet34	64.38 / 71.04	48.56 / 58.33	41.67 / 52.06	39.14 / 50.09	16.64 / 32.11	24.46 / 38.41	41.77 / 52.40	6.22 / 23.42
	Kinematic3D-$(\boldsymbol{G})$ [4]	DenseNet121	48.42 / 57.32	34.13 / 45.86	21.71 / 35.43	32.30 / 43.46	8.45 / 24.84	18.66 / 32.77	28.66 / 40.99	3.25 / 20.90
	MonoDLE-$(\boldsymbol{G})$ [25]	DLA-34	77.76 / 81.11	67.52 / 72.74	66.24 / 71.57	54.74 / 61.33	58.64 / 65.27	65.51 / 70.55	73.62 / 77.12	41.68 / 52.02
	MonoFlex-$(\boldsymbol{G})$ [50]	DLA-34	51.89 / 60.41	54.18 / 62.02	47.17 / 56.24	53.18 / 60.74	58.41 / 65.37	67.30 / 72.10	69.67 / 73.74	26.72 / 40.02
𝕀𝕀	M3D-RPN-$(\boldsymbol{G})$ [3]	ResNet34	15.51 / 31.51	5.96 / 23.65	23.74 / 37.29	7.50 / 23.94	1.79 / 19.83	3.41 / 20.87	10.75 / 27.14	1.78 / 19.62
	M3D-RPN-$(\boldsymbol{D})$ [3]	ResNet34	34.25 / 46.74	22.45 / 37.18	57.90 / 65.04	27.30 / 40.42	21.58 / 36.00	15.08 / 30.58	19.75 / 34.80	5.14 / 22.53
	Kinematic3D-$(\boldsymbol{G})$ [4]	DenseNet121	22.38 / 36.20	10.13 / 26.42	22.25 / 35.34	9.86 / 25.33	2.54 / 19.79	5.52 / 21.56	14.25 / 29.98	1.66 / 19.33
	MonoDLE-$(\boldsymbol{G})$ [25]	DLA-34	25.78 / 39.30	15.80 / 31.00	60.22 / 66.26	16.47 / 30.20	25.25 / 38.38	23.86 / 37.07	26.80 / 39.96	30.70 / 43.14
	MonoFlex-$(\boldsymbol{G})$ [50]	DLA-34	24.44 / 38.41	16.36 / 31.51	41.09 / 50.21	26.35 / 39.20	47.26 / 56.25	51.55 / 59.22	18.32 / 33.32	22.96 / 37.11

Table 8. Performance of the fine-grained categories on the Rope3D Dataset under the homologous (𝕀) and heterologous (𝕀𝕀) settings, respectively. -$(\boldsymbol{G})$ denotes adapting the ground plane, -$(\boldsymbol{D})$ means using the depth map of ground plane. The abbr. in the branch column denotes: A: anchor-based, K: keypoint-based. IoU = 0.25 for non-motor vehicles and pedestrian, IoU = 0.5 for motor vehicles.

Setting	Method	$\text{AP}_{3D\|R40}$[Mod] / $\text{Rope}_{\text{score}}$							
		IoU = 0.5		IoU = 0.7		IoU = 0.25		IoU = 0.5	
		Car	Big Vehicle	Car	Big Vehicle	Cyclist	Pedestrian	Cyclist	Pedestrian
𝕀	MonoDLE-$(\boldsymbol{G})$ [25]	51.70 / 60.36	40.34 / 50.07	13.58 / 29.46	9.63 / 25.80	24.26 / 37.35	4.14 / 21.85	4.68 / 21.70	0.44 / 18.91
	MonoDLE-$(\boldsymbol{D})$ [25]	77.50 / 80.84	49.07 / 57.22	54.53 / 62.48	17.25 / 32.00	61.81 / 67.57	35.72 / 47.22	32.60 / 44.22	12.96 / 29.03
	MonoFlex-$(\boldsymbol{G})$ [50]	60.33 / 67.86	37.33 / 47.96	33.78 / 46.12	10.08 / 26.16	65.63 / 70.78	36.83 / 48.10	24.25 / 37.70	7.58 / 24.70
	MonoFlex-$(\boldsymbol{D})$ [50]	59.78 / 66.66	59.81 / 66.07	35.64 / 47.43	24.61 / 38.01	74.09 / 77.45	50.46 / 59.03	39.33 / 49.64	13.55 / 29.50
𝕀𝕀	MonoDLE-$(\boldsymbol{G})$ [25]	19.08 / 33.72	19.76 / 33.07	3.77 / 21.42	2.31 / 19.55	10.93 / 26.44	3.72 / 21.42	2.02 / 19.32	0.47 /18.86
	MonoDLE-$(\boldsymbol{D})$ [25]	31.33 / 43.68	23.81 / 36.21	12.16 / 28.39	3.02 / 19.96	27.59 / 39.83	25.33 / 38.82	10.00 / 25.78	7.31 / 24.45
	MonoFlex-$(\boldsymbol{G})$ [50]	32.01 / 44.37	13.86 / 28.47	10.86 / 27.39	0.97 / 18.18	44.27 / 53.58	25.48 / 39.04	12.30 / 28.00	4.29 / 22.10
	MonoFlex-$(\boldsymbol{D})$ [50]	37.27 / 48.58	47.52 / 55.86	11.24 / 27.79	13.10 / 28.22	40.78 / 50.62	37.79 / 48.91	13.64 / 28.93	7.53 / 24.72

Table 9. Performance of the vanilla and improved approaches on the Rope3D Dataset under two train-val splitting settings: the homologous (𝕀) and the heterologous (𝕀𝕀). -$(\boldsymbol{G})$ denotes adapting the ground plane, -$(\boldsymbol{D})$ means using the depth map of ground.

hence only report the results by concatenation, a simple yet effective improvement strategy. We believe more sophisticated approaches might further improve the performance, which is out of the scope of this paper. In addition to the reported results of M3D-RPN (anchor-based) in the main paper, we also apply the depth map of the ground plane to MonoDLE and MonoFlex, the keypoint-based approaches. The comparison results are presented in Table 9. Comparing the vanilla and improved approaches, a consistent performance gain has been observed across all the baselines.